

UNIVERSITY OF CALIFORNIA
Los Angeles

***The Construction and Usage of a Microarray Data
Warehousing System***

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Human Genetics

by

Allen Jason Day

2008

© Copyright by
Allen Jason Day
2008

The dissertation of Allen Jason Day is approved.

Christopher J. Lee

Steve Horvath

Chiara Sabatti

Stanley F. Nelson, Committee Chair

University of California, Los Angeles

2008

TABLE OF CONTENTS

1	Introduction	1
1.1	DNA Microarray Design, Fabrication, and Assay Protocols	3
1.2	DNA Microarray Experiment Design	6
1.3	DNA Microarray Data Processing	9
1.4	DNA Microarray Informatics	15
2	Celsius: a community resource for Affymetrix microarray data	26
3	Gene Characterization Through Large-Scale Co-expression Analysis	40
4	Biopackages.net: Bioinformatics Libraries, Applications, and Data as Operating System Packages	50
4.1	Abstract	51
4.2	Introduction	51
4.3	Results	53
4.4	Discussion	55
4.5	Methods	58
4.6	Figure Captions	63
5	GMODWeb: A Web Framework for the Generic Model Organisms Database	68
5.1	Abstract	69
5.2	Introduction	69

5.3	Results	71
5.4	Discussion	78
5.5	Methods	80
5.6	Figure Captions	81
A	The Distributed Annotation System	87
B	Distinct Transcription Profiles of Primary and Secondary Glioblastoma Subgroups	95
C	Cartilage-selective genes identified in genome-scale analysis of non-cartilage and cartilage gene expression.	105

LIST OF FIGURES

4.1	Partial package dependency graph.	65
4.2	Package compilation process.	66
4.3	Biopackages.net build farm architecture.	67
5.1	GMODWeb and its relationship to Turnkey.	83
5.2	Overviews of the Turnkey::Generate and Turnkey::Render processes. .	85
5.3	An example gene feature rendered with the customized ParameciumDB skin.	86

LIST OF TABLES

5.1	The GModWeb application's software dependencies	84
-----	---	----

ACKNOWLEDGMENTS

I would like to acknowledge the guidance and mentorship of S.F. Nelson who has advised my research projects.

There are also many other people that have helped me with my research over the last five years and I sincerely appreciate their advice, contributions, and friendship, and support. For these reasons, I would particularly like to thank L. Lee, B. Merriman, M.R.J. Carlson, J. Dong, V.A. Funari, C.L. Tso, J. Braunstein, and J.M. Mendler.

Celsius, the microarray data warehouse project described in Chapter 2 , is the result of collaboration with M.R.J Carlson, J.Dong, and B.D. O'Connor. M.R.J. Carlson and B.D. O'Connor provided implementation of the web service through which Celsius is available to the public. B.D. O'Connor additionally provided implementation for the construction of the data warehouse. J. Dong provided analytical results which demonstrated that amalgamation of microarray data is methodologically sound. This project was mentored and originally conceived of by S.F. Nelson.

The gene analysis and annotation work done as part of Chapter 3 is the result of collaboration with J. Dong.

Biopackages.net, the scientific computing infrastructure project described in Chapter 4 , is the result of collaboration with B.D. O'Connor, J. Mendler and J. Fox. All collaborators provided new packages to the software repository. B.D O'Connor and J. Mendler were instrumental in implementing, testing, and maintaining the repository and the package building system. This project was mentored by S.F. Nelson and L.D. Stein.

GMODWeb, the web developer tool described in Chapter 5 , is the result of collaboration with B.D. O'Connor. Both A. Day and B.D. O'Connor provided software

implementation. This project was mentored by L.D. Stein.

In addition, I would like to thank L.D. Stein, G. Helt, S. Chervitz, S. Cain, V. Ruotti, L. Sperling, and O. Arnaiz for their advice on the informatics projects I have contributed to including Celsius, Biopackages, GMODWeb, and the Distributed Annotation System. These projects are detailed in Chapters 2 , 4 , and 5 , and in Appendix A . I would particularly like to acknowledge B.D. O'Connor for being my sounding board and an enthusiastic participant and partner in many scientific and software development endeavors.

Chapter 1 — Introduction

Chapter 2 — Celsius: a community resource for Affymetrix microarray data

A. Day was the first author of this chapter which was originally published in *Genome Biology*, volume 8, issue 6, in 2007. The corresponding author and adviser for this project was S.F. Nelson.

This article was reprinted with permission from BioMed Central Ltd. (copyright 2007).

I wish to thank M.R.J. Carlson and B.D. O'Connor for their work on making Celsius accessible via web services.

The authors thank J. Braunstein for critical comments on the manuscript. The work was supported by grants from the NINDS (U24HS052108) and the NHLBI (HL72367) with support from the NIH Neuroscience Microarray Consortium.

Chapter 3 —Gene Characterization Throught Large-Scale Co-expression Analysis

A. Day and J. Dong contributed equally to this chapter which is a manuscript in progress. S.F. Nelson is the corresponding author and adviser for this project.

Chapter 4 — Biopackages.net: Bioinformatics Libraries, Applications, and Data as Operating System Packages

A. Day and B.D. O'Connor contributed equally to this chapter which is a manuscript in progress. L.D. Stein is the corresponding author and adviser for this project.

The authors acknowledge the following individuals for their help in the development, documentation, testing and maintenance of the software and systems described here: P. Alger, A. Helsley, and V. Ruotti. Additionally, we thank S. Cain, S. Chervitz, and T. Harris for discussion related to the design of the project architecture. We thank L. Lee for designing the Biopackages.net website.

A. Day was supported by an Integrated Graduate Education and Research Traineeship grant (DGE-9987641).

Chapter 5 — GMODWeb: A Web Framework for the Generic Model Organisms Database

A. Day and B.D. O'Connor contributed equally to this chapter which is a manuscript in progress. L.D. Stein is the corresponding author and adviser for this project.

I wish to thank B.D. O'Connor for his significant contributions to the GMODWeb project which forms the basis for this chapter. I also wish to thank S. Cain, O. Arnaiz, and L. Sperling for their extensive testing of the GMODWeb software.

O. Arnaiz and L. Sperling were supported by the CNRS and by ACI IMPBio2004 contract 14. A. Day was supported by an Integrated Graduate Education and Research Traineeship grant (DGE-9987641).

Appendix A — The Distributed Annotation System

R.D. Dowell was the first author of this chapter which was originally published in BMC Bioinformatics volume 2, 2001. The corresponding author and adviser for this project was L.D. Stein.

This article was reprinted with permission from BioMed Central Ltd. (copyright 2001).

The initial ideas for DAS were developed in conversations with LaDeana Hillier of the Washington University Genome Sequencing Center. This work was supported by grants from the NHGRI (2-P01-HG00956) and an HHMI Predoctoral Fellowship grant to R.D. Dowell.

Appendix B — Distinct Transcription Profiles of Primary and Secondary Glioblastoma Subgroups

C.L. Tso was the first author of this chapter which was originally published in Cancer Research, volume 66, 2006. The corresponding author and adviser for this project was S.F. Nelson.

This article was reprinted with permission from the American Association for Cancer Research (copyright 2006).

The work was supported by grants from the National Cancer Institute (U01CA88173), NINDS (U24NS43562), Women's Reproductive Health Research Center (5K12HD001281), and an Integrated Graduate Education and Research Traineeship grant (DGE-9987641)

to A. Day.

Appendix C — Cartilage-selective genes identified in genome-scale analysis of non-cartilage and cartilage gene expression

V.A. Funari was the first author of this chapter which was originally published in BMC Genomics, volume 8, 2007. The corresponding author and adviser for this project was D.H. Cohn.

This article was reprinted with permission from the American Association for BMC Genomics (copyright 2007).

The work was supported by grants from the National Institutes of Health (HD22657, RR00425, HL072367, U24NS052108), a Joseph Down Foundation grant to Deborah Krakow, and an Integrated Graduate Education and Research Traineeship grant (DGE-9987641) to A. Day.

VITA

1977	Born, Rocklin, California
2000	B.A., Biology and Minor, Biochemistry, and Minor, Chinese, University of Oregon
2001	Scientific Programmer, Cold Spring Harbor Laboratory, New York
2002	Invited Panelist, Streaming Media East, New York
2002–2003	Invited Instructor, Course in Genome Informatics, Cold Spring Harbor Laboratory, New York
2002–2006	National Science Foundation Integrative Graduate Education and Research Traineeship (NSF IGERT) Training Grant, University of California, Los Angeles
2003	Teaching Assistant, Life Sciences, University of California, Los Angeles
2004	First Prize, MGED 9 Conference Poster Competition, Norway
2004	Invited Speaker, Brazilian Symposium on Bioinformatics & International Workshop on Genomic Databases, Instituto Militar de Engenharia, Brazil
2005	Intern, Chip Design, Affymetrix, Inc., California

PUBLICATIONS

R.D. Dowell, R.M. Jokerst, A. Day, S.R. Eddy, L. Stein. The Distributed Annotation System. *BMC Bioinformatics*. 2, 2001.

L.D. Stein, C. Mungall, S.Q. Shu, M. Caudy, M. Mangone, A. Day, E. Nickerson, J.E. Stajich, T.W. Harris, A. Arva, S. Lewis. The Generic Genome Browser: A Building Block for a Model Organism System Database. *Genome Research*. 12(10), 2002.

T.W. Harris, R. Lee, E. Schwarz, K. Bradnam, D. Lawson, W. Chen, D. Blasier, E. Kenny, F. Cunningham, R. Kishore, J. Chan, H.M. Muller, A. Petcherski, G. Thorisson, A. Day, T. Bieri, A. Rogers, C.K. Chen, J. Spieth, P. Sternberg, R. Durbin, L.D. Stein. WormBase: a cross-species database for comparative genomics *Nucleic Acids Research*. 31(1), 2003.

C.L. Tso, W.A. Freije, A. Day, Z. Chen, B. Merriman, A. Perlina, Y. Lee, E.Q. Dia, K. Yoshimoto, P.S. Mischel, L.M. Liao, T.F. Cloughesy, S.F. Nelson. Distinct Transcription Profiles of Primary and Secondary Glioblastoma Subgroups. *Cancer Research*. 66, 2006.

V.F. Funari, A. Day, D. Krakow, Z.A. Cohn, Z. Chen, S.F. Nelson, D.H. Cohn. Cartilage-selective genes identified in genome-scale analysis of non-cartilage and cartilage gene expression. *BMC Genomics*. 8, 2007.

A. Day, M.R.J. Carlson, J. Dong, B.D. O'Connor, S.F. Nelson. Celsius: a community resource for Affymetrix microarray data. *Genome Biology*. 8(6), 2007.

ABSTRACT OF THE DISSERTATION

***The Construction and Usage of a Microarray Data
Warehousing System***

by

Allen Jason Day

Doctor of Philosophy in Human Genetics

University of California, Los Angeles, 2008

Professor Stanley F. Nelson, Chair

The human genome project, which began in 1988 and completed in 2001, ushered in a new era of biology. This effort accelerated the development of biochemical assay and information technologies. As a result, biologists are now able to ask questions that were previously considered intractable. One example of a breakthrough in assay technology is the DNA microarray, a high-throughput measurement device which enables individual scientists to rapidly and simultaneously interrogate the RNA concentration levels of virtually all genes in the human genome for a single biological source. As with all advances, the advent of DNA microarray has created a new frontier of challenges. In this document, I describe an approach that addresses the problem of assembly, processing, and subsequent analysis of large volumes of data collected with DNA microarray. My work is presented in 5 chapters and 3 appendices.

Chapter 1 serves as a general introduction DNA microarray assay technology, idiosyncrasies of using this technology in biological experiments, methods for pre-processing the resulting experimental data, and techniques used in the informatic systems that enable the processing, representation, storage, and subsequent retrieval of these data.

Chapter 2 is the core of the dissertation and describes the Celsius project, a microarray data warehousing system that is an implemented solution to the informatic problems described in 1 . The completion of the Celsius project brought into existence the single largest publicly available source of primary and uniformly pre-processed DNA microarray data.

Chapter 3 builds upon Chapter 2 by describing an analysis of the data present in Celsius. Specifically, it describes the creation of gene-gene correlation matrices and their application in performing gene annotation and identifying disease genes within known linkage regions. While the idea of using gene-gene coexpression patterns is as old as DNA microarray technology itself, the scale of this analysis is unprecedented and the demonstrated applicability of the correlation data to a broad set of biological questions raises concerns about the validity of current microarray data deposition systems which rely heavily on experimental metadata.

Chapter 4 presents Biopackages.net, a technical subsystem of the data warehousing system described in Chapter 2 . Reproducibility is a shared pillar of both scientific and data warehousing methods. Because Celsius is very dependent on computing systems to process the data stored in the warehouse, it was essential to have a mechanism for making uniform and reproducible computing environments. This not only allows the system to scale as the volume of data inevitably increases, but also garners the benefits of being able to clone the system at other sites and to recover from failures.

Chapter 5 and Appendix A describe efforts for data modeling and dissemination. As we enter the post-genome era, new assay technologies continue to appear, and the growth in volume of existing and new data generated from each technology continues to accelerate. Thus, it imperative that protocols be developed for the encoding and distribution of these data to both individual scientists and the information systems and agents acting on their behalf.

Appendix B and Appendix C present analyses performed on previous iterations of Celsius, which is described in Chapter 2 . These early collaborations provided a glimpse of the utility of creating a micorarray data warehouse, without which the work described here would never have been completed.

CHAPTER 1

Introduction

The first human genome sequence, completed in 2001, ushered in a new era of biology. This effort accelerated the development of biochemical assay and information technologies. As a result, biologists are now able to ask questions that were previously considered intractable. One example of a breakthrough in assay technology resulting from genome sequencing efforts is the DNA microarray, a high-throughput measurement device which enables individual scientists to rapidly and simultaneously interrogate the RNA concentration levels of virtually all genes in the human genome for a single biological source. The microarray is one of the most commonly used of post-genomic high-throughput assay technologies, as is indicated by the number of abstracts in PubMed containing the word “microarray”. It has been widely adopted in the biological sciences and been used to address a large number of research questions across a diverse set of subdomains in biological research.

Experiments that collect data using microarray technology can be divided into several distinct categories, based on their overall goal and tactics employed in analysis of the data. These include: classification problems (supervised learning) [?], clustering problems (unsupervised learning) [?, ?], gene co-regulation and function identification studies [?], differential gene expression studies [?, ?], time-course and dose-response studies [?], gene pathway and regulatory network studies [?, ?], drug discovery and toxicology studies [?, ?, ?, ?], clinical diagnostics [?], and sequence-variation studies [?].

A successful microarray experiment is one which is able to provide insight into the hypothesis for which it was designed. Insights gained by experimentation are hard-earned, and require careful planning and execution of multiple key steps. The initial steps of feature selection, microarray design, microarray fabrication, experiment design, and biological sample processing, and image acquisition are described generally in Sections 1.1.1-1.1.2. A more exhaustive discussion of protocol surrounding these

steps is given in [?]. Following their acquisition, images are quantified, stored, distributed, analyzed, and analytical results shared. Challenges and findings related to these latter procedures are the subject of The Construction and Usage of a Microarray Data Warehousing System. They are thoroughly discussed in Sections 1.3-1.4. The entire process, from microarray design to data analysis has been formally described as an object model known as the Microarray Gene Expression Object Model (MAGE-OM) [?]. In this document I have borrowed vocabulary defined in the MAGE-OM for clarity and consistency. The MAGE-OM itself is discussed in Section 1.4.1.

1.1 DNA Microarray Design, Fabrication, and Assay Protocols

There are two basic types of gene microarray technologies currently at the disposal of researchers: single-channel and multi-channel microarrays. The methods used and form of measurements made for these two types differ, but the research questions that are addressable using either type are the same. In both systems, DNA is immobilized on a fixed surface in a specific and reproducible pattern. This is the *microarray*, sometimes abbreviated as an *array*. The pattern that allows distinct regions on the two-dimensional space of the surface to be mapped to the specific DNA sequences immobilized at that address is known as a *array design*. Each distinct region on the array is known as a *feature* and the DNA sequence present at a feature is called a *probe*. The source of the probe that is immobilized at the feature can be a cDNA clone deposited onto the microarray, or oligonucleotides that are synthesized on the microarray *in-situ* using photochemistry and light masks [?, ?].

1.1.1 Array Design and Fabrication

The first steps in performing a microarray experiment is deciding which genes to assay on the microarray, how many genes to add, and how to arrange the genes on the microarray. These steps are collectively known as *microarray design*.

In earlier times it was commonplace to manufacture multi-channel microarrays in the laboratory using a form of printer that deposited oligonucleotides onto a glass slide. This is no longer the case, as microarray production has been industrialized and is now done on a scale and level of precision that requires specialized apparatus that are not practical to host in a laboratory environment. However, the design principles remain the same. The array designer needs to select the sequences that are to be assayed by the microarray. These can correspond to regions that are transcribed to RNA, non-transcribed regions, or even oligonucleotides not expected to be observed in the biological source material to be hybridized. Historically, limitations in manufacturing technology were very restrictive, and limited the number of genes that could be assayed simultaneously. Current technology is sufficiently advanced to allow all genes to be simultaneously assayed for most organisms, so the question of which and how many genes to add is no longer relevant. However, as anyone who has performed an experiment will aver, the results obtained from the same experimental procedure performed twice will yield different results. This also holds true for measurements made with microarrays. There are many potential sources of variance in the gene measurements, and a good array design measures as much of this measurement variance as possible. Estimation of variance on a microarray essentially means making multiple measurements of the same gene. Each gene can be measured at multiple distinct locations along its sequence. Further, each location can be measured multiple times. A good array design uniformly distributes replicates across the array surface to compensate for local perturbations on the array surface itself.

Once the sequences to place onto the microarray have been determined, the sequences are either placed onto the physical substrate using printing technology, or synthesized directly onto the substrate using photolithography [?, ?]. Fabrication and microarray design are interdependent, in that improvements in manufacturing techniques allow for the fabrication of more powerful devices. Thus, this area of microarray technology is evolving especially rapidly and any specific details of the synthesis I can describe here will quickly be outdated.

Metadata are maintained that map the physical position of features in the array-design of the microarray back to the sequences they were designed to match during hybridization (Section 1.1.2). These are used by the data analyst after all data are collected to detect changes in gene expression that correspond to changes in experimental factors.

1.1.2 Hybridization

Once a microarray has been fabricated according to the array design, a solution of DNA fragments, or *targets*, can be labeled with a fluorescent dye and *hybridized*, or allowed to chemically anneal to the microarray's probes. The property of reverse complementarity between subsets of the solvent and immobilized DNA drives the solvent DNA targets to localize near to their corresponding immobilized probes.

After hybridization, the amount of target DNA hybridized to each feature on the array is assayed. A laser is used to excite the fluorescent dye associated with the targets, and a digital camera is used to measure the amount of fluorescence present at each feature, and at the specific wavelength of fluorescence known as a *channel*. The amount of fluorescent light is assumed to be a monotonic and increasing function of the amount of target DNA hybridized to that feature. In the case of single-channel assays, only a single fluorescent dye is used. In a multi-channel assay, multiple target solutions

are hybridized to the array, each labeled with a different fluorescent dye. Each digital image of a microarray's fluorescence pattern for a single channel is known as an *image*, and the process of observing and recording this image is called *acquisition*.

For a single-channel assay, the acquired image is processed to produce a matrix of numbers that has the same dimensions as the array design, and whose values correspond to the absolute intensity of light present at each feature. A similar matrix is produced for a multi-channel assay as well, but because of the competitive hybridization between the multiple labeled targets, the values reported are not absolute, but rather the intensity of each channel relative to all other target channels at a given feature. Typically one of the channels corresponds to a *reference sample* that is a common denominator to several arrays that will be analyzed as a set. This allows the relative values for the non-reference samples on each array to be compared.

For both single-channel and multi-channel arrays, the next step is *quantification*, a series of processes designed to convert the “raw” fluorescence intensities acquired from the scanner into values that are usable in higher-level data analyses such as classification. Methods for quantification of microarray data are an area of intense interest [?, ?, ?, ?, ?, ?, ?, ?, ?] because they can give very different estimates of the quantity of DNA present for a given feature/channel, and thereby have a large effect on the biological conclusions that can be drawn from analyses of those estimates. Quantification methods are discussed in greater detail in Section 1.3.

1.2 DNA Microarray Experiment Design

Scientific inquiries that use microarray technology to address research questions are referred to as *microarray experiments*. The majority of these experiments can be divided into two classes: pattern detection and predictive modeling. These are described

in greater detail in Sections 1.2.1-1.2.2, and specific examples of these experimental classes can be seen in [?, ?, ?, ?, ?, ?]. The method descriptions given here are described in terms of the characterization of biological samples, but it is important to bear in mind that the samples-by-genes data matrices can be transposed and that these same methods can be used for the characterization of genes across multiple phenotypic conditions. Indeed, this is also an area warranting further exploration given that gene-centric analyses are not stymied by the “curse of dimensionality”, discussed in greater detail in Section 1.2.3.

1.2.1 Pattern Detection

The first class of experiment is data-driven, in which an *a-priori* hypothesis is not explicitly stated and sample metadata in the form of labels are not provided. This data driven method is sometimes also called *pattern detection* because experiments employing it focus mainly on identifying biologically interesting and previously unobserved correlations between factors of the experiment [?]. A common type of pattern detection experiment seen in the literature examines biological samples that are indistinguishable using lower-resolution assay technologies to see if the samples can be broken into sub-groups by collecting more data using the higher-resolution microarray. A study by Freije, *et al.* used this approach to identify novel subclasses of malignant gliomas [?].

1.2.2 Predictive Modeling

In the second general class of experiment seen in the literature, the emphasis is on constructing a function that is able to *classify*, or predict class labels for unlabeled samples. This method is sometimes called *supervised learning* or *predictive modeling* [?]. A predictive modeling study begins with a group of samples for which assay data as well as class labels are available. This initial set of labeled samples is divided

into two sub-groups, *training data* and *test data*. The training data are processed using a classification algorithm to build a model of the relationship between class label and assay data. One major area of concern in the construction of the model and the application of that model to unlabeled samples is the number of features considered. Feature selection is discussed in greater detail in Section 1.2.3.

The performance of the classification model produced using the training data is typically assessed and iteratively optimized to minimize error using K -way cross validation for all class labels K . However, in some cases class labels K are nested meaning that the class k_i can be a refinement of the class k_{i-1} . In these hierarchical classification models, optimization and error minimization consider the graph structure [?]. Once a satisfactorily low error rate has been achieved, the predictive power of the classification model can be estimated by using it to predict labels for the test data that were set aside prior to model building.

1.2.3 Feature Selection

A prominent part of model building for the analysis of microarray data is *feature selection*, or the process of identifying which features of observations are relevant in the assignment of class labels. Selected features correspond to the microarray feature measurements, or some summarization thereof (Section 1.3.4).

Feature selection has been essential because the analytical methods applied to microarray experiments bear the “curse of dimensionality” [?]. In a typical experiment the number of samples is small ($1 \times 10^1 \dots 1 \times 10^3$), while the number of measurements made on each sample is relatively large ($1 \times 10^3 \dots 1 \times 10^6$). Because of this it is possible to correctly partition labeled data on one or multiple subsets of all observed features. Using a subset that robustly captures the sample labels is desirable because it increases classification accuracy and reduces the computation required required to cal-

culate a new sample label, as many classification algorithms scale exponentially with the number of features considered [?].

However, the work described in Chapter 2 presents a mechanism for aggregation and analysis of much larger numbers of samples than are commonplace in experiments to date.

1.2.4 Label Encoding

In both the pattern detection (Section 1.2.1) and predictive modeling (Section 1.2.2) classes of microarray studies, a class label from one or more multiple orthogonal experimental factors is attached to a sample. Class labels are typically “flat”, meaning that there is no hierarchical structure and that the distance between any two classes is uniform. However, It is well-established that such a flat representation is not suitable for representing gene annotation [?], and it has been demonstrated that statistical analyses that consider the structure of relationships between annotations bear more power than their flat counterparts [?, ?, ?]. Thus, analyses that explicitly model the relationship types and distances between class labels of samples are expected to become more commonplace as open, community-supported, standard structures for encoding sample annotations mature [?, ?, ?, ?, ?, ?, ?, ?]. Encoding metadata using open standards also has implications with regard to integration of results from multiple studies, or otherwise exchanging data. These aspects are discussed in Section 1.4.2.

1.3 DNA Microarray Data Processing

Microarray pre-processing, also known as *quantification*, is the process of estimating the quantity of each gene in the sample that was assayed in the hybridization step (Section 1.1.2) of the experiment. Quantification can be broken down into four distinct

sub-procedures, executed in the following order: image processing, background correction, normalization, and summarization. Much of the specific details in this section are idiosyncratic to the Affymetrix GeneChip platform, but general principles apply to all forms of microarray technology.

1.3.1 Image Processing

The term *image processing* is used to describe a set of steps that transforms the physical microarray into a digital file suitable for subsequent processing by a computer. The technical details of the protocol vary, but the general principles described here remain the same.

As discussed in Section 1.1.2, fluorescent dyes are linked to the synthesized DNA sequences that are hybridized to the microarray. The microarray is placed in a digital scanning apparatus that contains laser(s) that can emit light at the excitatory frequency of each dye. The apparatus also contains photosensors that are able to detect the fluorescent light from the dye(s). Each laser, then, is scanned across the surface of the microarray and the photosensors record the position and intensity of light for each dye. The raw form of these data is typically a series of images encoded using a lossless format, such as TIFF. The image is then processed by an alignment algorithm that aligns the captured image to the coordinate system known to have been printed/synthesized onto the microarray surface.

Finally, platform-specific protocols are used to represent the raw image in a format more suitable for downstream processing. For example, with single-channel Affymetrix microarrays, the TIFF image is converted to an alternate format, called *CEL* format, that describes the attributes of each microarray feature, such as mean intensity, variance, and number of pixels that represent the feature.

1.3.2 Background Correction

Background correction is a statistical procedure that estimates and removes low levels of noise on the microarray. Background noise can have many sources.

The simplest and most common source of background noise is optical. It can be caused by general cross-hybridization of target to all probes, mis-calibration of the microarray scanner's photo-sensor, and diffused or reflected light from the laser used to excite the fluorescent dyes. Optical noise can be estimated by measuring the level of fluorescence from featureless regions of the microarray and negative control probes that are not reverse-complementary to any sequences in the hybridization mixture. These measure background-level reflected light and the level of non-specific hybridization, respectively.

Manufacturing and hybridization artifacts, such as surface scratches and salt residues, are another source of noise. A simple form of location-based background correction is described in the Statistical Algorithms Description Document [?]. Briefly, the chip is broken into a 4×4 grid of 16 rectangular regions. The lowest 2% of each region's probe intensities are used to compute a background value for that region. Each probe (PM and MM) is then adjusted based upon a weighted average of the backgrounds for all regions. The weights are based on the distances between the location of the probe and the centroids of all regions. More sophisticated methods attempt to detect areas of the microarray containing high levels of manufacturing and hybridization noise. Noisy areas can be identified because the probes located there will be outliers relative to probes for the same target located elsewhere on the microarray. Probes in these areas are considered unreliable and are either given a very low weight parameter or are removed from normalization (Section 1.3.3) and other downstream processing (Section 1.3.4) altogether [?].

Newer, multi-array background correction methods have leveraged existing data to

build a models of how background noise is generally distributed. The gcRMA model [?] includes a parameter the sequence composition of each probe, while other models such as those used in the RMA and MBEI (dChip) [?, ?, ?, ?, ?] methods only include a parameter for concordant each probe is with other probes in the same set. The RMA background correction method is the *ide facto* standard, and corrects perfect match (PM) probe intensities by using a global model for the distribution of probe intensities [?, ?].

Multi-array background correction methods are able to detect background noise due to the manufacturing and hybridization artifacts described above, but the size of the array artifact can be as small as a single feature. This should in principle do a better job of noise estimation. A major drawback to multi-array background noise models is that the noise estimates are only valid in the context of the co-processed set of microarrays. This is because the noise estimates are derived from parameter estimates specific to that set of microarrays. While this is not a problem for small-scale analysis on individual experiments, it creates difficulties when merging data from multiple experiments because all microarrays will need to be re-processed to re-fit the parameters of the noise model. This re-processing problem can become intractable for background correcting a very large number of arrays. Further attention is given to tractability in Section 1.3.5.

1.3.3 Normalization

After correcting for background noise (Section 1.3.2), microarrays are normalized. The purpose of normalization is to transform the distribution of microarray measurements so that properties of the distribution of measurements match expectations (e.g., a log-normal distribution).

The most simple form of microarray normalization is a linear scaling. The Affymetrix

MAS 5 algorithm [?] performs linear scaling by (1) setting aside the top and bottom 1% of measurements as outliers, adjusts the mean of the remaining measurements to a constant value, then multiplies each measurement, including the outliers, by the factor used to adjust the mean.

The normalization method used in the dChip software [?, ?] selects a baseline microarray. Then, all microarrays are normalized by selecting invariant sets of probes within each of the “treatment” and “baseline” conditions. These are used to fit a non-linear relationship between the two conditions, and this relationship is used to carry out the normalization.

Many other normalization methods exist. The essentially differ in two aspects: the theoretical model of how the microarray behaves, and the techniques used to fit the observed data to that model. In this regard, normalization techniques are similar to those used for background correction (Section 1.3.2) and summarization (Section 1.3.4).

A comparison of the performance of a large number of normalization methods at correctly estimating RNA concentration on a standard, synthetic data sets published by Affymetrix is described in [?].

1.3.4 Summarization

The last step in microarray data preprocessing is to combine the measurements from all probes in a probe set into a single value. This procedure is called *summarization*.

The simplest summarization algorithm, called “average difference” [?] computes the mean of difference between each PM/MM probe pair (Equation 1.1),

$$y_k = I_k^{-1} \sum_{i=1}^{I_k} |PM_i - MM_i| \quad (1.1)$$

where the probe set k has PM perfect match and MM mismatch probe pairs $i = 1, \dots, I_k$.

Summarization methods parallel background correction and normalization methods in that there are two varieties, the single-array methods and the multi-array methods. “Average difference” is an example of the former. Multi-array methods consider the distribution of probe measurements across all arrays, and in some cases assign an array-specific parameter used to compute the probe set summary. The summarization component of the MBEI method introduced by Li and Wong [?, ?] is given in Equation 1.2,

$$y_{ij} = \phi_i \theta_j + \epsilon_{ij} \quad (1.2)$$

where y_{ij} is PM_{ij} or the difference between $PM_{ij} - MM_{ij}$. The ϕ_i parameter is a probe response parameter and θ_j is the expression on microarray j .

The summarization component of RMA pre-processing [?] performs a multi-array linear fit to data from each probe set. Specifically, for probe set k with $i = 1, \dots, I_k$ probe pairs and microarrays $j = 1, \dots, J$ the model given in Equation 1.3 is fit,

$$\log_2 \left(PM_{ij}^{(k)} \right) = \alpha_i^{(k)} + \beta_j^{(k)} + \epsilon_{ij}^{(k)} \quad (1.3)$$

where α_i is a probe effect and β_j is the \log_2 expression value, and the method is known as *median polish*, named after Tukey’s algorithm used to perform the calculation.

It is noteworthy that summarized probe set values from all popular multi-array summarization methods, including those described here, are dependent upon the probe set and microarray effect parameters calculated as part of the model fit. While this is not a problem theoretically, it introduces unique challenges in the implementation of a

pre-processing pipeline for a large number of arrays. This is discussed in greater detail in Section 1.3.5.

1.3.5 Scalability

While the superiority of multi-array techniques described in Sections 1.3.2-1.3.4 has been established, a major impediment to implementing these in practice is that they are not designed to scale to large data sets due to the amount of system resources consumed by a single quantification process. One approach to solving this problem is “divide and conquer”. By breaking the matrices up row- or column-wise, larger batches of microarrays can be co-processed [?]. While this is useful for performing analysis on a static data set, it does not address the needs of users performing analysis on constantly growing data sets, because each time an additional microarray is added the entire set must be reprocessed. Katz *et al.* [?] recognized that the true advantage of multi-array over single-array quantification methods is that they fit probe and probeset behavior to a model, and that if a theoretical (as opposed to empirical) model is used and reasonable estimates can be calculated and saved, arrays can be added one by one without reprocessing the whole set. A similar system was implemented in Celsius, and is described in greater detail in Chapter 2 [?].

1.4 DNA Microarray Informatics

1.4.1 Data Modeling

In order to be able to perform analyses on the results of a DNA microarray experiment, the digitized assay data (Section 1.3.1) must be structured and stored. This is also true of the experimental *metadata*, or description of the experimental design and procedures. Principles of scalability and mass-production prescribe that the encoding

of these data be uniform, meaning that the structure used to encode the experiment must be sufficiently flexible and descriptive to capture the full range of experiments that can be conceived.

The *MicroArray Gene Expression (MAGE)* Group was established by the *MicroArray Gene Expression Data Society (MGEDS)*, sometimes also referred to as *MGED*) to develop a standard for the representation of microarray data, with the intent of making those data exchangeable between different information systems [?]. The work of the MAGE Group can be divided into two types of projects: those dealing with the syntax used to concretely represent microarray data, and those dealing with the semantics used to describe the microarray experimental materials and processes, produced data, and data derived from processing [?, ?, ?].

The semantic developments of the MAGE Group [?] are more important for *The Construction and Usage of a Microarray Data Warehousing System* than the syntactic developments. This is because semantics are abstract and generally useful for encoding experimental information, while the syntactic developments of the MAGE Group have optimized for the purposes of data interchange using technologies such as the *eXtensible Markup Language (XML)* that are not well-suited for systems whose primary purpose is to store and retrieve large volumes of quantitative data.

Application of the MGED semantic technologies, such as the MGED Ontology [?], are discussed elsewhere in this document (Section 1.2.4). There are two major semantic developments from MGEDS. The first of these is the *MAGE Object Model*, or *MAGE-OM*. The purpose of the MAGE-OM is to provide a standard set of classes that can be used to represent any object or process that may be included in, or referenced from, a microarray experiment [?]. It draws heavily from a seminal work by Brazma *et al.* describing the “Minimal Information About a Microarray Experiment”, or *MI-AME* that should be collected [?]. The model employs concepts common to object-

oriented software design and knowledge engineering, such as subclass/superclass relationships between object classes, the notion of abstract classes, and the possibility for directional and cardinal relationships between objects. The specification of the MAGE-OM was developed using the *Unified Modeling Language (UML)*, a standard technology used by knowledge and software engineers for the composition of object models. Fitting microarray experiment data into the MAGE Object Model is described in greater detail in Section 1.4.1.1. Encoding specific information about objects under investigation or that are used to facilitate the conduct of an experiment are discussed in Section 1.4.1.2.

Unique syntax-related challenges that have arisen as part of *The Construction and Usage of a Microarray Data Warehousing System* are also presented. Internal data representation and scalability issues are discussed in Section 1.4.3.2 and interoperability concerns are discussed in greater detail in Section 1.4.2.

1.4.1.1 Experimental Metadata

As a concrete example of how an experiment might be encoded into a MAGE-ML document, consider the now-classical study of leukemias by Golub, et al [?]. Briefly, this study is has both pattern-detection and predictive-modeling methodology (Sections 1.2.1-1.2.2), and describes a method for identifying features that discriminate between two leukemia subclasses and how they can be used to identify the subclass of previously unlabeled cancer samples.

Encoded into the MAGE-OM, the initial cancer samples in Golub, et al [?] are represented as *BioSource* objects, a class used for biological material prior to any treatment. Each BioSource object then goes through a series of modifications, ultimately resulting in a *LabeledExtract* object that represents the fluorescently-labeled cRNA that is hybridized onto a microarray. In the series of modifications, a combination of

a *BioSample* object and one or more *Treatment* objects is used to represent the transformed Biosource object. Further, each LabeledExtract refers back to the object from which it was derived all the way back to the BioSource object so that the full path of derivation via treatment is modeled.

To represent the microarray hybridization, a *BioAssay* object is created. The BioAssay is a central connection point in the object model. It refers to an *Array* object representing the microarray itself, the LabeledExtract that is hybridized, and to one or more *Factor* objects. The factor objects are in turn related to a network of other objects that encode the design and variables used in the microarray experiment. The Array object is associated to a series of other objects that describe the microarray itself, including the information about specific sequences and their physical locations on the microarray, as well as information about the grouping of features on the array used as reporters for a common cRNA target sequence. Further, each LabeledExtract is associated with a *Channel* object. The Channel is used to link a specific LabeledExtract to a specific *Image* object that results from the scanning of the hybridized array.

The Image object is combined with a *FeatureExtraction* object to produce a *BioAssayData* object. This association of objects represents that transformation of the microarray image acquired by the scanner (Section 1.1) into a numerical form that can be processed (Section 1.3) for further analysis. BioAssayData objects may also be derived from other BioAssayData objects, similar to the way BioSource, BioSample, and LabeledExtract objects may be related. This is how the MAGE-OM represents arbitrary data transformations, and is sufficient for describing microarray data pre-processing (Section 1.3), as well as additional downstream summarizations or other transformations of these data, such as sample or probe set clustering.

1.4.1.2 Object Metadata

Objects in the MAGE-OM may have attributes attached to them to provide more specific detail about the microarray experiment. A design decision was made to reference, from the MAGE-OM, objects from an *ontology* for the description of objects. Doing so constrains the scope of MAGE-OM's purpose to the structure of the microarray experiment and associated quantitative data.

Ontologies are a key crossover into the biological sciences from computer science, information theory, and artificial intelligence research. According to T. Gruber [?], an ontology is “an explicit specification of some topic. For our purposes, it is a formal and declarative representation which includes the vocabulary (or names) for referring to the terms in that subject area and the logical statements that describe what the terms are, how they are related to each other, and how they can or cannot be related to each other. Ontologies therefore provide a vocabulary for representing and communicating knowledge about some topic and a set of relationships that hold among the terms in that vocabulary.”

As biologists have always faced the problems of nomenclature and classification, ontologies are a natural extension of these activities. Many high-profile ontologies have been created to annotate and classify a wide variety of biological concepts. The *MGED Ontology (MO)* was developed by MGED specifically for describing the attributes of objects used or studied as part of an experiment [?], and as a formal encoding of the concepts of the “Minimal Information About a Microarray Experiment” (*MIAME*) developed by Brazma, et al [?]. The MO is used in *The Construction and Usage of a Microarray Data Warehousing System* for high-level grouping of experiments by their general design (time-course, dose/response, etc).

Other ontologies are also useful in *The Construction and Usage of a Microarray Data Warehousing System* for attaching concepts to objects that are outside the scope of

the MO. For example, we have used the Sequence Ontology (SO) [?] to encode information pertinent to the sequences present on, or used to select the sequences present on, microarrays. In the case of using the SO, the concepts from the ontology are non-orthogonal to those in the MO and therefore there is some redundancy in the annotation stored. In most cases, however, the ontologies used to annotate objects in the data warehouse are orthogonal. We have used the Adult Mouse Anatomy Ontology (MA) [?] for attaching tissue information onto mammalian samples, the Mammalian Pathology Ontology (MPATH) [?] for annotating disease states, and the Cell Ontology (CL) [?] for annotating cell type onto samples that have been grown *in-vitro*.

The use of ontologies provides a mechanism for attaching consistent and unambiguous descriptions to objects relevant to *The Construction and Usage of a Microarray Data Warehousing System*. The encodings are useful in that they allow for the accurate representation, storage (Section 1.4.3.2), retrieval, and exchange (Section 1.4.2) of information about these objects. Encoding object descriptions as ontology terms also presents interesting data analysis opportunities because of the ontology structure. The relationship between terms in the ontologies is structured as a directed graph, a data structure whose properties are well understood and for which a large body of theory and analytical methods already exist. Indeed, several prior studies have leveraged the annotation of genes into the Gene Ontology (GO) [?] to predict gene function [?, ?, ?], and these techniques can be extrapolated to annotation of the hybridized biological source materials as well.

Ontology construction continues to be an active area of development in the biological sciences, and many important projects have spearheaded the effort to establish ontologies for the biological research community. The *Open Biomedical Ontologies (OBO)* project provides a range of ontologies designed for use in the biomedical fields [?]. Other related projects include the Gene Ontology, and the Sequence Ontology

[?, ?]. Recently, the National Institutes of Health recognized the critical role ontologies are playing in the organization and interchange of biological information and founded the National Center for Biomedical Ontology [?] as a coordinating center for ontologies, similar to what the *National Center for Biological Information (NCBI)* [?] and *PubMed* [?] do for sequence and bibliographical data.

1.4.2 Interoperability & Exchange of Data

Related to the concept of open APIs for interacting with biological data is the concept of web services. Sharing of biological data, whether it is raw, primary data, or processed results, is of the utmost concern for biologist. Yet journal articles provide a poor repository for large datasets. Simply downloading large data files is a better approach, yet many times only a few pieces of information are needed and the downloading of an entire dataset is unnecessary. For example, if a researcher wants a particular protein structure from the protein data bank (PDB) [?] downloading the thousands of structures in bulk from the PDB is unnecessary.

The emerging web services approach used across the Internet points to a better solution. This model includes technology such as SOAP, the Simple Object Access Protocol and ReST concepts for interacting with an API remotely over the hypertext transport protocol (HTTP) on which the web is based [?]. This idea is that simple requests for data or information calculated on the fly can be made by a researcher and the result is calculated or retrieved remotely and returned over the Internet. This type of approach is extremely flexible and can be used in a variety of contexts to present biological data to other researchers.

The distributed annotation system, or DAS, is a popular bioinformatics web services project geared towards the sharing of genome annotations with the larger research community [?]. Organizations looking to share genome assemblies, gene anno-

tations, and other genomic features use DAS to make this information available over the web. Implementations of DAS use the standard HTTP protocol and XML as an exchange standard. The next version, DAS/2, expands on the genomics focus of DAS by including capabilities to exchange ontologies, download experimental assay results such as microarray data, and perform on-demand sequence analysis such as BLAST [?]. The success of DAS as a project is due to the ease of which scientists can utilize information published with DAS. Many clients exist, such as the Generic Genome Browser because the web services model affords programmatic access to the servers [?, ?]. This allows additional applications to be built on top of these public repositories. For example, the Celsius project web interfaces, described in Chapter 2, were created on top of a DAS/2 server which provided the raw data. These web tools let an end user query the microarray data available via ontology annotations and download the corresponding data in a variety of different processed forms.

More information on web services can be found in Chapter 2 and Chapter 5, which examines a model organism website generation framework that includes web services tools.

1.4.3 Computing Infrastructure

One of the key principles that has enabled the creation of high-throughput methods is scalability. Simply put, scalable methods are able to process large amounts of data at the same level of performance as small amounts. For biological assays this is largely a problem of parallelization of chemical reactions and miniaturization of devices. In terms of computational infrastructure, building effective, scalable systems also requires parallelization. This is often referred to by scalability engineers as *horizontal scaling* [?, ?].

When setting up an infrastructure system using the *horizontal scaling* pattern, com-

puter resources can be treated as groups of resources, or *clusters*. Each cluster is responsible for one or more types of tasks, and each component of the cluster is uniform. This allows the throughput of the cluster to be scaled simply by adding or removing components.

The effort needed to maintain each of a cluster of computers may exceed that needed to maintain computers individually. For example, additional software may be necessary to orchestrate computation across the cluster [?]. Contention for storage and network bandwidth may also be a concern. Hardware that is employed, while at face-value appears to be of identical make and model, may in fact contain different versions of microchip sub-circuits, resulting in inconsistent behavior. Finally, there is the issue of managing the software present on each computer within the cluster so that they produce equivalent results. As part of *The Construction and Usage of a Microarray Data Warehousing System*, we have had to deal directly with the problems of storage and maintaining consistency of software on a computing cluster. Storage issues are discussed in Section 1.4.3.2, and the problem of maintaining software consistency was addressed using Puppet [?] along with the construction of an RPM [?] software repository that is described in greater detail in Chapter 4 .

1.4.3.1 Standardization

A closely related concept to web services is the concept of software standardization. In the model of web services, a researcher can focus on the analysis of data and its biological meaning rather than figuring out how to store data locally. This approach affords abstraction between the researcher and the entity providing the web service, making it easier for others to either validate existing work by performing the same analysis or expand on the work using the same web services. It allows researchers to easily standardize a given dataset or analysis server. Another technique familiar to all com-

puter users on standardization is the versioning of computer programs. When research is being performed on a particular dataset or with a particular software program, it is extremely important to track which version was used. Otherwise it becomes impossible to replicate the work. The idea of software packaging, borrowed from the field of information technology, is of key importance to bioinformatics. In addition to simply versioning software, many comprehensive systems exist for specifically tracking, installing, and updating both software and data in a particular computing environment. The Linux system, for example, uses one of several different package managers to perform this task.

The Biopackages project looks to standardize many tools used commonly in bioinformatics projects. It encompasses an automated build system that creates software packages for particular Linux distributions. These include packages for APIs such as BioPerl [?] and BioConductor [?], web services such as DAS/2 [?], and databases such as Chado [?]. Details of the construction, public availability, and benefits of this standardization tool can be found in Chapter 4 .

1.4.3.2 Storage

The data and metadata produced as part of pre-processing (Section 1.3) and fitting the microarray experiment to a uniform data model (Section 1.4.1) must be stored and made available for retrieval at a later time to analysts to perform further processing.

Storage solutions borrowed from computer science and the information technology industry include physical media on which to store the data, such as hard drives, and also database systems to structure the data in accessible and searchable contexts. Hard drive storage systems have advanced considerably over the decades in response to the demand for safe, reliable, and cost effective ways of storing large amounts of data. Systems that link together many individual hard drives or stor-

age on groups of computers into contiguous virtual volumes are available, making it possible to group together large datasets in a common repository. Implementations of contiguous volumes, such as redundant arrays of inexpensive disks (RAID) and storage area networks (SANs) make the storage of critical scientific data secure and retrieval speedy. Database systems represent another strategy for meeting the storage needs of bioinformatics projects. Unlike hard disk-based solutions, databases actively index information to improve the retrieval of structured data. Advances in open source projects such as MySQL (<http://www.mysql.com>) and PostgreSQL (<http://www.postgresql.org>), have allowed researchers to use very high performance relational database systems in their research for minimal cost.

Many database solutions exist for representing biological data. The Generic Model Organism Database Project (GMOD, <http://www.gmod.org>) provides the modular Chado schema [?] for storing a wide variety of biological data. This schema has been used by a variety of projects, and is used as the primary relational database schema in *The Construction and Usage of a Microarray Data Warehousing System*. Specific details of how the Chado database schema is used is given in Chapter 2 and Chapter 5 .

CHAPTER 2

Celsius: a community resource for Affymetrix microarray data

Celsius: a community resource for Affymetrix microarray data

Allen Day, Marc RJ Carlson, Jun Dong, Brian D O'Connor and Stanley F Nelson

Address: Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, California, 90095, USA.

Correspondence: Stanley F Nelson. Email: snelson@ucla.edu

Published: 14 June 2007

Genome Biology 2007, **8**:R112 (doi:10.1186/gb-2007-8-6-r112)

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2007/8/6/R112>

Received: 2 February 2007

Revised: 9 May 2007

Accepted: 14 June 2007

© 2007 Day et al.; licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Celsius is a data warehousing system to aggregate Affymetrix CEL files and associated metadata. It provides mechanisms for importing, storing, querying, and exporting large volumes of primary and pre-processed microarray data. Celsius contains ten billion assay measurements and affiliated metadata. It is the largest publicly available source of Affymetrix microarray data, and through sheer volume it allows a sophisticated, broad view of transcription that has not previously been possible.

Background

DNA microarrays have become the most important source of experimental genomic information that are applied in a large scale. They are widely used for tissue/disease classification as well as gene function discovery. Applications of this technology are routinely and widely published within almost all aspects of biology and human disease studies, with more than 14,000 PubMed citations containing the word 'microarray' published between 1996 and 2007. Even in the early years of microarray experimentation, it was widely recognized that a central repository of this information should be created to house these data. This enables potentially important additional information to be gleaned by re-interpretation by other researchers, perhaps in different contexts or in relation to new data. Thus, major efforts to house such data were made, namely the Gene Expression Omnibus (GEO) [1] and ArrayExpress (AEX) [2]. These repositories contain more than 82,000 and 50,000 microarray hybridizations of data, respectively. Primary data are expensive and time consuming to generate. In spite of the high cost, such experiments are rarely fully mined for their information content. Indeed, several meta-analyses have been reported that were based on archived data [3,4]. These studies demonstrate the benefit of

data repositories and that additional inferences are possible with reanalysis.

Although gene expression microarray technology has been implemented in a variety of formats (spotted cDNAs, spotted column-synthesized oligos, and *in situ* synthesized oligos), the leading commercial supplier of microarrays has been Affymetrix Inc. (Santa Clara, CA, USA) since 1996. Within the GEO repository Affymetrix platforms account for 35% of all arrays deposited, but they represent approximately 60% of the genome-scale gene expression data. For instance, Affymetrix platform arrays account for the top seven array platforms in terms of the number of arrays deposited in GEO. Thus, in the public domain within repositories, this platform type forms the richest set of expression information that can be most readily combined in a useful manner for meta-analyses spanning multiple experiments. Furthermore, the Affymetrix platform has a standard set of protocols for probe generation and labeling, uses a single color detection system, and has a relatively reliable array fabrication process. The Affymetrix platform is widely applied to a variety of biologic problems. Thus, this platform is highly attractive as the basis for amalgamation of data from many different sources. In theory,

historic arrays can be directly compared with additional experiments and provide an important tool for comparative analyses. However, because of the large number of analytical procedures for normalization and quantification from the oligonucleotide level data, it is greatly preferable to reanalyze primary data in the form of processed image files (termed CEL files, or CELs). This permits substantially more robust comparisons between datasets because the same analytical metric can be applied to the joint data and will ultimately permit more thorough vetting of algorithms to assess gene expression levels from this platform.

Based on the popularity and ease of use of the Affymetrix platform we began to construct a combined resource for the storage of publicly available CELs for ongoing comparison with data generated at the University of California, Los Angeles (UCLA) DNA Microarray Core Facility as part of the National Institutes of Health Neuroscience Microarray Consortium (NNMC). The purpose of this assembly of CELs was to create a substantial reference set of primary data that would then be available for all ongoing projects. As we examined the available CEL file resources, it became apparent that fragmentation of public data into multiple small repositories has effectively occurred despite the presence of two major repository efforts and deposition requirements of journals. Of the more than 30,000 instances of CELs that were collected from 11 institutional servers (Figure 1) [1,2,5], fewer than 5% are present as CELs in either GEO or AEX, the two official public repositories. We estimate that up to 90% of generated CELs are not yet deposited in AEX or GEO. In fact, most public CELs are not easy to find. This suggests that the number of publicly available CELs is much larger than that used in our study, but these CELs are not accessible using standard bulk-mode data retrieval network protocols such as network file transfer protocols FTP and Rsync.

We further note that inconsistent annotation of experiments impedes meta-analysis. Re-use of these data is compromised by the low quality of clinically or experimentally relevant annotated metadata actually available for many datasets, as well as the inconsistent and incomplete implementation of the standards for encoding these metadata [6,7]. For instance, no repository uses controlled vocabularies, and therefore the annotation of experiments can be ambiguous and difficult to use when integrating datasets.

Here, we present a community-oriented structure to permit massive amalgamation of microarray data for joint analyses. We have termed this resource 'Celsius', to reflect both the intended community spirit and the restriction to image files generated from the Affymetrix platform. Celsius has four major goals: to import all available Affymetrix primary data, whether published or not, specifically gene expression, genotyping, and tiling CELs; to process imported data using best-of-breed statistical methods made available by the community; to facilitate and encourage community involvement in

annotation of deposited samples using controlled vocabularies; and to make available for re-export consistently quantified and normalized data that can be combined without further processing. In this article we describe the methods employed to create this resource, a snapshot of its contents, nascent systematic approaches to annotate samples and genes solely using expression data, and growth rate.

Results and discussion

Data overview

Celsius contains an agglomeration of more than 61,000 CEL files, each of which represents a single microarray hybridization performed using Affymetrix technology on one of 156 different array designs. The majority (67%) of CELs are derived from only ten array designs, as shown in Table 1. Of all CELs in Celsius, 95% contain gene expression measurements, 4% contain human DNA allelic copy number measurements, and the remainder of the CELs contain tiling and re-sequencing data. Within the gene expression data, nearly 50% were collected from human tissues or cell lines and nearly 20% from mouse tissues or cell lines (Figure 2).

Only primary data are imported, all of which were collected using the Affymetrix platform, a popular technology that represents more than 70% of all microarray data in GEO. The primary data in Celsius are the union of CELs collected from more than 11 institutions, including the two central repositories for microarray data: GEO and AEX. Celsius is continuously updated, and has a growth rate of 1,000 CELs/week, as observed from January 2006 to January 2007 (Figure 3). The size and growth rate of this dataset are corrected for inter-repository as well as intra-repository file-level redundancy, because approximately 5% of CELs are available from more than one institution or are available as replicates but by multiple database accession identifiers from a single institution. As of January 2007, Celsius is the world's largest publicly accessible resource for microarray data derived from the Affymetrix platform and contains three times as many CELs as GEO and ten times as many CELs as AEX, which are the largest public microarray data repositories. An illustration of the CEL load process is given in Figure 4. This illustrates redundancy checking and assignment of the serial number database identifiers (SNIDs; the primary database accession identifiers used by Celsius).

Data processing

CELs loaded into the Celsius system are processed using many best-of-breed statistical quantification algorithms, including dChip, BRLMM, GC-RMA, MAS5, PLIER, RMA, and VSN [8-12]. Some of these algorithms are in the multi-array class of algorithms and require co-processing a batch of CELs to provide a confident signal estimate for each of the probesets. Each CEL loaded into the Celsius system is processed together with a selected 'quantification pool' of 50 CELs of the same array design that is held constant for all

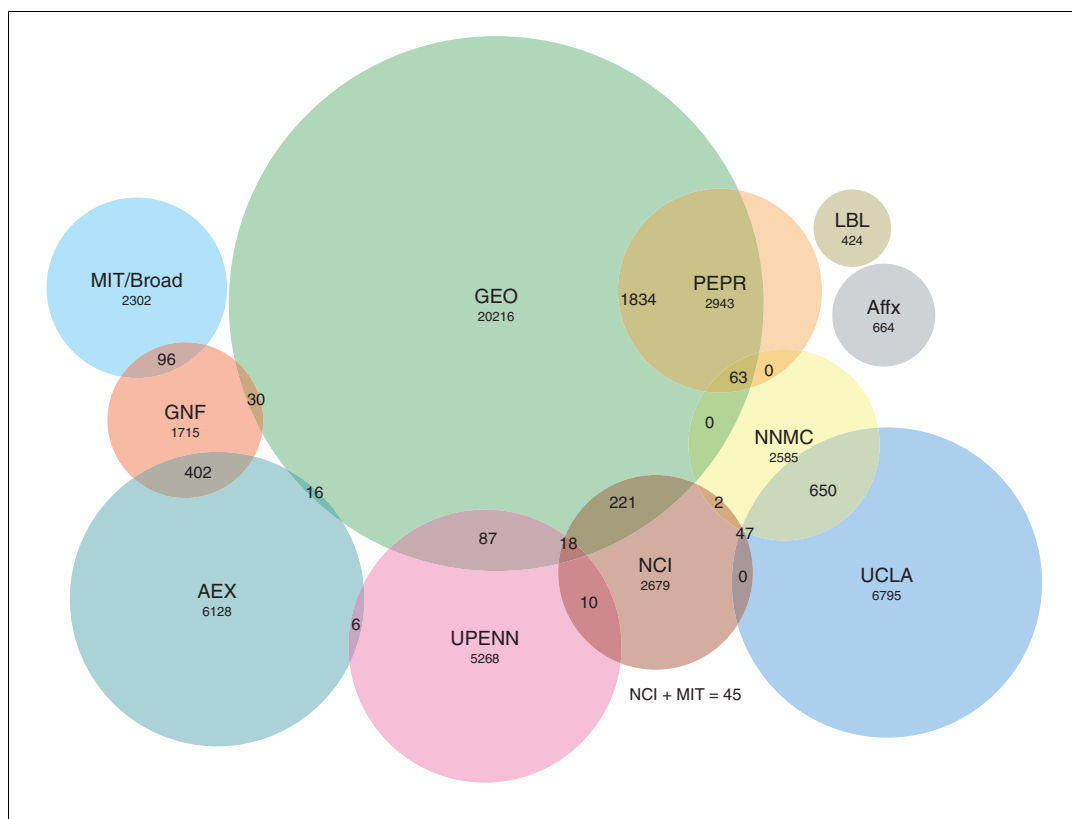


Figure 1

Summary of data sources present in Celsius. Data have been imported from several sources, 11 of which are shown. Numerals indicate the number of files within each source. Circle overlap is proportional to CEL overlap between data sources. AEX, EBI ArrayExpress [49]; AFFX, Affymetrix [50]; GEO, NCI Gene Expression Omnibus [51]; GNF, Genomics Institute of the Novartis Research Foundation [52]; LBL, Lawrence Livermore National Laboratory; MIT, Broad Institute [53]; NNMC, NIH Neuroscience Microarray Consortium [54]; PEPR, Public Expression Profiling Resource [55]; UCLA, University of California, Los Angeles DNA Microarray Core Facility [56]; UPENN, University of Pennsylvania Microarray Core Facility [57].

quantification events. We chose this method based on our observation that a quantification pool of this size is sufficient for all algorithms to estimate a signal stably, provided the pool was created from a heterogeneous mixture of samples. Corroborating findings using a similar approach were recently reported [13]. This 'quantification pool' technique allows Celsius to grow the dataset incrementally while ensuring that quantified values from each CEL are compatible for analysis with values from all other CELs. The code for managing CEL quantification is modular so that as new algorithms become available for processing microarray data extensions to the Celsius quantification pipeline can be readily implemented.

Data access

All contents of Celsius may be accessed through use of the Celsius software library written in the R statistical programming language. It may be downloaded as the Celsius from the Comprehensive R Archive Network [14]. The Celsius library provides an application programmer interface (API) to a subset of the Celsius web services for both reading and write data, and is designed for seamless operation with components of the Bioconductor project [15,16]. Specific instructions for how to obtain, install, and use this library are provided at the Celsius project homepage [17]. We chose to focus on opening public access to Celsius through a programmatic API written in R because it is the *de facto* standard environment for the analysis of microarray data.

Table 1**Most common Affymetrix array designs represented in Celsius**

Platform	Number of CELs	Percentage of CELs	Organism
HG-U133A	11,296	19%	Human
HG-U133 Plus 2	5,954	10%	Human
HG U95Av2	4,600	8%	Human
ATH1-121501	3,952	6%	Plant
MG U74Av2	3,840	6%	Mouse
Mouse430 2	3,580	5%	Mouse
MOE430A	3,154	5%	Mouse
HG-U133B	2,522	4%	Human
RG U34A	2,258	4%	Rat
RAE230A	1,247	2%	Rat
Total	42,403	70%	

Experimental metadata and community participation

The Celsius policy of microarray data sharing is liberal and inclusive when contrasted with the status quo. Rather than adhering to the Minimal Information About a Microarray Experiment [18] guidelines recommendation on data deposition (namely, that metadata for an experiment be provided concomitantly with primary data submission to a repository), we instead adopt the successful data sharing model of the International Nucleotide Sequence Database Collaboration (INSDC) [19]. In the INSDC model, primary data can be contributed to a public repository with or without metadata.

In contrast to other web-accessible microarray resources, additional metadata can be subsequently provided to the repository by anyone, not only by the contributor of the primary data. Celsius places strong emphasis on community participation. This is most evident in the system's ability to accept community contributions in the form of primary data as well as metadata. Indeed, public users of the Celsius system are able to upload, either anonymously or with attribution, primary data in the form of CEL files. These are processed as all other CELs in the system; they are archived, quantified, and then made publicly visible and annotatable.

Users may annotate all SNIDs and probeset records present in Celsius, either through the use of ontology terms approved by the National Center for Biomedical Ontology (NCBO) [20], such as the Gene Ontology (GO) and Mouse Anatomy Ontology [21,22], or by using free-text 'tags'. These activities are possible through programmatic web service APIs (described below under Web services). Likewise, records for CELs and probesets may be retrieved from Celsius using ontology identifiers. We created these interfaces to allow the community to import and export data and metadata as easily as possible and in a distributed manner. Our aim in creating these interfaces is to create a metadata resource with broad coverage that permits analysis over an integrated set of data produced through the efforts of the community.

The community annotation features of Celsius have already been used to manually encode annotation for more than 30% of all HG-U133A CELs (Figure 5a). This number continues to grow as driven by user demand, both inside and outside our group. These CELs are annotated for tissue of origin, cell type of origin, pathologic state, or phenotypic state of the hybridized biologic sample. The current state of annotation for HG-U133A CELs for tissue and neoplastic pathologic state is shown in Figure 5b,c. After careful review of the available descriptions of experiment design and sample treatment, these annotations were manually encoded using controlled vocabularies provided by public ontology efforts [22-26]. The process of encoding annotation with controlled vocabularies is difficult and time consuming, because it frequently requires review of the primary literature to obtain key facts about the biologic samples. Our intention is that the community annotation features will promote distribution of the effort to annotate CEL files gathered into Celsius, both by manual curation and by programmatic extraction of annotation from literature and GEO/AEX annotation deposition.

The past few decades have shown that, in general, this policy of open and minimal participation is beneficial. By allowing primary data deposition even without any metadata, and *vice versa*, a flood of primary data has entered public sequence repositories. This condition fostered the growth of a large and active sequence analysis research community. We believe the INSDC data sharing policy can be successfully applied to data and metadata derived from all high-throughput genome-scale assays. We demonstrate the application of this policy in Celsius.

Web services

To facilitate the sharing of data from Celsius, a series of programmatic interfaces have been developed following the web services model of information exchange. This design is attractive because of its platform neutrality, so researchers can interact with the Celsius services using a wide variety of pro-

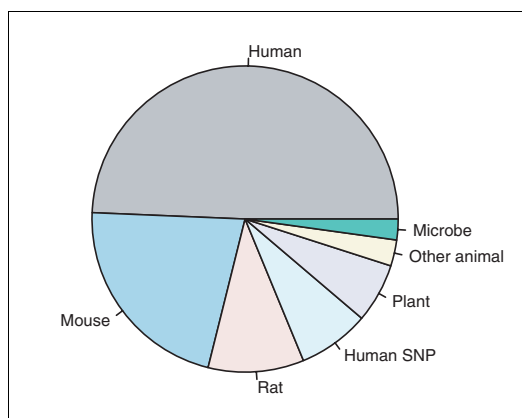


Figure 2
Tally of CELs by organism as of January 2007. SNP, single nucleotide polymorphism.

gramming languages. XML (extensible markup language) and web-based protocols were used to facilitate this ease of access to Celsius for researchers embracing large-scale microarray experimentation and data analysis through bulk data access.

The services available through Celsius provide a wide range of abilities to query, transform, and upload microarray data. For example, Celsius web services provide an identifier transformation service that allows researchers to query the

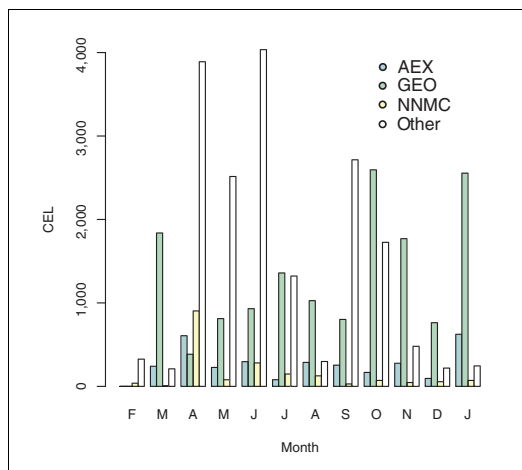


Figure 3
Monthly tally of CEL file import into Celsius from February 2006 to January 2007. AEX, EBI ArrayExpress; GEO, NCBI Gene Expression Omnibus; NNMC, NIH Neuroscience Microarray Consortium.

system with one database accessor and retrieve all others with which a given sample is associated. This allows the mapping of a GEO identifier to an AEX identifier via a SNID identifier intermediate, checking whether a given CEL is present, and, perhaps in the future, automatic import of experimental metadata. Given the distributed nature of microarray datasets, and the possibility of the same dataset being present in multiple repositories, this accession transformation service is an invaluable resource for the microarray community.

Several query web services complement the look-up services and provide a mechanism for searching experimental metadata within Celsius. These services take advantage of the extensive use of ontology annotations throughout Celsius and allow for the rapid identification of all annotated SNIDs of a particular type and level of certainty. For instance, a query for manually curated nervous tissue uses the structure of the controlled vocabulary to return not just SNIDs annotated as nervous tissue, but also all SNIDs annotated with controlled vocabulary terms that are part of the nervous system, such as spinal cord. Other search services include identification of samples based on platform, data retrieval by normalization algorithms, and searching of free-text tags.

Exemplifying the inclusive position of Celsius toward microarray data, we provide a deposition service that can be used either anonymously or with attribution. This function permanently archives CEL data, and all uploaded CELs are assigned a SNID and quantified. They can subsequently be retrieved in SOFT format and submitted to GEO, thus meeting current journal data deposition requirements. This upload service is complemented by a curatorial service that allows Celsius contributors to attach both ontology-based and free text annotations to any sample records.

Other web services are also available from Celsius and more will be added in the future. Up-to-date documentation of what web services are available can be found at the Celsius project homepage [17]. Data from each service are available in both XML and tabular formats so that they may easily be imported into many programming environments.

Annotation examples

The Celsius community features have been used programmatically to annotate a large number of samples and genes. We present two examples to demonstrate the wealth of information that can be gleaned from a data resource of this magnitude. Automated annotation algorithms such as these will become increasingly common in Celsius, similar to the way vector trimming and gene predictions algorithms are routinely run on nucleotide data as they are produced by sequencers.

Assignment of sex annotation to CELs

We assigned each of the 8,915 HG-U133A SNIDs present in Celsius as of October 2006 into male/female classes. This was

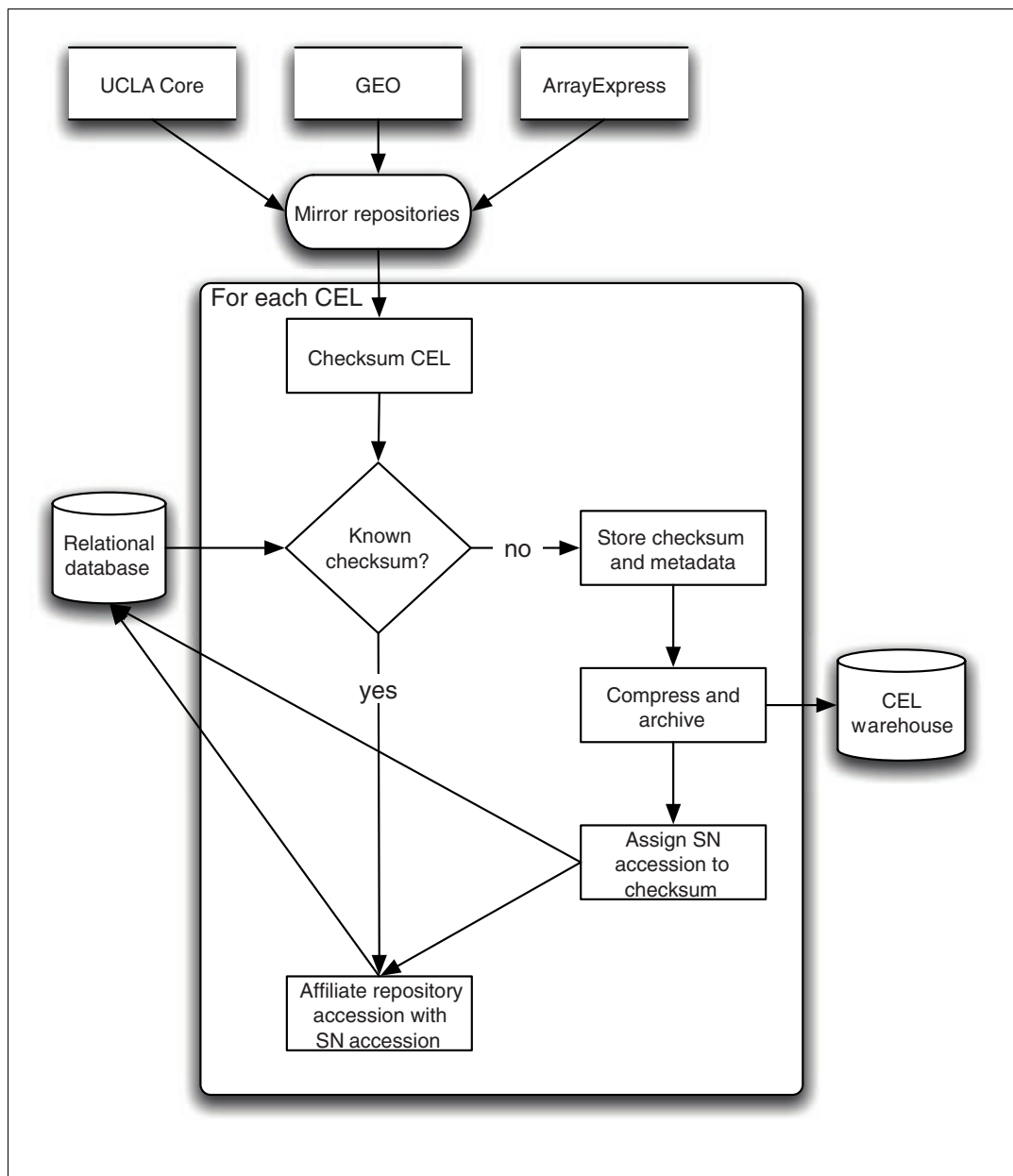


Figure 4
 Process for importing microarray data from other repositories. Potentially novel CELs are checksummed and associated with a Celsius serial number database identifier (SNID) database accession identifier. Metadata from the source repository (sample accession, dataset accession), as well as metadata from the CEL (checksum, array type), are archived to a relational database. If a CEL not currently present in Celsius is detected, a then a SNID is assigned and the CEL is compressed and archived. Quantification is performed and resulting data are stored in a relational database. GEO, NCBI Gene Expression Omnibus; SN, University of California, Los Angeles DNA Microarray Core Facility.

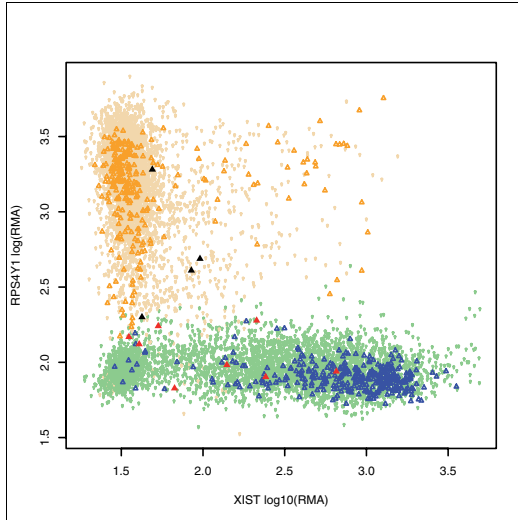


Figure 5

Human HG-U133A CELs are automatically classified for sex of the tissue or cell line of origin. Orange points are manually curated as male and are also correctly classified as male. Red points are manually curated male that are falsely classified as female. Wheat points are classified as male but do not have manually curated results. These three types of points are also denoted by different shapes in the order of triangle, filled triangle, and circle respectively. All points are classified by assigning two clusters in five-dimensional probeset space, two of which are shown. x-axis, 221728_x_at, XIST; y-axis, 201909_at, RPS4Y1.

achieved using the R package mclust's Mclust function with default options to assign points into two clusters. Cluster assignment was based on RMA processed gene expression values of two probesets on the X chromosome and three probesets on the Y chromosome (Figure 6). Of these 8,915 SNIDs, 624 were previously manually assigned to one of the male/female classes using external information, which we regard as accurate. Details for retrieving these data are described below under Materials and methods.

Table 2 shows the fraction of correctly and incorrectly assigned male/female labels by the clustering method. For the female class, the false-positive rate is $8/349 = 0.0229$ and the false-negative rate is $8/279 = 0.0287$. For the male class, these rates are $4/275 = 0.0145$ and $4/345 = 0.0116$, respectively. The rand index and rand index corrected for agreement by chance [27,28] for Table 2 are 0.9622 and 0.9244, respectively. The male class has a lower false-positive and a higher false-negative rate. This is probably due to the strong dependence of female classification on XIST expression, which is typically high in all female derived cell lines and tissues but can be down regulated in some disease states. Overall, the classification method based on gene expression data works very well in assigning sex class, and it enables large-scale analyses based on sex that were not previously possible

using the manually encoded annotations, even though a small error rate is added.

Assignment of Gene Ontology biological process annotation to probesets

Genes with similar expression patterns are thought to be more likely to be functionally associated [29]. They may form structural complexes, participate in the same biochemical pathway, or be regulated by a common transcriptional mechanism. Gene co-expression networks are constructed on the basis of microarray data from the transcriptional response of cells to changing conditions [30,31]. In these networks a node corresponds to an individual probeset-based measurement of a given gene. We constructed such a network of 3,600 probesets with the greatest coefficients of variation measured across 1078 HG-U133A SNIDs that were annotated as pathologically normal using previously described methods [31,32]. We identified 35 modules within this network that correspond to well separated branches of the resulting hierarchical clustering tree. They are visualized as blocks along the diagonal of the topologic overlap matrix (TOM), as shown in Figure 7. The TOM measure uses the neighbor information instead of just their direct connection strength (adjacency) and is thus a robust measure of interconnectedness. This is similar to a gene cluster. More details about the topologic overlap measure, along with a tutorial using freely available R software to construct gene co-expression networks and to identify modules, can be found in the Materials and methods section, below. The parameters and other settings specifically used in this application are listed there for readers to replicate this analysis.

Modules in Figure 7 are color coded by the most significantly enriched GO biologic process (BP) [21] as computed with EASE [33]. Modules that did not have any significantly enriched BP (Bonferroni P value > 0.05) were not considered for further analysis ($n = 5$). Many of the remaining 30 modules shared common BP and were merged, leaving 15 distinct annotation groups. All 15 of these groups are shown in Figure 7. The green group of probesets ($n = 282$) is enriched for probesets involved in muscle contraction ($P = 2.33 \times 10^{-42}$). Of the 188 probesets in this group that are annotated for any BP, 59 (31%) were previously known to be involved in muscle contraction, which correspond to 68% of the 87 probesets contained in the analyzed population of 3,600 probesets that are associated with muscle contraction. The green group corresponds to a bright block along the diagonal of the TOM plot. It indicates that the probesets within this group have high topologic overlap measures as well as highly correlated expression profiles.

Use of the primary BP assigned by EASE to annotate uncharacterized or partially characterized probesets warrants further exploration, but these data cannot be used for classification using conventional methods.

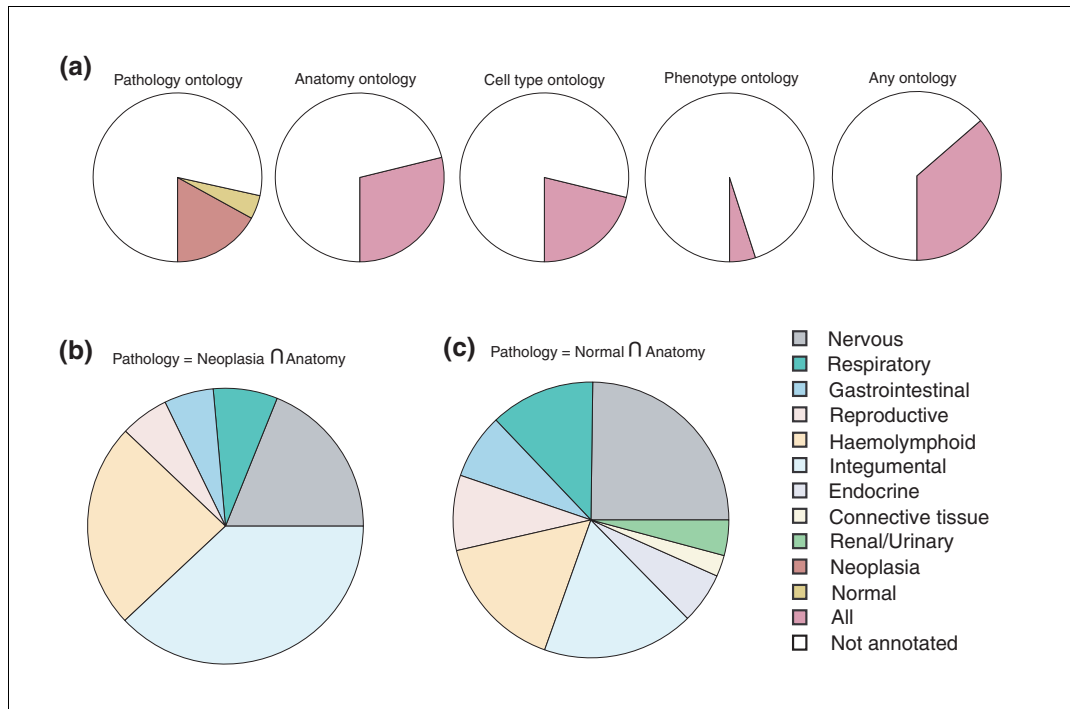


Figure 6
Annotation coverage and depth for the Human HG-UI33 platforms. **(a)** Filled wedges indicate the fraction of CELs for which annotation is present. The red and yellow wedges of the left-most pie indicate fraction of diseased and normal samples, respectively. The right-most pie's wedge indicates the fraction of CELs for any annotation from the preceding columns have been given (excluding sex). **(b)** Human HG-UI33A samples grouped by tumor type and normal. Annotation was manually assigned after literature review. Many integumental system tumors are breast tumors. **(c)** Human HG-UI33A samples grouped by tissue of origin. Annotation was manually assigned after literature review.

This is because gene annotation is incomplete, and it is therefore not possible to estimate the false-positive rate in assigning an annotation to a probeset. However, the 223 probesets within the green module not known to participate in muscle-specific processes are highly correlated with probesets known to be involved in these processes, and therefore they may play a role in muscle tissue. Numerous other gene-gene correlations provide additional information about the specific expression of genes within specific tissue types.

Conclusion

Celsius is a substantial data resource that contains more primary and derivative microarray measurements than all public repositories combined. Celsius was assembled and continues to grow by means of permissively importing Affymetrix CEL files and assigning SNID database accession identifiers to them. Initially, data from 11 independent institutions were imported. Celsius continues to add more institutions to this list and imports data from all available

institutions on a weekly basis. Imported data are processed using best-of-breed signal estimation algorithms. Metadata are acquired and associated with SNIDs through both manual and automated curatorial processes. Access to all contained data and metadata are provided to the public through easy-to-use programmatic interfaces, along with online documentation and prefabricated software libraries for data extraction. Celsius is a useful amalgamation of primary data for development and testing of quantification algorithms, individual probe level analyses, and identification of gene-gene relationships and gene networks; furthermore, it provides reference material for ongoing work using these array platforms. We encourage further enhancement of this dataset by the community through its programmatic and manual interfaces for upload of primary data and metadata. We continue to stimulate the growth of an active and collaborative environment for the development of gene expression inference algorithms, similar to that created by the establishment of large nucleic and peptide sequence databases. By assembling, redistributing, and creating mechanisms for large-scale com-

Table 2

Assignment of sex-annotated HG-UI33A SNIDs by clustering

	Curated female	Curated male	Total
Classified female	341	8	349
Classified male	4	271	275
Total	345	279	624

SNID, serial number database identifier.

munity involvement in working with these data, a turning point will be reached such that high-throughput genomic data will be reused, mined, and analyzed to its full potential.

Materials and methods

Data import

Initially, all public CELs identifiable at AEX and GEO were copied to UCLA by FTP mirror. This was performed in January 2006 in bulk. Subsequently, additional data sources were added as institutions have been willing to permit upload into Celsius. Since the initial data upload in bulk, additional data from these sources are automatically mirrored on a weekly basis (Figure 4). The data import process begins by creating a local mirror for each of the sites from which data will be imported. The contents of these repository mirrors are then scanned for all CELs present. For each CEL, an MD5 checksum is calculated and associated with the CEL's accession from the remote data source (for instance, a CEL from GEO is associated with a GSM [GEO sample] accession). Most new data from this mirroring process derive from AEX, GEO, the UCLA DNA Microarray Facility, and NNMC. When Celsius detects a CEL that has not previously been imported, file-level metadata (file format, platform, and checksum) are extracted, a permanent SN database accession identifier is assigned, and the file is compressed and permanently stored on an archival file system. The assigned SN database accession identifier (SNID) can be used by both internal and external applications to refer to that CEL's data in Celsius. Unique CELs are identified using a checksum algorithm. This is necessary because a single CEL may exist in multiple repositories but under a different name at each repository. Finally, each unique CEL is processed on a 16-node cluster of computers administered using Sun Grid Engine. We use several common quantification algorithms, namely dChip, gcRMA, RMA, PLIER, MAS5, and VSN [8-11]. These algorithms are available from the Bioconductor suite of bioinformatics utilities [34]. Quantified expression values for each probeset from each CEL are produced by processing it along with a platform-specific pool of 50 other CELs, where the pool is held constant for each CEL that is processed. A similar procedure was recently described and validated [13].

Programmatic access

All data in Celsius can be accessed using the Celsius R library. The package itself and instructions on its use are available at

the Celsius project homepage [17]. This service utilizes an extension to the DAS2 (Distributed Annotation System 2.0) protocol [35,36], which allows assay data to be provided as hyperlinked Microarray Gene Expression Markup Language (MAGE-ML) fragments. In addition to the DAS2 service, other Celsius-specific services also documented at the homepage [17] are also available. Notable among these are the following: an identifier transformation service for mapping external database accession identifiers such as the GEO GSM sample identifier to and from Celsius SNIDs; a matrix label generation service, which can be used to create textual descriptions suitable as sample descriptors, for instance as row/column labels in a heatmap; a curatorial service that enables users to contribute to Celsius through attaching ontology and free-text metadata to existing CEL and probeset records; and a CEL deposition service, which allows primary data to be deposited anonymously and can be subsequently extracted in SOFT format suitable for upload to GEO.

Open source libraries for interacting with the Celsius web services have been written in the R and Java computer programming languages. These libraries are available from the Comprehensive R Archive Network [14] and Genoviz web-sites [14,37,38].

Data representation

At its core, Celsius is a relational data warehouse based on PostgreSQL [39] and is designed for online analytical processing. Given the scale of data to be stored, the ability to respond to user queries in minimal time has been a major design goal in all aspects of system design. Celsius is implemented using the Chado database schema, which is a component of the Generic Model Organism Database Project [40]. The MAGE module of the Chado schema that is pertinent to the representation and storage of microarray data is presented in Additional data file 1. The schema has been optimized to accommodate several classes of user requests: to retrieve signal estimates calculated with algorithm A for all probesets on quantified CEL Q; to retrieve signal estimates calculated with algorithm A for all CELs on probeset P; to calculate distance from signature p to all samples using metric D and signal estimates calculated with algorithm A; to calculate distance from signature q to all probesets using metric D and signal estimates calculated with algorithm A; to retrieve all annotations on CEL Q; to retrieve all CELs annotated at or below ontology term T; and to annotate CEL Q with term T.

If implemented as a single table, it is impossible to simultaneously minimize query time for both cases 1 and 2. Minimization is achieved through clustering, or physically ordering disk blocks by an index, and it is not possible to have more than one physical ordering for a table. Thus, minimizing query time for case 1 necessarily increases query time for case 2. Optimization of query time for cases 3 and 4 presents the same problem because these cases are dependent upon cases 1 and 2. We overcame this obstacle by storing identical data in two tables and clustering each table on a different index. By using this technique, the size of the tables containing signal estimates is doubled. However, the advantage is that the retrieval time is reduced by several orders of magnitude by lessening hard disk activity. We have also chosen to partition the table that holds all signal estimates based on quantification algorithm, denoted A. This optimization reduces the number of rows in each table and results in a proportional query time of $\log_2(n/N)$, where n is the number of CELs processed with algorithm A and N is the number of CELs multiplied by the number of quantification algorithms used.

Typical functions to be performed on Celsius involve matrix manipulations using the R programming language. Although it is possible to perform these calculations through a call to an external instance of R, it is inefficient. We make use of PL/R [41], a procedural extension to the PostgreSQL database that allows an embedded R instance to run inside the PostgreSQL environment. This technique allows R functions to be called as part of a standard structured query language (SQL) query. For instance, the calculation of correlation coefficients for all pairs of probesets from a particular array design can be performed. This method is used to infer gene interaction networks [30] from gene expression data, which otherwise is very costly to perform.

Cases 4 through 7 can be accommodated for a single user's ontology-based annotations by using the stock Chado schema for representing CELs, their biologic source materials, and associated annotations. We extended the schema to support storage of annotations from multiple users through the creation of a user module that associates all annotations with a particular user. We also added the ability to both attach and search free-text 'tags' using PostgreSQL's Tsearch2 extension [42].

Sex annotation

We assigned each of the 8915 HG-U133A SNIDs present in Celsius as of October 2006 into male/female classes using the R mclust package's Mclust function with default options to assign points into two clusters. Cluster assignment was based on RMA processed gene expression values of five probesets: 214218_s_at and 221728_x_at on chromosome X and 201909_at, 206769_at and 205000_at on chromosome Y. Of these 8,915 SNIDs, 624 SNIDs were previously manually assigned to one of the male/female classes using external information, which we regard as accurate. All HG-U133A samples may be retrieved from Celsius by using the Celsius R library referenced in the section Programmatic access (above) and retrieving all HG-U133A measurements for these five probesets.

Gene coexpression network construction

Using previously described methods [31], we calculated the Pearson correlation matrix for the gene expression profiles of the 3,600 probesets across 1078 HG-U133A SNIDs annotated as pathologically normal. To reproduce these results, these data may be retrieved from Celsius using the following R commands after installing the Celsius R library referenced in the section Programmatic access and searching for HG-U133A samples matching the 'normal' (MPATH:458) ontology term.

We raised all elements in the matrix to the power 6 to create the adjacency matrix. The adjacency matrix is equivalent to an undirected weighted network. Using the topologic overlap measure [32], we calculated a TOM from the adjacency matrix. The topologic overlap measure between two nodes is approximately the proportion of the number of shared neighbors divided by the total number of neighbors of the node with fewer neighbors. The TOM measure uses the neighbor information instead of just their direct connection strength (adjacency), and is thus a robust measure of interconnectedness. The TOM was converted to a dissimilarity matrix by subtracting all elements from 1 and then used as the input to an average linkage hierarchical clustering function. Modules were identified as well separated branches in the resulting dendrogram. We used dynamic height cut-off 0.995 to cut the clustering tree with minimum module size 40 to reach the 35 proper modules. This module detection approach has led to biologically meaningful modules in several applications [30,32,43-47], but we make no claim that it is optimal.

Figure 7 (see following page)

A gene network constructed from 3600 most varying human probesets. The hierarchical clustering tree and the heat map of the topologic overlap matrix for the 3600 HG-U133A probesets with the largest coefficients of variation measured across 1078 HG-U133A serial number database identifiers (SNIDs) that were annotated as pathologically normal. The color breaks in the colored annotation bar above the heat map mark annotation groups of probesets based on EASE, and tick marks mark the individual modules of highly interconnected probes before being merged into a single annotation group. Colors, left to right are defined as follows: red, transcription; black, response to biotic stimulus; turquoise, ectoderm development; magenta, regulation of metabolism; blue, nervous system development; green, muscle contraction; dark orchid, digestion; chocolate, organic acid metabolism; brown, acute-phase response; dark khaki, complement activation; orange, pregnancy; yellow, sexual reproduction; midnight blue, mitotic cell cycle; deep sky blue, skeletal development; tan, phosphate transport.

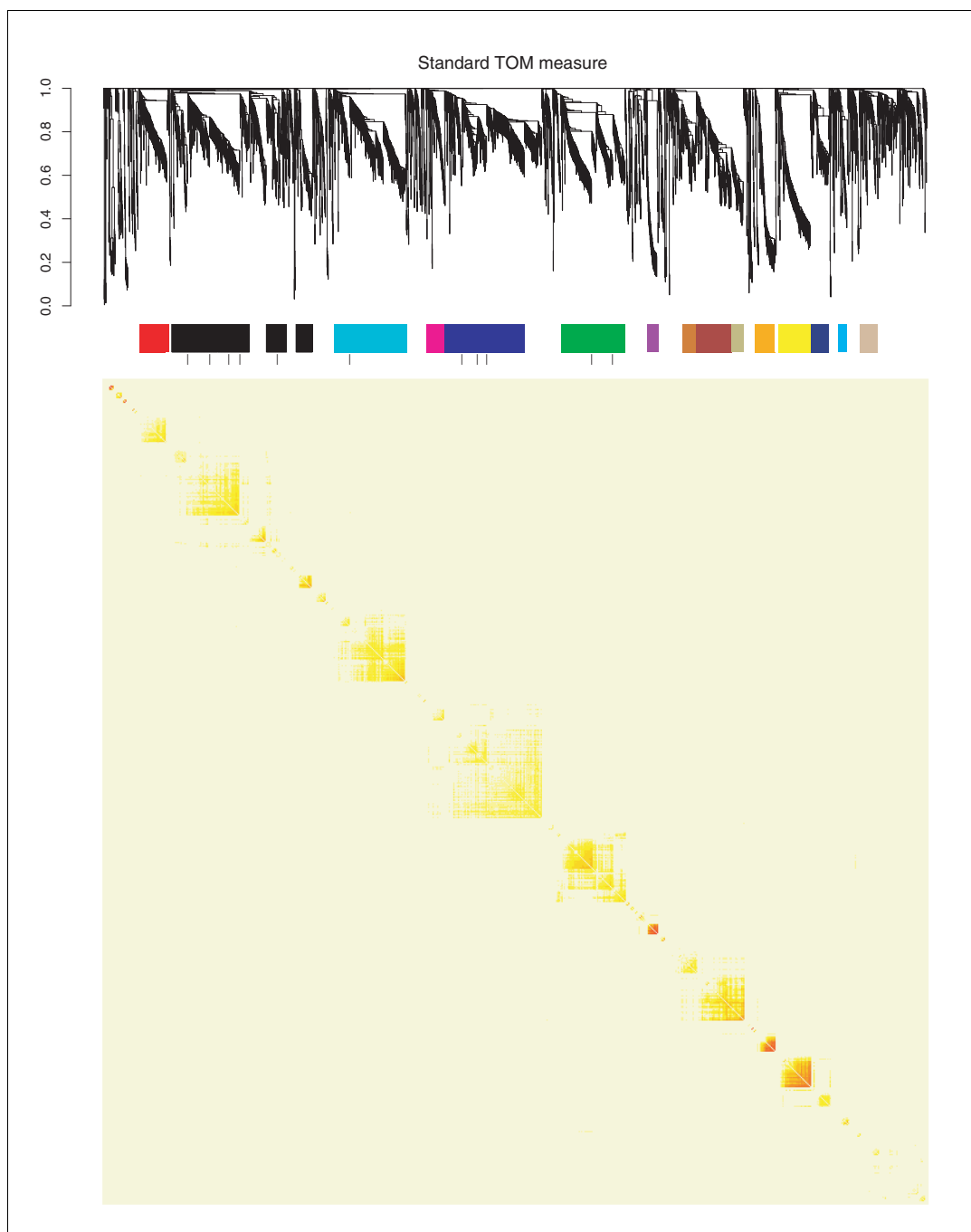


Figure 7 (see legend on previous page)

Tutorials using freely available R software to construct gene co-expression networks and to identify modules are available in the report by Zhang and Horvath [31] and on the internet [48].

Additional data files

The following additional data are available with the online version of this paper. Additional data file 1 provides The MAGE module of the Chado schema that is pertinent to the representation and storage of microarray data in Celsius.

Acknowledgements

We thank Jerome Braunstein for critical comments on the manuscript. The work was supported by grants from the NINDS (U24HS052108) and the NHLBI (HL72367), with support from the NIH Neuroscience Microarray Consortium.

References

- Barrett T, Suzek T, Troup D, Wilhite S, Ngau W, Ledoux P, Rudnev D, Lash A, Fujibuchi W, Edgar R: **NCBI GEO: mining millions of expression profiles-database and tools.** *Nucleic Acids Res* 2005, **33**:D562-D566.
- Sarkans U, Parkinson H, Lara G, Oezcimen A, Sharma A, Abeygunawardena N, Contrino S, Holloway E, Rocca-Serra P, Mukherjee G, et al.: **The ArrayExpress gene expression database: a software engineering and implementation perspective.** *Bioinformatics* 2005, **21**:1495-1501.
- Rhodes D, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan A: **Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression.** *Proc Natl Acad Sci USA* 2004, **101**:9309-9314.
- Stuart J, Segal E, Koller D, Kim S: **A gene-coexpression network for global discovery of conserved genetic modules.** *Science* 2003, **302**:249-255.
- Chen J, Zhao P, Massaro D, Clerch L, Almon R, DuBois D, Jusko W, Hoffman E: **The PEPR GeneChip data warehouse, and implementation of a dynamic time series query tool (SQGT) with graphical interface.** *Nucleic Acids Res* 2004, **32**:D578-D581.
- Spellman P, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, Ball C, Lepage M, et al.: **Design and implementation of microarray gene expression markup language (MAGE-ML).** *Genome Biol* 2002, **3**:research0046.1-research0046.9.
- Whetzel P, Parkinson H, Causton H, Fan L, Fostel J, Frago G, Game L, Heiskanen M, Morrison N, Rocca-Serra P, et al.: **The MGED Ontology: a resource for semantics-based description of microarray experiments.** *Bioinformatics* 2006, **22**:866-873.
- Li C, Wong W: **Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.** *Proc Natl Acad Sci USA* 2001, **98**:31-36.
- Wu Z, Irizarry R: **Preprocessing of oligonucleotide array data.** *Nat Biotechnol* 2004, **22**:656-658. author reply 658
- Irizarry R, Bolstad B, Collin F, Cope L, Hobbs B, Speed T: **Summaries of Affymetrix GeneChip probe level data.** *Nucleic Acids Res* 2003, **31**:e15.
- Huber W, von Heydebreck A, Sltmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression.** *Bioinformatics* 2002:S96-S104.
- Hua J, Craig D, Brun M, Webster J, Zismann V, Tembe W, Joshipura K, Huettelman M, Dougherty E, Stephan D: **SNiPer-HD: improved genotype calling accuracy by an expectation-maximization algorithm for high-density SNP arrays.** *Bioinformatics* 2007, **23**:57-63.
- Katz S, Irizarry R, Lin X, Tripputi M, Porter M: **A summarization approach for Affymetrix GeneChip data using a reference training set from a large, biologically diverse database.** *BMC Bioinformatics* 2006, **7**:464.
- Comprehensive R Archive Network** [http://cran.r-project.org/]
- Bioconductor Project Homepage** [http://www.bioconductor.org/]
- Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al.: **Bioconductor: open source development for computational biology and bioinformatics.** *Genome Biol* 2004, **5**:R80.
- Celsius Project Homepage** [http://genome.ucla.edu/projects/celsius]
- Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball C, Causton H, et al.: **Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.** *Nat Genet* 2001, **29**:365-371.
- International Nucleotide Sequence Database Collaboration Homepage** [http://www.insdc.org/]
- Rubin D, Lewis S, Mungall C, Misra S, Westerfield M, Ashburner M, Sim I, Chute C, Solbrig H, Storey M, et al.: **National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge.** *OMICS* 2006, **10**:185-198.
- Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, et al.: **Gene ontology: tool for the unification of biology.** *Nat Genet* 2000, **25**:25-29.
- Hayamizu T, Mangan M, Corradi J, Kadin J, Ringwald M: **The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data.** *Genome Biol* 2005, **6**:R29.
- Burger A, Davidson D, Baldock R: **Formalization of mouse embryo anatomy.** *Bioinformatics* 2004, **20**:259-267.
- Smith C, Goldsmith C, Eppig J: **The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information.** *Genome Biol* 2005, **6**:R7.
- Schofield P, Bard J, Booth C, Boniver J, Covelli V, Delvenne P, Ellender M, Engstrom W, Goessner W, Gruenberger M, et al.: **Pathbase: a database of mutant mouse pathology.** *Nucleic Acids Res* 2004, **32**:D512-D515.
- Bard J, Rhee S, Ashburner M: **An ontology for cell types.** *Genome Biol* 2005, **6**:R21.
- Cohen J: **A coefficient of agreement for nominal scales.** *Educ Psychol Measure* 1960, **20**:37-46.
- Hubert L, Phipps A: **Comparing partitions.** *J Classification* 1985, **2**:193-218.
- Eisen M, Spellman P, Brown P, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95**:14863-14868.
- Carlson M, Zhang B, Fang Z, Mischel P, Horvath S, Nelson S: **Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks.** *BMC Genomics* 2006, **7**:40.
- Zhang B, Horvath S: **A general framework for weighted gene co-expression network analysis.** *Stat Appl Genet Mol Biol* 2005, **4**: Article17
- Ravasz E, Somera A, Mongru D, Oltvai Z, Barabasi A: **Hierarchical organization of modularity in metabolic networks.** *Science* 2002, **297**:1551-1555.
- Hosack D, Dennis G, Sherman B, Lane H, Lempicki R: **Identifying biological themes within lists of genes with EASE.** *Genome Biol* 2003, **4**:R70.
- Gentleman R, Carey V, Bates D, 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**:R80.
- Dowell R, Jokerst R, Day A, Eddy S, Stein L: **The distributed annotation system.** *BMC Bioinformatics* 2001, **2**:7.
- Distributed Annotation System Protocol Specification, Version 2** [http://biodas.org/documents/das2/das2assay.html]
- Genoviz Project Homepage** [http://www.sourceforge.net/projects/genoviz/]
- Lorraine A, Helt G: **Visualization techniques for genomic data.** *Proc IEEE Comput Soc Bioinform Conf* 2002, **1**:321-326.
- PostgreSQL Project Homepage** [http://www.postgresql.org/]
- GMOD Project Homepage** [http://www.gmod.org/]
- PL/R Project Homepage** [http://www.joeconway.com/plr/]
- Tsearch2 Project Homepage** [http://www.sai.msu.su/megera/postgres/gist/tsearch/V2/]
- Ye Y, Godzik A: **Comparative analysis of protein domain organization.** *Genome Res* 2004, **14**:343-353.
- Horvath S, Zhang B, Carlson M, Lu K, Zhu S, Felciano R, Laurance M,

- Zhao W, Qi S, Chen Z, et al.: **Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target.** *Proc Natl Acad Sci USA* 2006, **103**:17402-17407.
45. Ghazalpour A, Doss S, Zhang B, Wang S, Plaisier C, Castellanos R, Brozell A, Schadt E, Drake T, Lusis A, Horvath S: **Integrating genetic and network analysis to characterize genes related to mouse weight.** *PLoS Genet* 2006, **2**:e130.
 46. Gargalovic P, Imura M, Zhang B, Gharavi N, Clark M, Pagnon J, Yang W, He A, Truong A, Patel S, Nelson S, Horvath S, Berliner J, Kirchgessner T, Lusis A: **Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids.** *Proc Natl Acad Sci USA* 2006, **103**:12741-12746.
 47. Oldham M, Horvath S, Geschwind D: **Conservation and evolution of gene coexpression networks in human and chimpanzee brains.** *Proc Natl Acad Sci USA* 2006, **103**:17973-17978.
 48. **Weighted Gene Co-Expression Networks** [<http://www.genetics.ucla.edu/labs/horvath/GeneralFramework/>]
 49. **ArrayExpress Homepage** [<http://www.ebi.ac.uk/arrayexpress/>]
 50. **Affymetrix Homepage** [<http://www.Affymetrix.com/>]
 51. **NCBI Gene Expression Omnibus** [<http://ncbi.nlm.nih.gov/geo/>]
 52. **Genomics Institute of the Novartis Research Foundation** [<http://www.gnf.org/>]
 53. **Broad Institute** [<http://www.broad.mit.edu/>]
 54. **NIH Neuroscience Microarray Consortium** [<http://arrayconsortium.tgen.org/>]
 55. **Public Expression Profiling Resource Microarray Center** [<http://pepr.cnmcresearch.org/>]
 56. **UCLA DNA Microarray Core** [<http://microarray.genetics.ucla.edu/>]
 57. **Penn Microarray Facility** [<http://www.med.upenn.edu/microarr/>]

CHAPTER 3

Gene Characterization Throught Large-Scale Co-expression Analysis

.png —

\Z

.png —

\Z

.png —

\Z

.png —

\Z

.png —

\Z

.png —

✓

.png —

\Z

.png —

✓

.png —

\Z

CHAPTER 4

Biopackages.net: Bioinformatics Libraries, Applications, and Data as Operating System Packages

4.1 Abstract

Background: As biological science becomes more quantitative, it is increasingly dependent on large and complex computational infrastructure. Large-scale systems exist upon which analyses can be done, but efforts are often foiled by the analytical software itself, which is frequently of pre-production quality, poorly documented, and difficult to locate, install and configure.

Results: We have created Biopackages.net: a repository of libraries, applications, and data sets that are commonly used in computational biology. This resource solves the problem of software installation by providing pre-configured packages for many of the data and software most commonly used in the biological sciences. We have also created a “build farm” – a software application that allows software packages in the repository to be automatically adapted to new operating systems and hardware architectures.

Conclusions: Using software available from the Biopackages.net repository in conjunction with a mature cluster management product lowers the barrier to entry for addressing large-scale problems in computational biology and bioinformatics. It streamlines the initial setup and maintenance of complex software collections allowing scientists to focus more time on research and less time on configuration. More information is available at <http://biopackages.net>.

4.2 Introduction

As biological science becomes more quantitative, it is increasingly dependent on large and complex computational infrastructures. Scalable and modular clustered computer

systems are now essential for analysis and hypothesis generation from current high-throughput assay technologies. While advances in the scientific computing community have led to the creation of systems that use clustered computer systems to increase productivity, e.g. the Rocks environment [?]. These efforts seldom address another part of the problem, namely *deploying* research-grade software on these systems. This latter problem is not trivial, and the associated productivity loss is a major obstacle in the advancement of computational life science.

The degree to which software increases productivity can generally be evaluated along three axes: utility, accessibility, and usability [?]. Applications and software libraries produced in life science research environments are of great utility. They provide means to represent, store, retrieve, and manipulate biological data. However, even the best of these applications and libraries typically have low marks for accessibility and usability.

Research code is often not readily accessible because it is of prototype quality, with poorly documented interface, installation and configuration mechanisms, and comes with no formal support or guarantee that it will compile, execute, or be otherwise usable on the platforms and operating system combinations of end users. Usability suffers from poor documentation, as well as from inconsistent and often rough-cut user interfaces. Ease of use has also been lacking in the research code environment. Software installation has historically been a painful process that involves README files, configuration settings, and other processes requiring active human interaction. This condition frequently persists even within popular and widely used scientific software, likely due to a deficiency in knowledge of and adherence to software “best practices” as previously posited [?].

The primary aim of the Biopackages.net project is to increase the productivity of life science informatics environments through software packaging. By compiling, test-

ing, and configuring software, we are able to greatly increase the accessibility and usability of it. Another benefit of packaging is that it is possible to compile and configure complementary packages to be aware of and interact with one another and the underlying OS. This creates an environment with greater utility than the sum of its parts.

4.3 Results

The Biopackages.net software repository is available at <http://www.biopackages.net/>. At the time of writing more than 3800 software packages are available for several versions of the Fedora Core and CentOS Linux operating systems on i386 and x86_64 architectures, as well as a shared set of architecture-independent (noarch) packages. All packages currently provided are prepared using the Biopackages.net build farm which in turn employs the RPM Package Manager [?], a suite of utilities for installing and maintaining software on a computer system.

4.3.1 Provided Applications and Libraries

Biopackages.net aims to provide a software repository as broad as possible within the discipline of open-source bioinformatics. Research interests of the contributing authors that are outside the scope of this manuscript include complex disease gene mapping and gene expression analysis. As such, there is a trend for the content of the repository to reflect these topic areas. Even so, we have attempted to include many of the commonly used bioinformatics data sets, libraries, and applications. Data sets provided include: ontologies provided by the National Center for Biomedical Ontology, standard Affymetrix GeneChip data sets, and recent genome assemblies for several commonly studied organisms pre-formatted for use with the sequence analysis

programs BLAST, ePCR, and BLAT [?, ?, ?, ?]. For libraries, the popular Bioconductor and Bioperl [?, ?] libraries are available, as well as the NCBI Toolkit libraries supporting the BLAST application already mentioned. Other applications available include the Generic Genome Browser, the BLAT server, Textpresso, EMBOSS, and Turnkey/GMODWeb [?, ?, ?, ?, ?].

4.3.2 Using the Biopackages.net repository

Software provided by the Biopackages.net servers can be browsed and downloaded directly using a web browser, then installed on the target system with the `rpm` command-line utility. Using the RPM Package Manager reduces the amount of human interaction to two steps in most cases: users can simply retrieve the RPM file and then install with a single command. In addition to ease of installation, RPM also provides a wide array of features that make it superior to many other software packaging formats [?].

While this method of installing software is possible we also support – and recommend – using automated software utilities for package dependency resolution and installation, such the Yellow dog Update Manager (Modified), or `yum` [?]. `yum` allows users to install desired packages in an automated fashion with a single command. The Biopackages.net website provides instructions and a single RPM that can be downloaded and installed to make the user’s installed version of `yum` configured to be aware of how to obtain and install packages from the Biopackages.net repository. `yum` is well suited for bioinformatics research because many of our applications and libraries are considered high-level from the point of view of the OS and have a deep and broad graph of software dependencies. For example, consider the partial software dependency graph shown in Figure 4.1. Installation of the `das2-server` package requires the `chado-Hsa` package, which in turn requires two other packages to be installed, and so on. Installation of these high-level packages frequently requires the

manual retrieval and installation of possibly hundreds of other packages in the dependency graph [?].

4.3.3 Build Farm

Software packages for the biopackages project are built using a software system spread across multiple machines. In the system a single machine acts as the master control node (MCN) and orchestrates the process of building a given package. Individual machines act as build nodes where a package and its dependencies are built. The build farm's MCN manages the building of packages and their dependencies by assigning the jobs to particular build nodes. On each build node the `resolve_deps.pl` script manages the process of recursively building a given package and all its dependencies. This "build farm" allows the same package to be built on multiple operating systems and architectures in a parallel fashion. This flexible framework allows new platforms and operating systems to easily be added. This build farm approach is flexible and requires little human interaction to build all packages for a particular platform and operating system. In this way, biopackages is more than just a collection of software and data packages. It provides a complete framework for adding and building new software packages across a variety of platforms and operating systems with minimal effort.

4.4 Discussion

The Biopackages.net repository is a large collection of software developed that is relevant to – and in many cases specifically developed for – information management and computational analysis in the biological sciences. Software is packaged using the RPM Package Manager and distributed using the Yellow dog Updater (Modified), both of which are commonly used software management utilities for Linux-based comput-

ers. At the time of this writing the repository contains more than 3800 packages for Fedora Core Linux, Centos Linux, and Apple's OS X Darwin 10.4.

A large fraction of the package growth rate can be accounted for by (a) the recompilation of packages for all platforms, and (b) minor adjustments to package contents and metadata, analogous to debugging and documentation in the software development process. While this type of growth is necessary for the Biopackages.net repository to remain usable on a variety of platforms, it is less interesting and less desirable than repository growth to include new software from the research community, as well as new versions of existing software.

Substantial effort must be invested to maintain the repository in a useful and up-to-date state, but it is an overall win for the bioinformatics community if users buy in to the system, as it allows bioinformatics-specific system updates to be treated as a service, effectively factored out by delegation to a group of specialists [?, ?]. The amount of effort necessary to identify and incorporate new software from the community may also be partially mitigated if users of the Biopackages.net repository reaches a critical mass. In this scenario, authors of software are able to attain more prominence and citation rank by ensuring their software is readily available to the largest number of users.

4.4.1 Licensing

Academic software is commonly available through a dual-licensing model in which academic or not-for-profit research use is freely granted, but enterprise-class users are required to pay for use. The licensing terms are typically prominently displayed on the authors' software download pages, and are difficult to miss. One potentially adverse effect of using the Biopackages.net repository through `yum` is that hundreds of license terms can be implicitly agreed to with only a few keystrokes. While this

is typically not a problem for academic users, it may create legal difficulties for users in the enterprise. We encourage all users to closely examine all package metadata to ensure they are entitled to use the software under the terms in which it is distributed via the Biopackages.net repository. One possible mechanism to ease the burden on end users is to partition the repository's software based on license permissiveness, and is a future goal of the Biopackages.net project. Future goals also include contributing to larger and more general-purpose repositories, such as <http://pbone.net/> and <http://atrpms.net>.

4.4.2 Non-RPM Systems

We are also interested in extending the scope of the project beyond operating systems that use the RPM Package Manager natively. In our preliminary work in this area, we focused on Apple MacOS X 10.4 and Cygwin for Microsoft Windows XP, and used the RPM and `yum` ports available on these systems. While it was possible to use RPM and `yum` to manage installed software on these systems, it was neither trivial nor elegant. The essence of the problem lies in the fact that at their lower levels, software dependency graphs rely on basic software, such as the Bourne shell, or the Apache HTTP server. These basic packages are available and usable by RPM-installed software on non-RPM-native systems, but the RPM-based utilities cannot verify dependency integrity. Our solution to this problem is to provide virtual packages that claim to provide software (such as Apache), but actually do nothing and defer to the base system packages to provide the functionality.

There is also potential to convert Biopackages RPMs into non-RPM formats for use on a wide breadth of UNIX types. While RPM support has been ported to most versions of UNIX and Linux, many of these OSes were initially developed with other binaries in mind. Distributions based on Debian Linux, for instance, are native to `deb`

packages while Sun's Solaris and many other UNIXes are native to pkg. Slackware Linux uses .tgz packages, while Stampede Linux uses .slp packages. Fortunately for maintainers tools has been developed to convert packages between these many formats. One such software is Alien which allows conversion between the above named package types. This creates the prospect of developing Biopackages branches for different package formats to attract users for whom RPM is not a viable option.

4.5 Methods

4.5.1 Package Creation

The first step in creating a software package is to obtain the package's source material, or sources, from an upstream provider. For a software library, this may be a tarball of C++ files, and for a genomic data set may be several FastA files.

The second *optional* step is to create patches that modify the package sources to be compatible with the target platform for installation. Modifications typically do not change the core functionality of the software, but customize the software or software installer slightly to be compatible with the target platform or related software libraries and applications.

The third step is the creation of a metadata file, which for RPM based systems is called a spec file. The spec file references the source and patch files and includes instructions for performing the configuration and installation of the software. Additionally, the spec file lists all other packages that are required in the build- and install-phase. The packages depended upon may differ for build and install phases, and may also be conditional by platform . The spec file also contains other package-related metadata, such as license details and type of package (library, database, application, etc).

4.5.2 Package Version Control and Platform Targeting

After being initially written, the spec file is then converted to a spec.in file and imported into the Biopackages.net Concurrent Versions System (CVS) repository. The Biopackages.net project abstracts the spec file metadata one step further for two reasons.

First, the spec file contains a revision metadata field for package version information, but provides no mechanism for automatic management of version numbers. Likewise, it contains a changelog metadata section for describing how a package has evolved over time, but does not provide a utility for tracking changes. CVS does both automatic revision and changelog tracking, but in a different format than is required by RPM. A simple search-and-replace preprocessing step is sufficient to map the CVS sections for versioning and changelogs (`Id` and `Log`) to the corresponding RPM spec file sections (`%revision` and `%changelog`). The Biopackages.net build system includes a utility to perform this task.

Second, spec file syntax does not allow all portions of the file to be conditional based on aspects of the platform. With the use of a single spec file for multiple platforms, it is important to optionally specify different build and installation requirements for the same package on different Linux distributions. This is because of the variation of libraries included with a given package across distributions. As an example, a default Perl on Fedora Core 2, may include a Perl library that is not included in the default Perl installation of Fedora Core 5. Therefore it is important to specify this library as an additional requirement for Fedora Core 5, while unnecessary for Fedora Core 2. RPM is limited in its ability to include conditional statements in the build and installation requirements sections of the spec file. Therefore the solution we have implemented is the inclusion of a platform specific `if` control structure in the spec.in file. This is translated from a conditional statement to the resulting package dependency when the

spec.in file is converted to a spec file.

4.5.3 Package Compilation

Package compilations are performed on a self-contained Build farm that contains all of the OS distributions we target. Using VMWare virtualization on a 64-bit AMD based machine, we are able to simultaneously run multiple 32- and 64-bit x86 architecture virtual machine (VM) instances, each with a different OS and virtual architecture. This allows us to build packages for many different platforms simultaneously.

The RPM construction process uses the Sun Grid Engine (SGE) cluster management utility to orchestrate construction of a given *target* package for multiple platforms, and is illustrated in 4.3. This process is initiated by iterating over all platforms targeted for construction, and interacting with the SGE Master Control Node (MCN) to build those packages. MCN manages all VMs and ensures that each request to build *target* for a particular platform is delegated to the correct VM. Once delegated to a VM, package construction process can be divided into three phases: staging, building, and cleanup. These are illustrated in Figure 4.2.

In the staging phase, *target*'s metadata is converted from the CVS RCS to `rpmbuild` format, and all packages that are required to build *target* are identified and installed. This process is coordinated by the `resolve_deps.pl` script. It begins by parsing *target*'s metadata to identify all packages that are depended upon for build or install of *target*. Each of these dependencies in turn has its dependencies identified, and so on, recursively. When the terminal packages of the dependency graph are reached, the recursive process traverses back toward *target*, and at each step installs the dependency to ensure that all requirements in the graph have been installed.

In the building phase, *target*'s metadata is used to construct the RPM package. This is done with the `rpmbuild` utility, which typically invokes several other scripting

tools and software compilers to prepare the *target* software for the VM's platform.

Finally, cleanup phase takes place. All packages that were installed to satisfy the dependency are uninstalled. This restores the build system to a “clean” state ready for the next scheduled build. The rpm file for *target* is archived to a repository where all complete packages are moved to shared storage and made available to the public via a web server. Log files created during the staging and build phases are also archived to shared storage for post-processing. Examples of post processing include the generation of build reports, identifying VM-specific problems, and detection and re-queueing of failed builds.

After all packages have been iterated over and all requests scheduled with the MCN have completed, the packages in the RPM repository are indexed using the `createrepo` command. This enables `yum` clients to connect to the web server and install packages.

4.5.4 Package Deployment

Built RPMs are automatically written to a shared NFS volume that is accessed by the Biopackages web server. Packages are initially placed into a “testing” repository corresponding to distribution and architecture. After enough testing of package stability, RPMs are migrated to the “stable” repository. On the web server, a nightly cron job automates the creation of `yum` and `apt` headers for all platforms of both the stable and testing biopackages repositories. From this server, packages are served over HTTP and NFS to both local and external clients that make requests.

4.5.5 Package Installation

Once headers are generated on the web server, the user has the choice of several installation methods for a given RPM. The native Red Hat method is to use RPM to perform the installation directly from the web server, or from a local disk the package has been downloaded to. While this allows automated downloading and installation of a package, RPM is limited in its inability to perform dependency resolution and therefore requires manual installation of dependencies. Therefore an easier installation method is through the use of `yum` or `apt`. `yum` allows for easy installation from a software repository by locating and downloading the desired packages and all of its dependencies in one automated transaction. `apt` serves a similar purpose as `yum`, originally being ported from Debian Linux for use with RPMs. Therefore some users are more comfortable with and prefer the syntax and capabilities of `apt`.

Another decision to be made by the user is if they would like to use “testing” RPMs. `yum` and `apt` make it possible for a user to specify which repositories they prefer, and which repositories are only to be used with manual input. This allows a user to disable regular use of the “testing” repository, while allowing manual overriding when a package is present only in the “testing” repository.

4.5.6 Package Quality Assurance

In maintaining a software repository, a crucial issue is adopting a method of ensuring stability and quality of packages and identifying problematic software. A two tiered repository approach helps separate production tested packages from more newly built RPMs. Newly built RPMs are placed into the Testing repository, where they are served to both local and external users for functional testing to be performed. This allows assurance of proper documentation, configuration files, log files, runtime scripts and

other software functionality before a package is deemed “production ready” and migrated to the “stable” repository. It is in this testing period that bugs are ideally discovered and reported by users and developers, to allow fixes to be made before the package is moved to the “stable” repository. Most Testing packages are considered to be mostly stable and are immediately implemented on local compute nodes, however a separate Stable branch adds an extra level of quality control for production environments.

4.6 Figure Captions

4.6.1 Figure 3.1

Packages are represented by boxes, with grey provided by the OS distribution, and white provided by Biopackages.net. Packages may rely on other packages, either at build- or install-time. Solid head dashed lines indicate a n installation dependency, hollow head dashed lines indicate a build dependency, solid head solid lines indicate both an installation and build dependency.

4.6.2 Figure 3.2

A spec.in file is preprocessed to form a spec file, which is processed by the `rpmbuild` utility along with the package source files to form architecture-specific packages and a source package that contains the .spec file and source files.

4.6.3 Figure 3.3

A controlling script observes the state of the package repository, and spawns jobs to build missing packages via a master/slave cluster managment system. Packages are build and uploaded to a shared filesystem, where they are then made available over the

network to clients wishing to install software.

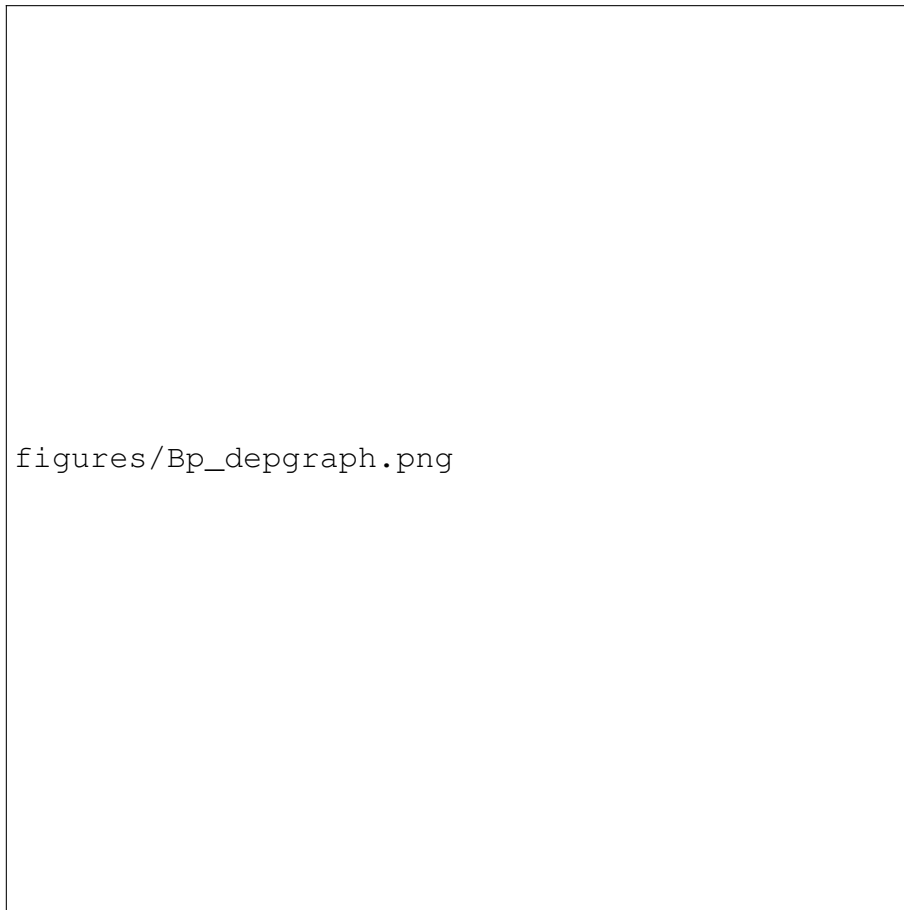


Figure 4.1: Partial package dependency graph.

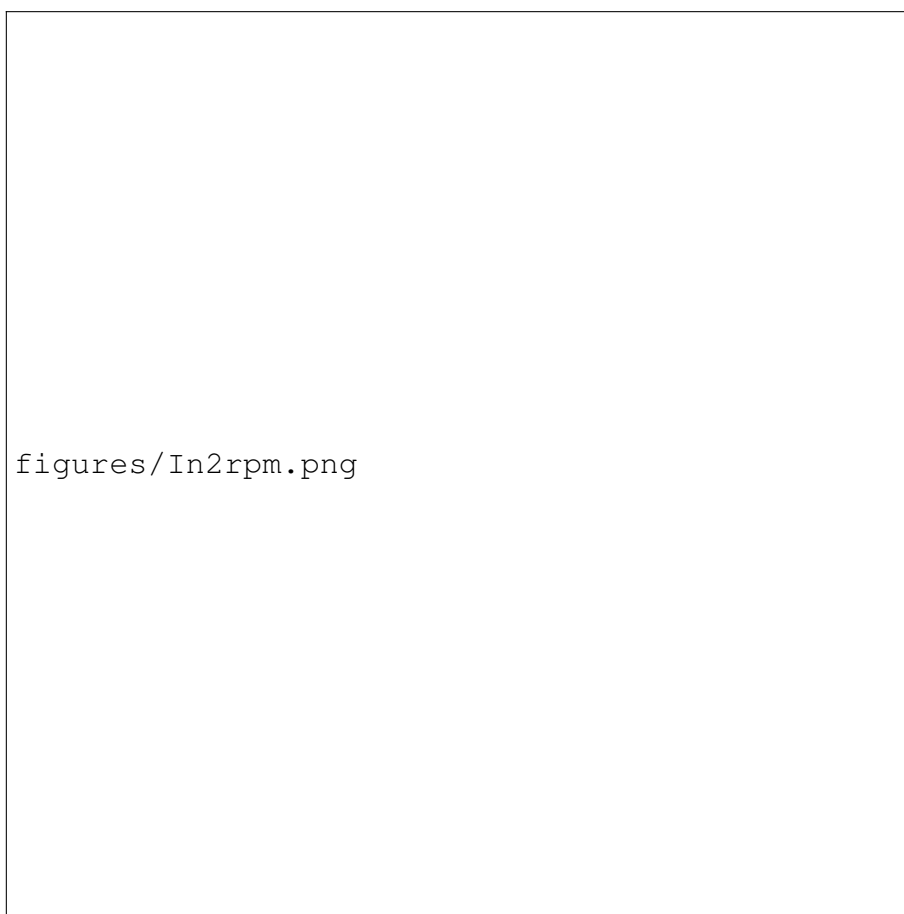


Figure 4.2: Package compilation process.

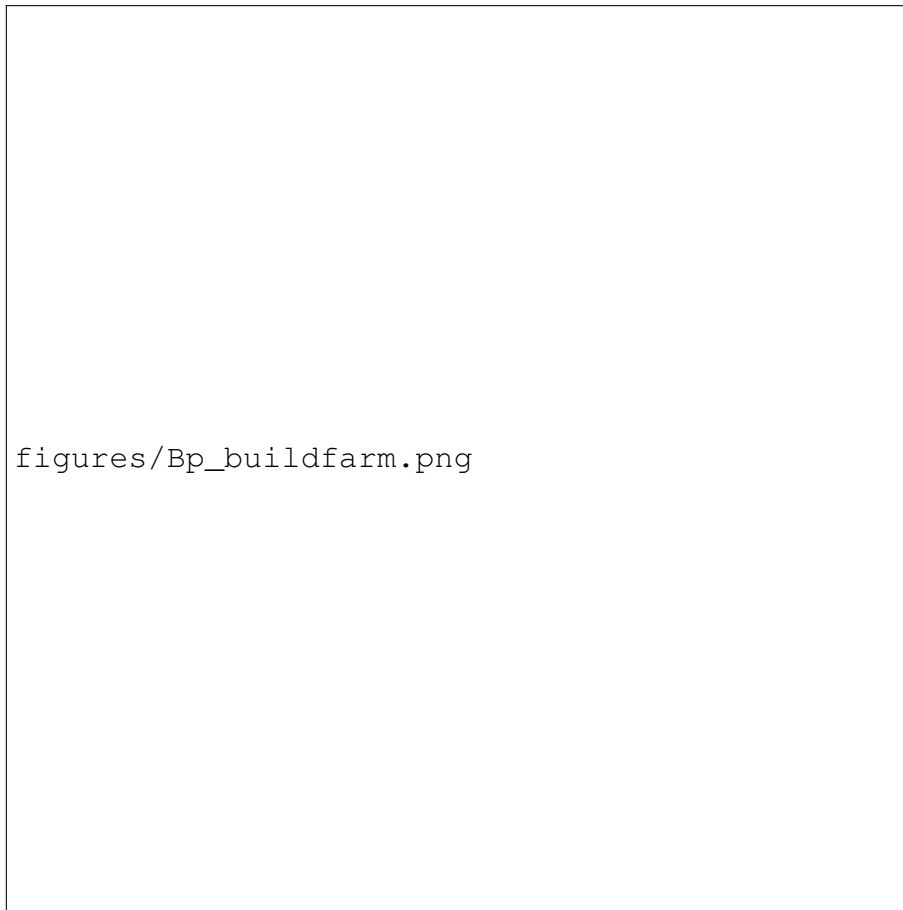


Figure 4.3: Biopackages.net build farm architecture.

CHAPTER 5

GMODWeb: A Web Framework for the Generic Model Organisms Database

5.1 Abstract

The GMODWeb project is a software framework designed to speed the development of websites for Model Organism Databases (MODs). GMODWeb is part of the larger Generic Model Organisms Database (GMOD) initiative which provides species-agnostic software tools and data models for representing curated model organism system data. Users of GMODWeb can browse and search through many different data types including genomic features and annotations, stocks, literature references, and genomic maps. It is also integrated with other GMOD tools such as GBrowse, AmiGO, and Textpresso. To assist development of new MOD sites end-to-end examples of fully customized and integrated GMODWeb instances for the *H. sapiens* and *S. cerevisiae* genomes have been created. The recently inaugurated ParameciumDB also uses GMODWeb as its core website technology for presenting key data of interest to this MOD community. GMODWeb is built on the flexible Turnkey web framework and is freely available under an open source license from <http://turnkey.sourceforge.net>. User documentation, support forums, and source downloads are available at this site while pre-packaged versions for Linux distributions are downloadable at the Biopackages repository for bioinformatics software (<http://biopackages.net>).

5.2 Introduction

Model Organism Databases (MODs) are built around the information needs of scientists working on a single model organism or group of closely related organisms. Examples of MODs include Flybase (<http://www.flybase.org>) [?], Wormbase (<http://www.wormbase.org>) [?], the Mouse Genome Informatics (MGI) Database (<http://www.informatics.jax.org>) [?], the *Saccharomyces* Genome Database (SGD) (<http://www.yeastgenome.org>) [?], Gramene (<http://>

www.gramene.org), a monocot genomics database [?], and ParameciumDB (<http://paramecium.cgm.cnrs-gif.fr>) [?]. MODs provide scientists with access to information about genomic structure, phenotypes, and mutations along with large-scale datasets such as those generated by gene microarray experiments, SNP analyses, or protein-protein interaction studies. A key concern for any MOD is to provide well-designed and convenient community tools for accessing this information. All MODs create websites to fulfill these needs, an expensive and time-consuming prospect. As many more model organisms are sequenced the costs, in terms of both time and funds, of independently developing schemata and web-based tools will become prohibitive.

Recognizing this duplication of work, the NIH and the USDA Agricultural Research Service funded the Generic Model Organisms Database (GMOD) project with the goal of developing flexible applications that can be used across all MODs. The result is a collection of database and web tools that can be mixed and matched to meet the requirements of new MODs. To date, this effort has produced several high-profile components. A generic and modular relational database schema, called Chado, provides the core mechanism to store genomic features, information on gene function, genomic diversity data, literature references, and other common data types. Other popular GMOD tools include Apollo [?], an application for genomic curation, GBrowse [?], a web-based genomic browser that can effectively display genomic features across megabases of sequence, and Textpresso [?], a web tool for literature archiving and searching. While several solutions exist for representing genome annotation data on the web, such as Ensembl [?] and the UCSC genome browser [?], no solution exists for representing the full variety of data types needed for a MOD. In this paper we describe GMODWeb, a flexible and extensible framework for creating a MOD website that integrates with other GMOD tools and accommodates all the data types needed for a model organism database.

5.3 Results

5.3.1 GMODWeb Architecture

GMODWeb is based on the Turnkey (<http://turnkey.sourceforge.net>) site generation and rendering framework which consists of two distinct components. The first is a code creation tool (Turnkey::Generate) that produces a Model View Controller (MVC) based website given a database schema file [?]. In the case of GMODWeb this is the Chado schema. The second Turnkey component (Turnkey::Render) is a page-rendering module that links the generated MVC code to an Apache webserver (<http://apache.org>). This portion of the Turnkey framework uses a collection of open source perl modules and the popular mod_perl webserver plugin (<http://perl.apache.org>). Each Turnkey component is used in a different phase of website construction. While the MVC generator automates the creation of most site code, the page-rendering module handles the response to user requests received by the webserver.

Turnkey, and GMODWeb by extension, is strictly divided into MVC layers. This style of abstraction is a useful tool for organizing a web application into manageable layers and improves the overall organization of the software. Likewise, the active code generation approach used by Turnkey, which is similar to the Object Management Group's (<http://www.omg.org/mda>) Model Driven Architecture (MDA) proposal, is especially useful for the GMODWeb project because underlying changes in the data model are quickly and easily integrated into the application [?]. For example, the inclusion of new database modules in Chado can be easily accommodated by regenerating the Turnkey site from the database schema file. GMODWeb is produced by simply applying customizations, a GMODWeb "skin", to the Turnkey website auto-generated using the Chado schema. This decoupling of user interface customization

from underlying data structure makes the GMODWeb application easy to extend, customize, and maintain. Figure 1 shows the close relationship between GMODWeb and Turnkey.

5.3.2 GMODWeb Site Generation and Rendering

The creation of a Turnkey site, such as GMODWeb, begins with a SQL schema file used to define the tables in a database and how they relate to each other. This file is abstracted into relationships between objects forming a Directed Graph (DG). Turnkey::Generate uses the perl module SQL::Translator to perform this conversion from a SQL schema file to a DG object model (<http://sqlfairy.sourceforge.net>). For example, in the Chado schema a `feature` table stores information about genomic features such as mRNAs or genes. This table is linked to many other tables, such as the `synonym` table via the `feature_synonym` table. The Turnkey::Generate script creates objects representing each table (`feature`, `synonym` and `feature_synonym`) and their individual data fields. It then creates links between these objects to mimic the relationships encoded by the schema, in this case linking the `feature` and `synonym` tables.

Using the relationships encoded by the DG, Turnkey::Generate produces an MVC mod_perl website. Each layer of the MVC framework is created using Template::Toolkit templates. The model layer, which handles the flow of information to and from the underlying database, is created using a template to produce Class::DBI-based objects (<http://wiki.class-dbi.com>). Class::DBI is a convenient tool to connect and retrieve information from the database because it abstracts complex SQL queries into easy-to-use object calls. Controller objects, called atoms in the Turnkey framework, wrap the model objects and provide an abstraction between the view and model objects. They also include the logic necessary to bring these two layers together. The view layer is, itself, implemented in Template::Toolkit and uses HTML with embed-

ded tags to extract information from controller objects for display to the end user. Turnkey::Generate also creates the Turnkey.xml controller document that describes how model and view objects are to be combined by the atom controller objects. Figure 2a illustrates the MVC-based architecture created with the Turnkey::Generate software.

Once created, the output of Turnkey::Generate is configured to work in an Apache server using the mod_perl framework. The process of rendering a page is handled by Turnkey::Render. When a user requests a certain URL, the Turnkey.xml document is examined by Turnkey::Render and the appropriate Class::DBI model and controller atom objects are instantiated. For example, the `feature` table described previously has an entry in this XML linking it to the `synonym` table through the `feature_synonym` table. This provides Turnkey::Render with enough information to create atom and model objects for both the `feature` and `synonym` tables. Following this, the appropriate template view objects are created and Turnkey::Render uses the atom controller objects to manage the handoff of objects and template files to the Template::Toolkit engine for rendering. The resulting HTML output is then returned to the client (Figure 2b).

5.3.3 GMODWeb Customization

Customization is an important feature that all MODs require in their web interfaces. To accommodate this, key design features were integrated into the Turnkey framework affecting both the site generation and page rendering processes. These include template customization through overriding and Cascading Style Sheet-based (CSS) layouts (<http://www.w3.org/Style/CSS>). Template overriding provides the ability for MOD developers to create a customized look and feel for a given type of information being displayed in a GMODWeb site. For example, the default `feature` page in GMODWeb was swapped out for a custom templates that included a `GBrowse`

panel if the feature had a genomic location, as is the case for a gene or an mRNA transcript feature. These custom templates would normally be overwritten in an MDA web framework but Turnkey allows site designers to create and persist these modifications. In addition to template customization, layout and styling in a Turnkey site is accomplished with flexible CSS documents, allowing the MOD developer to dramatically change the look and feel of the entire site. Not only can colors and fonts be changed, but element layouts can be reordered.

A combination of these customizations can be grouped together into a "skin" which can easily be parameterized and switched on the fly. This makes it possible for a MOD website to be context dependent and support a "print" view or completely different color scheme with the same underlying website and database. For example, a clade-oriented database that provides information on 12 different beetle species could apply a different page color to each species to avoid user confusion. GMODWeb was created by taking a Turnkey website generated from the Chado database schema and applying these types of customizations to the codebase (Figure 1).

Demonstration GMODWeb sites have been created for *H. sapiens* and *S. cerevisiae* and include the basic functionality associated with a typical MOD's homepage. These sites illustrate the common layout for a Turnkey website and show the effects of a customized GMODWeb skin. The sample websites include the ability to search by features and controlled vocabulary (CV) terms indexed from the underlying Chado database using the open source search engine Lucene (<http://lucene.apache.org>). Since many data types in Chado are annotated and linked together through CV terms using various ontologies, such as the Gene Ontology (GO) [?], it was important to be able to query by both data types. Search results will take an end user to either a feature or CV term page rendered using customized GMODWeb templates.

Browsing a feature reveals several customizations to the default templates. Figure 3

shows a typical gene feature page using the GMODWeb skin from the ParameciumDB MOD website. In this example, the basic layout of a Turnkey page is evident: the item being rendered, in this case a row from the `feature` table, is present as the major content panel while linked tables are represented as minor panels on the left-hand side. For this gene feature, two types of linked data were presented on the left: external references (via the `feature_dbxref` table) and relationships to other features in the database (via the `feature_relationship` table). Customizations of links and panel headings in both the major and minor panels are shown in this example as well.

Further customization was used in the major panel to organize information about the gene feature in an intuitive and helpful way. Related content, such as GO term annotations, genomic location, synonyms, and other information, was included as a summary. The Turnkey framework's flexibility allows custom template authors to easily extract this information using the underlying `Class::DBI` model objects. In this customized template, simple nested method calls on the feature model object were used to extract linked information such as synonyms. Together these modifications have created a gene page that can be leveraged across MODs and provide many of the key pieces of information about genomic features that end users will require. Turnkey pages also contain an edit link which provides a limited but useful facility for editing record data. Authentication is provided by standard HTTP access controls in Apache.

5.3.4 GMODWeb Integration

The example in Figure 3 shows how GMODWeb's templates were directly integrated with other GMOD projects. In this page, a GBrowse instance was embedded and provided not only a graphical view of the genomic neighborhood but also linked out to nearby genes and other annotations. In addition to GBrowse, the sample GMODWeb sites for *H. sapiens* and *S. cerevisiae* include integration with Textpresso for litera-

ture tracking, BlastGraphic for performing Blast analysis, and AmiGO for controlled vocabulary term visualization. These dependencies, which are available from the Biopackages software repository, have been pre-configured to work with the GMOD-Web demonstration sites. Packaging the sample applications and their dependencies makes installation and configuration a quick and easy task for site developers and jumpstarts the process of setting up new MOD websites.

In addition to web interfaces, GMODWeb also provides Simple Object Access Protocol (SOAP) (<http://www.w3.org/TR/soap>) bindings for accessing data in an automated, programmatic way. This web services approach is designed to allow savvy end users to interact directly with the underlying GMOD Chado database, affording bulk access to features contained within the database. Providing this tool for GMODWeb's model objects makes data access platform agnostic so developers can interact with the service using the language of their choice. Apache2::SOAP (<http://search.cpan.org/~rkobes/Apache2-SOAP-0.72>) was used to bind Class::DBI-based model objects to a SOAP interface. Unlike XML genome feature annotation services, such as the Distributed Annotation System (DAS) [?], the SOAP bindings present low-level interfaces to database tables. This SOAP interface is pre-configured and immediately available for all MOD sites based on GMODWeb.

5.3.5 Case Study: Creating a New MOD Website with GMODWeb and Turnkey

Paramecium, a unicellular eukaryote that belongs to the ciliate phylum, has served as a genetic model organism for over half a century and is also widely used to teach biology. The genome of *Paramecium tetraurelia* was recently sequenced and annotated at the Genoscope French National Sequencing Center [?]. In anticipation of public release of the data from the sequencing initiative, a project was started in 2005 to develop a Paramecium community MOD, ParameciumDB. Its immediate objectives were to

integrate the genome sequence and annotations with available genetic data and coordinate the manual curation of the gene models by members of the research community. Ultimately, ParameciumDB should provide a useful resource for the classroom as well.

GMOD's Chado database schema was well suited for this project because of its genetic module, which ensured the integration of both the genetic and sequence data, and its support for describing phenotypes using controlled vocabulary terms. Another important factor in choosing the GMOD toolkit for ParameciumDB was the availability of Turnkey and GMODWeb to generate the MOD's website, since it was anticipated that this would be the most difficult part of the project.

GMODWeb was first tested on a generic installation of Chado, populated with published data from a previously sequenced and annotated Paramecium chromosome [?]. The next step was modeling the genetic data, which involved writing a stock sub-module for the Chado genetic module to make it possible to incorporate data about Paramecium stock collections. Since the Chado database schema was modified, Turnkey::Generate was used to create a custom website for ParameciumDB.

The last and most time-consuming step for building ParameciumDB was customization of the auto generated website layout. The overall design of the site components (header, footer, feature page, etc) was achieved using templates and CSS modifications of the GMODWeb skin, resulting in a custom ParameciumDB skin. Within the auto-generated view code, the feature-relationship atom object was modified to make it possible to recover the complete hierarchy of relationships (e.g. gene → mRNA → exon) from the top-level gene feature, even if the feature page being rendered concerned a feature type lower in the hierarchy. Additional static content was added to the site using Microsoft Frontpage including a help section, project documentation, and announcements (<http://office.microsoft.com/frontpage>). Finally, commonly used applications were integrated into the ParameciumDB by link-

ing to other bioinformatics tools, such as NCBI's BLAST tool [?], and to forms for data submission by the community.

The templates for pages within ParameciumDB were customized using many of the ideas taken from the GMODWeb sample sites and, in particular, the layout of the sample gene page. The elegance of the Turnkey-based MVC site was most apparent at this level: customization of ParameciumDB was focused on the template view layer while relatively few changes were required to either the model or controller objects. The bulk of the code produced for ParameciumDB was automatically generated and untouched by the customization process. This freed developers to work on the effective presentation of MOD data rather than low level database access or website rendering code.

5.4 Discussion

Model organism databases (MODs) gather together biologic information on a variety of important organisms for the scientific community. A key concern for any MOD is to provide well-designed and easily accessible tools for sharing this information. The GMODWeb project was started to provide a simple and generic solution for quickly creating new MOD websites using the Turnkey framework. GMODWeb, by running directly off the flexible and extensible Chado schema, can accommodate the wide variety of data types and usage patterns that model organism communities require. It offers both a clean MVC framework and pre-built sample websites configured to work with other GMOD tools. Together these features can greatly streamline the process of new MOD website development.

One of the challenges for any framework based on Model Driven Architecture techniques is balancing auto-generated code tied to a particular underlying schema with

customized layouts for creating compelling and effective user interfaces. Turnkey, the technology underling GMODWeb, attempted to solve this limitation by providing a mechanism for bundling specialized skins with an auto-generated website. Since most of the website is automatically created, designers can focus on the quality of the user interface and not on the underlying rendering code. In GMODWeb this translated to extensive customization of the default feature templates. With the adoption of GMODWeb by the ParameciumDB project, we anticipate incorporating UI changes and customization into GMODWeb for the benefit of all future GMODWeb-based MODs. As additional MODs adopt GMODWeb, we envision the availability of a large library of site-specific customized templates, which can be adopted, altered and expanded by subsequent MOD projects.

The rapidity with which ParameciumDB was built, by a very small development team, is encouraging. For this MOD, data modeling was much more time-consuming than building the GMODWeb-based website. In fact, the only difficulty encountered in implementation of ParameciumDB was not with GMODWeb or Turnkey *per se*, but with the installation and tuning of `mod_perl` and the Apache web server for use in a production environment. The current availability of GMODWeb sample sites and installation dependences as pre-compiled software packages on Biopackages should make this part of a new project much easier for future MODs.

As the model organism database community continues to expand, there will be an increased need to leverage existing tools to store, query, and present MOD data. The GMOD project was created to engineer generic tools to meet these needs. GMODWeb was designed to quickly create MOD websites based on the easy to use, customize, and update Turnkey framework. A GMODWeb MOD site provides not only the ability to browse and search MOD data but it also forms a key link to other tools. As future applications and components become available, GMODWeb will continue to act as a

natural point of integration and a central hub for the display of MOD data to end users.

5.5 Methods

5.5.1 Availability

Turnkey and GMODWeb are both available as a source code files from Turnkey's website (<http://turnkey.sourceforge.net>) and as pre-compiled packages for various Linux distributions from the Biopackages repository for bioinformatics software (<http://biopackages.net>). All dependencies are provided using the Red Hat Package Manager (RPM) (<http://www.rpm.org>) and integration with other programs, such as GBrowse, is accomplished using this same package management system. When a MOD installs GMODWeb via RPMs, a pre-configured GBrowse, Blast server, Textpresso, and other applications are installed and configured to work within GMODWeb immediately. Table 1 shows the software dependencies for GMODWeb, all of which are available either from specific Linux distributions or through Biopackages.

5.5.2 User Feedback

The GMODWeb project is supported with help from the larger open source development community. Information on installation, troubleshooting, and optimization can be found on the Turnkey website at <http://turnkey.sourceforge.net>. MODs setting up GMODWeb can also solicit help from the email lists either at this site or at GMOD's homepage (<http://gmod.org>) where a community of users is very active.

5.5.3 Unresolved Challenges

As with any open source development project, challenges remain for GMODWeb. Although the project has been available for two years, it has only recently released a 1.0 version. It has been a challenge to attract new MOD users and developers when the project was in this pre-release stage. With the release of ParameciumDB as a proof of concept for GMODWeb in a production environment, the prospects for attracting both new users and developers have improved. Another challenge for the project is the very integrative nature of the GMODWeb application. Since it attempts to bring together several very large web applications, the dependencies for the project are daunting. Maintenance on the various GMOD components integrated into GMODWeb has taken up a large percentage of the development time. However, as a beneficial side effect of this effort, many useful GMOD tools have been packaged as RPMs for distribution through the Biopackages repository, making them available for other projects and uses.

5.6 Figure Captions

5.6.1 Figure 4.1

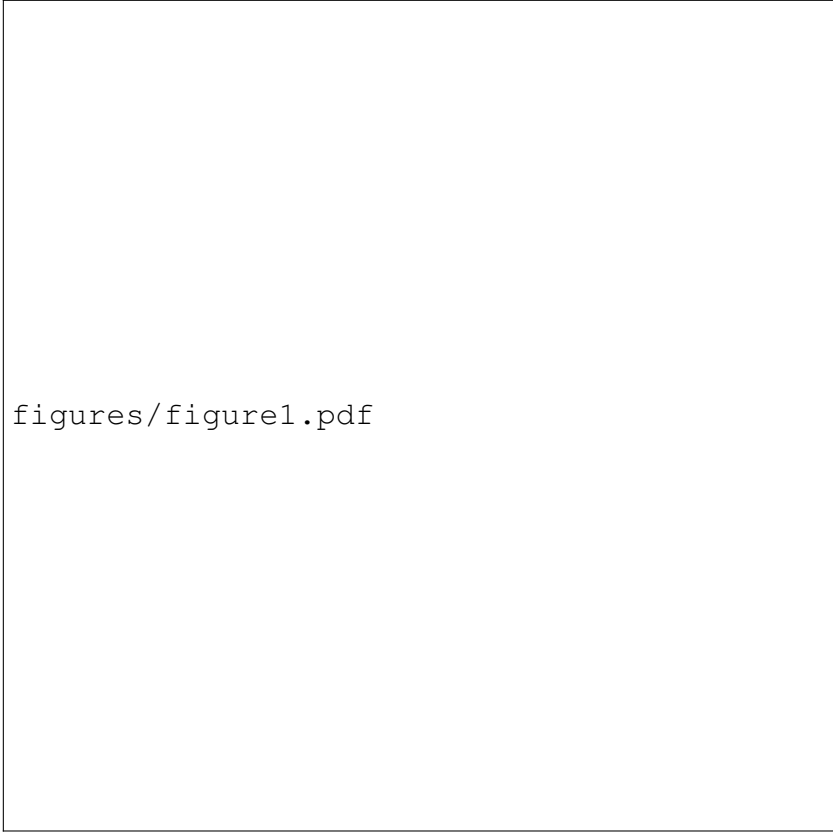
GMODWeb is the result of customizations to a Turnkey website built with the Chado schema. The GMODWeb skin was the product of modifications mainly to the view layer. This included changes to the template view layer including overriding default templates and CSS changes. Enhancements were also performed with layout changes through controller XML files modifications.

5.6.2 Figure 4.2

Overviews of the Turnkey::Generate and Turnkey::Render processes. A. The process of creating a Turnkey-based website via Turnkey::Generate is shown. A SQL schema file is processed using SQL::Translator to create a directed graph representation of the relationships between tables. These are used by Turnkey::Generate to create an MVC-based web application. B. The rendering of a Turnkey page by Turnkey::Render is shown. When a client request is received an XML document describing the relationships between objects is consulted. Model objects are created and combined with templates by the atom controller layer to produce a rendered page. This is returned to the client.

5.6.3 Figure 4.3

An example gene feature rendered with the customized ParameciumDB skin.



figures/figure1.pdf

Figure 5.1: GMODWeb and its relationship to Turnkey.

Table 5.1: The GMODWeb application has many software dependencies. This table shows the immediate GMODWeb and Turnkey dependencies on the Fedora Core 2 Linux distribution. All dependencies are available from the Biopackages software repository (<http://biopackages.net>) or are provided by the underlying operating system.

Package Name	Version	GMOD Tool	Description
postgresql-server	≥ 7.3		The PostgreSQL database server.
postgresql			Client libraries for PostgreSQL.
perl-Apache-ParseFormData			A perl library for accessing form data in mod_perl.
perl-Apache2- perl-Class-Base			A perl base class for other modules.
perl-Class-DBI			A perl tool for abstracting database access.
perl-Class-DBI-ConceptSearch			A flexible perl module for searching databases.
perl-Class-DBI-Pager			A perl tool for breaking database query results into pages.
perl-Class-DBI-Pg			A PostgreSQL driver for Class::DBI.
perl-Class-DBI-Plugin-Type			A perl tool for determining data type information.
perl-DBD-Pg			A PostgreSQL driver for perl.
perl-DBI			A generic database interface for perl.
perl-Log-Log4perl			Logging software for perl applications.
perl-SQL-Translator			A perl tool for translating SQL schema into an object model.
perl-Template-Toolkit			A template engine for perl.
perl-XML-LibXML			An XML parsing library for perl.
perl-Lucene	$\geq 2.0.1$		A Lucene search engine interface for perl.
perl-Apache2-SOAP			Automatic SOAP bindings for mod_perl.
perl-Cache-Cache			A perl tool used to cache web pages in GMODWeb.
httpd			The Apache webserver.
mod_perl			A plugin for Apache that executes perl code.
perl			An interpreted language used throughout the Turnkey/GMODWeb project.
gbrowse		yes	A genome feature browser web application from the GMOD project.
textpresso		yes	A literature search web application from the GMOD project.
AmiGO		yes	An ontology browser web application from the GMOD project.
chado		yes	A sample Chado database from the GMOD project.
chado-schema		yes	The Chado schema from the GMOD project.
gmod-web	≥ 1.3	yes	A GMODWeb site generated with Turnkey for the Chado schema.
turnkey	≥ 1.3		The website generation tool used to create GMODWeb.

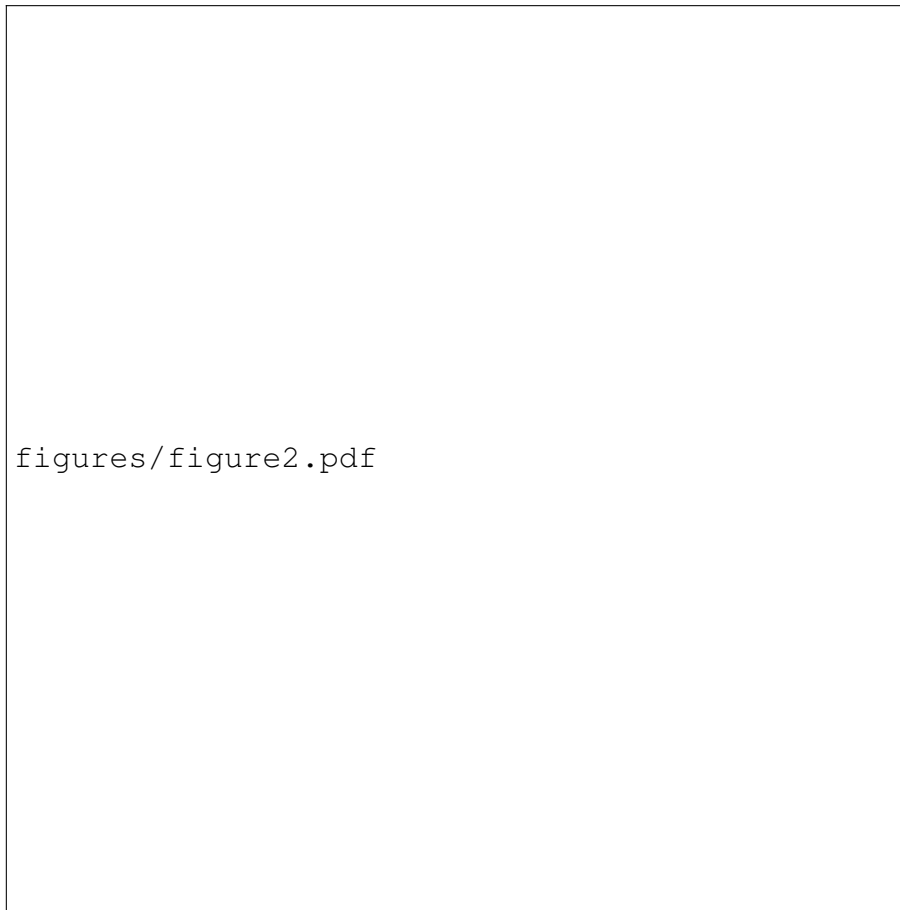


Figure 5.2: Overviews of the Turnkey::Generate and Turnkey::Render processes.

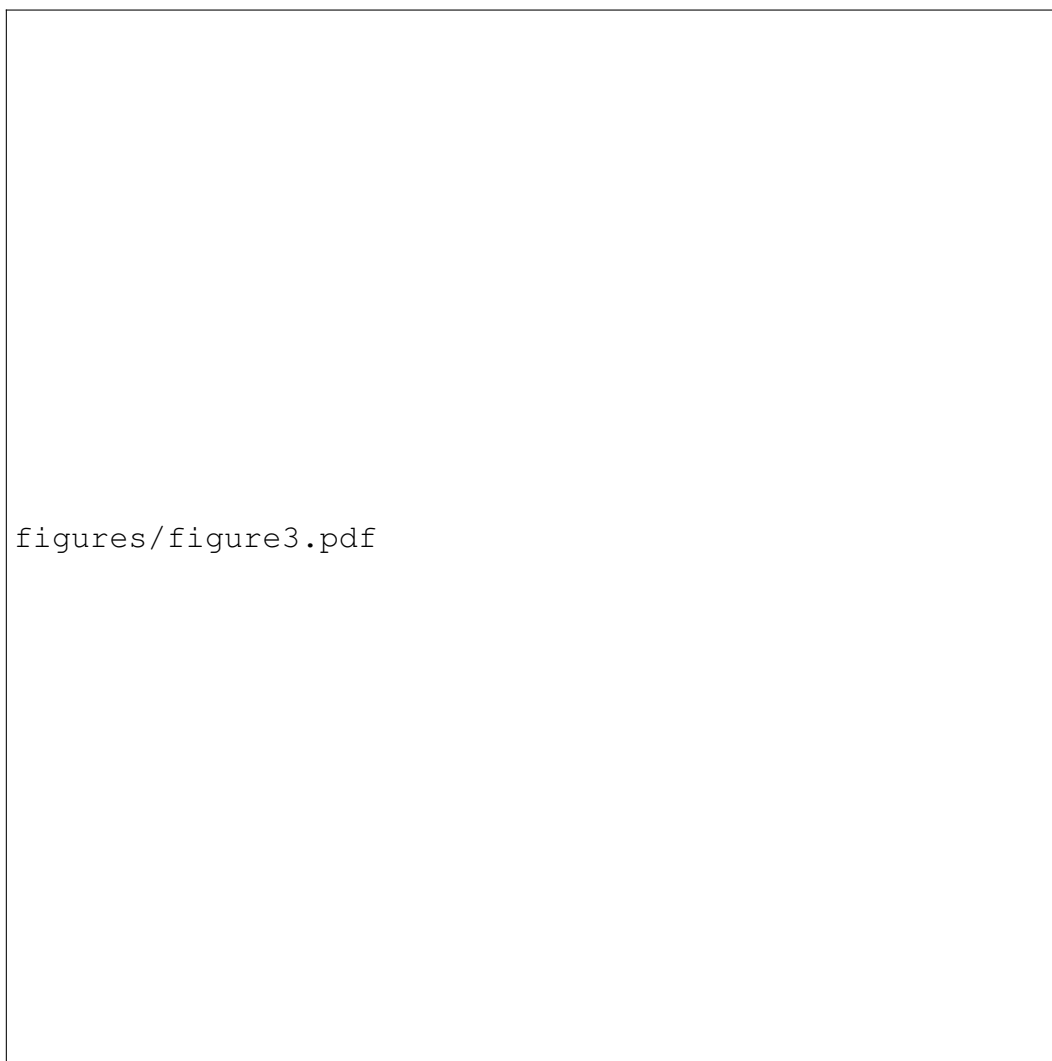


Figure 5.3: An example gene feature rendered with the customized ParameciumDB skin.

APPENDIX A

The Distributed Annotation System

Research article

The Distributed Annotation System

Robin D Dowell¹, Rodney M Jokerst¹, Allen Day², Sean R Eddy¹ and
Lincoln Stein^{*2}

Address: ¹Howard Hughes Medical Institute and Department of Genetics, Washington University, St. Louis, MO 63110 USA and ²Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724 USA

E-mail: Robin D Dowell - robin@genetics.wustl.edu; Rodney M Jokerst - jokerst@genetics.wustl.edu; Allen Day - day@cshl.org; Sean R Eddy - eddy@genetics.wustl.edu; Lincoln Stein* - stein@cshl.org

*Corresponding author

Published: 10 October 2001

BMC Bioinformatics 2001, 2:7

This article is available from: <http://www.biomedcentral.com/1471-2105/2/7>

Received: 10 August 2001

Accepted: 10 October 2001

© 2001 Dowell et al; licensee BioMed Central Ltd. Verbatim copying and redistribution of this article are permitted in any medium for any non-commercial purpose, provided this notice is preserved along with the article's original URL. For commercial use, contact info@biomedcentral.com

Abstract

Background: Currently, most genome annotation is curated by centralized groups with limited resources. Efforts to share annotations transparently among multiple groups have not yet been satisfactory.

Results: Here we introduce a concept called the Distributed Annotation System (DAS). DAS allows sequence annotations to be decentralized among multiple third-party annotators and integrated on an as-needed basis by client-side software. The communication between client and servers in DAS is defined by the DAS XML specification. Annotations are displayed in layers, one per server. Any client or server adhering to the DAS XML specification can participate in the system; we describe a simple prototype client and server example.

Conclusions: The DAS specification is being used experimentally by Ensembl, WormBase, and the Berkeley Drosophila Genome Project. Continued success will depend on the readiness of the research community to adopt DAS and provide annotations. All components are freely available from the project website [<http://www.biodas.org/>].

Background

With the rise of computational biology and the decrease in hardware costs, high throughput annotation is now possible within many laboratories. They can now annotate entire genomes relatively quickly and efficiently. What has not kept up with the pace of annotation is the ability for multiple groups to exchange and compare their data, leading to fragmentation of annotation information among multiple databases and web sites, and to a certain level of frustration among the bench biologists who are the intended beneficiaries of this data.

Ideally, an annotation system should give individual experts the ability to contribute to the collective annotation in a quick, robust, and mostly painless fashion. They should have complete control over their annotations in order to keep them current and relevant. These annotations should not need approval from a central authority. Simultaneously, it should be easy for a user to obtain and visualize the most recent data about their particular region of interest. Users would also prefer not to be swamped by bogus information. Unfortunately, these goals seem to be at odds in the current sequence annotation environment.

Initial database efforts were largely centralized repositories such as GenBank, established in 1982 [1]. These databases act primarily as archival storage of sequence information. Consequently, each entry is owned by the sequence provider and integrating annotation information is, by design, nearly impossible.

A number of specialized databases have developed to serve a curatorial role within particular communities, such as Swissprot [2], Refseq [3], and WormPD [4]. A *C. elegans* database (ACeDB) is one particularly successful community database [5] [<http://www.acedb.org/>]. It has served as the central database of phenotyping, bibliographic, mapping, and sequencing information for the *Caenorhabditis elegans* community since 1990 [6]. Individuals are encouraged to submit annotations and changes to the central database curatorial group. The group then reviews the request and decides what and how it is to be incorporated into the next official release. With limited numbers of curators available, these databases find it difficult to keep up with the requests of many expert annotators.

To overcome the restrictions of archival databases and the bottlenecks of curatorial databases, a number of groups have attempted to develop third party annotation systems. Examples include the Worm Community System [7], the Genome Sequence Database [8], and GDB [9,10]. These systems typically require global coordination by either keeping all annotations in a centralized open repository or by forcing all parties to adhere to a common database format or by requiring a controlled vocabulary.

Another recent experiment with third party annotation has been the "annotation party," exemplified by Celera's Fly Jamboree and the Human Genome Project Consortium's Analysis Group (HGPCAG). Parties gather together a large number of experts to produce the best annotations possible in a limited time frame. However, it is not clear that the annotation party model is sustainable once the initial flush of enthusiasm has worn off.

The HGPCAG model has a notion of annotation "tracks", where a track contains a particular kind of annotation produced by a particular participating group. For example, the Eddy lab provides a noncoding RNA track that annotates the positions of RNA genes in the human genome. Annotation tracks are independent of each other and therefore easy to integrate into a single display. The concept is essentially identical to the independent columns of annotation displayed by an ACeDB browser, except that the tracks in the HGPCAG annotation are curated by a variety of groups at different institutions, as opposed to a centralized curation group. However, the

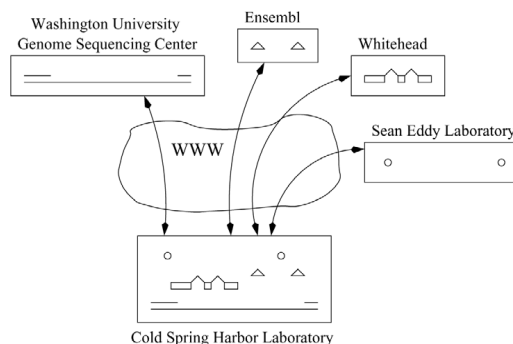


Figure 1
Basic distributed annotation system architecture One server is the designated reference server, in this case the Washington University Genome Sequencing Center. One or more annotation servers, shown above as Ensembl, Whitehead, and the Sean Eddy Laboratory, provide annotations relative to the reference sequence. The client, at Cold Spring Harbor Laboratory in our example, fetches data from multiple servers and automatically generates an integrated view.

data for every track are still kept on a single centralized server; updating an annotation track after it has been submitted is cumbersome.

Here we introduce a genome annotation strategy that enables third-party annotation in a way that allows annotators to control and update their work, and which does not require much centralized coordination. The Distributed Annotation System (DAS) was designed as a lightweight system for integrating data from a number of heterogeneous distributed databases. The DAS system has a notion of annotation "layers", which are essentially identical to tracks, except that now the data for each layer are on "third party servers" that are controlled by each annotation provider. The key idea was to produce a data exchange standard (the DAS XML specification) that enables layers to be provided in real time from 3rd party servers and overlaid to produce a single integrated view by a DAS client.

Figure 1 shows a cartoon example of the DAS paradigm. The client selects a single reference genome server and any number of annotation servers. The display layers the data returned from each server. A particular annotation can then be queried to retrieve more information from its providing server, as HTML pages.

Implementation

The basic system is composed of a genome server, one or more annotation servers, and an annotation viewer. The

genome server is responsible for serving genome maps, sequences, and information related to the sequencing process. Annotation servers are responsible for responding to requests on a region and delivering annotations. The client, an annotation viewer, is a lightweight application whose behavior is analogous to a web browser. The viewer communicates with the genome and annotation servers using a well defined language specification.

At a fundamental level, all annotations can be reduced to their coordinates relative to a particular sequence landmark. The DAS viewer retrieves annotations from the various annotation servers and uses the sequence coordinates to generate an integrated index of what is on the genome. This integration is then presented to the user in tabular or graphical form. Annotation providers can provide a suggestion of how their annotations should be rendered in a graphical display, and can provide links back to their databases and web sites to allow the researcher to retrieve further information about the annotation.

Because it relies entirely on sequence coordinates to achieve integration, DAS does not attempt to resolve semantic contradictions between different data sources. The goal of the system is to provide indexing and visualization, thereby making contradictions between annotations visible.

Reference sequence

The distributed annotation system relies on there being a common "reference sequence" on which to base annotations. The reference server consists of a set of "entry points" into the sequence, and the lengths of each entry point. Entry points will vary from genome to genome. For some genome projects, entry points correspond to entire chromosomes. For others, entry points may be a series of contigs.

The entry points describe the top level items on the reference sequence map. It is possible for each entry point to have substructure, basically a series of subsequences (components) and their start and end points. This structure is recursive. Annotations take the form of a statement about a region of the reference sequence. Each annotation is unambiguously located by providing its position as the start and stop positions relative to a "reference sequence."

To give a concrete example, the *C. elegans* reference map consists of six top level entry points, one per chromosome. Each chromosome is formed from several contigs called "superlinks," and each superlink contains one or more smaller contigs called "links." Links in turn are composed of one or more fully-sequenced clones [11]. One could refer to an annotation by specifying its start or

stop positions in clone, link, superlink, or chromosome coordinates.

The reference sequence server is responsible for providing the reference sequence map and the underlying DNA. The server can provide a list of sequence entry points or given a component of the map it can return its parent and children components. The reference server can provide arbitrarily long stretches of raw DNA sequence given a reference subsequence, start position, and stop position. Needless to say, bandwidth becomes a limiting factor for retrieving multi-megabase segments of DNA. However, in practice it is rare for users to retrieve more than a gene's worth of raw DNA at a time.

Annotation servers

Annotation servers are specialized for returning lists of annotations across defined regions of the genome. Each annotation is anchored to the genome map by way of a start and stop position relative to one of the entry points. Annotations have an identifier that is unique to the providing server and a structured description of its nature and attributes. The general description of an annotation follows loosely the general feature format (GFF) which intentionally aims for a basic lowest common denominator description [<http://www.sanger.ac.uk/Software/formats/GFF/>]. Annotations may also be associated with URLs where additional human or machine readable information about the annotation can be found.

The annotator is free to describe his annotations using any terms which he feels are appropriate, as DAS does not impose a controlled vocabulary. Annotations have *categories*, *types*, and *methods* defined by the annotator. The annotation **type** corresponds to a biologically significance description. In the Eddy Lab RNA track of the HGP three types are defined, "tRNA", "snoRNA", and "miscRNA". The annotation **method** is intended to describe how the annotated feature was discovered, and may include a reference to a software program. The annotation **category** is a broad functional category. "Homology", "variation" and "transcribed" are example categories. This structure allows researchers to add new annotation types if the existing list is inadequate without entirely losing all semantic value. It is intended that larger annotation servers provide URLs to human-readable information that describes its types, methods and categories in more detail.

Another optional feature of annotation servers is the ability to provide hints to clients on how the annotations should be rendered visually. This is done by returning a DAS "stylesheet." Stylesheets use the **type** and **category** information to associate each annotation with a particular graphical representation, a glyph.

Table 1: Server Status Codes Server status codes are modeled after the familiar status codes of the HTTP 1.0 protocol.

Code	Meaning
200	OK, data follows
400	Bad command (command not recognized)
401	Bad data source (data source unknown)
402	Bad command arguments (arguments invalid)
403	Bad reference object (reference sequence unknown)
404	Bad stylesheet (requested stylesheet unknown)
405	Coordinate error (out of bounds/invalid)
500	Server error, not otherwise specified
501	Unimplemented feature

Although the servers are conceptually divided between reference servers and annotation servers, there is in fact no key difference between them. A single server can provide both reference sequence information and annotation information. The main functional difference is that the reference sequence server is required to serve the coordinate map and the raw DNA, while annotation servers have no such requirement.

Specification

The main component of DAS is the XML specification, which defines all valid DAS communication. As with HTML, our goal is a language which is human readable, easily parsed, and extensible. The additional file [appendix.pdf] provides a summary of version 1.01 of the DAS specification.

While a client can query multiple servers simultaneously, the communication between the client and any single server follows a simple client server model. Clients query the reference and annotation servers by sending a formatted URL request to each server. Each URL has a site-specific prefix, followed by a standardized path and query string. The standardized path begins with the string **/das**. This is followed by URL components containing the data source name and a command. For example:

`http://stein.cshl.org/das/elegans/features?segment=ZK154:1000,2000`

In this case, the site-specific prefix is `http://stein.cshl.org/`. The request begins with the standardized path **/das**, and the data source, in this case **/elegans**. This is followed by the command **/features**, which requests a list of features relative to a given set of named arguments (*?segment=ZK154:1000,2000*). The data

source component allows a single server to provide information on several genomes.

Servers process the request and return a response as defined by the DAS specification, typically a formatted XML document. The response from the server to the client consists of a standard HTTP header with DAS status information within that header followed optionally by an XML file that contains the answer to the query. The DAS status portion of the header consists of two lines. The first is X-DAS-Version and gives the current protocol version number, currently DAS/1.0. The second line is X-DAS-Status and contains a three digit status code which indicates the outcome of the request. The defined status codes are listed in Table 1.

An example HTTP header: (*provided by server*)

HTTP/1.1 200 OK

Date: Sun, 12 Mar 2000 16:13:51 GMT

Server: Apache/1.3.6 (Unix) mod_perl/1.19

Last-Modified: Fri, 18 Feb 2000 20:57:52 GMT

Connection: close

Content-Type: text/plain

X-DAS-Version: DAS/1.0

X-DAS-Status: 200

DATA FOLLOWS ...

The specification outlines seven basic queries which a client can use to interrogate a DAS server. The valid queries are briefly summarized in Table 2. Two queries, "dsn" and "entry points", essentially provide information to the client about the structure of the server and the reference sequence. The "dna" query can be used to fetch a segment of DNA from a reference server. A client can request annotations, "features", or a summary of the annotations available, "types", from any DAS server. The main annotation content query, "features", basically follows the general feature format (GFF). The servers provide a "stylesheet" to suggest representations to the client's graphical display. When more information is desired about a particular annotation, the client makes a "link" request. The "link" request, the only query which does not return a structured XML document, returns HTML. It is anticipated that DAS clients will hand off the link requests to the local web browser or other web-accessible genome database.

Table 2: Queries Summary The basic seven queries of the DAS 1.0! specification.

Command	Basic Format	Scope
dsn	PREFIX/das/dsn	both
entry-points	PREFIX/das/DSN/entry points	reference
dna	PREFIX/das/DSN/dna?segment=SEG	reference
types	PREFIX/das/DSN/types?segment=SEG	both
features	PREFIX/das/DSN/features?segment=SEG	both
stylesheet	PREFIX/das/DSN/stylesheet	both
link	PREFIX/das/DSN/link?field=TAG;id=ID	both

Prototypes

A series of prototypes for both the client and server components were developed to test various versions of the DAS specification.

Servers

A server is expected to respond to the DAS specification's defined queries with the appropriate content, usually XML. The details of server implementation are left to the various annotation source providers. We provide a sample Perl script for converting ACeDB-based databases into DAS servers, and the Dazzle Java library does the same thing for annotation databases based on the Ensembl code base (T. Down, personal communication, 2001).

The first reference DAS server was written for WormBase [11] and piggybacks on the WormBase software architecture: an Apache/mod_perl web server communicating with an ACeDB database via the AcePerl database access library. The Perl DAS server accepts incoming DAS requests, translates them into the ACeDB query language, reformats the results as XML, and returns them. The WormBase DAS server is currently serving as the *C. elegans* reference server at [http://www.wormbase.org/db/das/]. A set of servers containing test data, one reference and four annotation, are available at [http://skynet.wustl.edu/cgi-bin/das/].

Viewers

We have developed two prototype DAS client programs. One, called Geodesic, is a stand alone Java application. It connects to one or more DAS servers, retrieves annotations, and displays them in an integrated map, as seen in Figure 2. The other, called DasView, is a Perl application that runs as a server-side script. It connects to one or more DAS servers, constructs an integrated image, and serves the image to a web browser as a set of click-able image map, as seen in Figure 3.

Geodesic is mouse and menu driven. The user can choose which data sources to display. The user identifies a segment of the genome to view by browsing through entry points or entering a region name directly. By clicking on a feature, the user obtains additional information in the Feature Details tab and can optionally follow available links back to the original data source. The user can save displayed data as FASTA, GFF, or DAS XML. The user can, to a limited extent, customize the display within the preferences menu.

The DasView prototype implements an alternative mode of using DAS, browserless server side integration. A database can hook into trusted third party servers behind the scenes. The third party data are then integrated into the normal data displays of the database. In this scenario, no DAS client software would be needed.

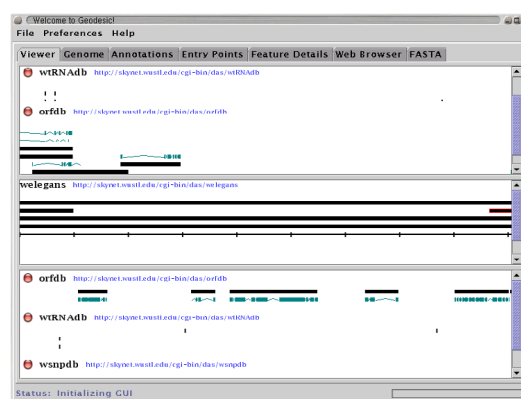


Figure 2
Geodesic A screen-shot of the current version of Geodesic. The view is on clone ZK154 using sources from the *C. elegans* test server set.

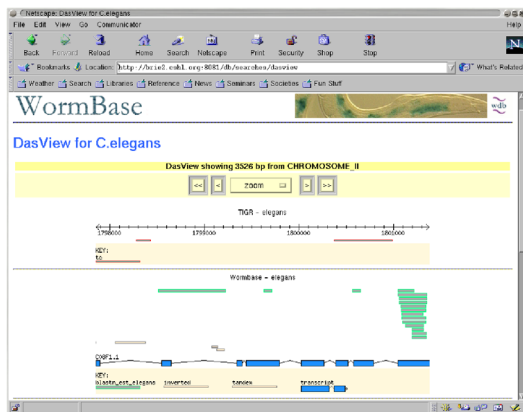


Figure 3
DasView A screen-shot of the current version of DasView. The view is on Chromosome II of WormBase.

Both viewers provide the user with one-click linking back to the primary data sources where they can learn more about a selected annotation, and are sufficiently flexible to accept a wide range of annotation types and visualization styles. The stand alone Java viewer is appropriate for extensive, long-term use. The Perl implementation is suitable for casual use because it does not require the user to preinstall the software.

Discussion

DAS distributes data sources across the Internet improving scalability over monolithic systems. This distribution of data encourages a divide-and-conquer approach to annotation, where experts provide and maintain their own annotations. It also permits annotation providers to disagree about a particular region, encouraging informative dissension and dialogue. The separation of sequence and map information from annotation allows them to be stored and represented in a variety of database schema. A number of different database backend alternatives could arise.

The use of links as a method of referencing back to the data provider's web pages provides even greater power of expression and content control. Annotation providers can make available complex query mechanisms for fine access to more information about the data provided to DAS. Alternatively they can link directly to webpages.

DAS does not enforce third party annotations to be peer reviewed. A strict requirement of peer review would block data sharing activities between collaborating labs.

However, nothing prevents DAS layers from being "blessed" by a data provider, peer reviewer, or by both.

We made a design decision to use an XML-based format. This gives us a strongly typed, extensible data exchange format, but at the cost of non-trivial bandwidth demands. Bandwidth requirements are a substantial concern in the continued design and development of DAS. A user browsing a large genome can easily request more information than their network connection can reasonably handle. The DAS spec attempts to minimize bandwidth demands by representing each annotation with the minimal set of attributes needed for integration. Further bandwidth reductions will be useful, and the extreme redundancy of XML suggests that compression methods are a natural way forward. The HTTP protocol allows web clients to request byte-level compression of the response by sending the HTTP header "accept-encoding". Web servers can reply with a "content-transfer-encoding" header and a compressed body. The Dazzle server and Bio::Das client have already utilized this feature to reduce their bandwidth requirements. Other compression schema are possible including DAS specific approaches that take advantage of the structure of DAS data.

The World Wide Web Consortium has developed a number of technologies to support XML based systems. A number of these technologies should be considered for future integration into DAS. The Simple Object Access Protocol (SOAP) 1.1 describes a lightweight protocol for the exchange of information in a decentralized, distributed environment. A DAS request may be replaced with a SOAP-style XML-encapsulated document in future versions of this specification. Each annotation is identified by its site-specific database identifier. The combination of this identifier with the server URL and data source produces an feature identifier which is globally unique. Future versions of DAS could utilize this identifier with XPATH and XLINK technologies to permit meta-annotations.

In large part, the continued success of this project will depend on the readiness with which the research community creates annotation sources. To facilitate this, we are working with the BioPerl and BioJava software developer communities [<http://open-bio.org/>] to develop a core set of servers, clients and software modules to support DAS. It is particularly important that the general biological community should be enabled to develop their own DAS annotation servers, without learning XML and Web software development. Easy, well-documented DAS annotation servers that take input data in simple flat file formats and convert it automatically to DAS XML are currently under development.

The DAS specification is under continued development. It does not detail how data source URLs will be published.

Table 3: Summary of DAS URLs For the latest information on the DAS project, see the project website. To learn more about one of the prototype components of DAS, see the appropriate website.

Site	URL
DAS project website	[http://www.biodas.org/]
Current specification	[http://www.biodas.org/documents/spec.html]
Wormbase reference server	[http://www.wormbase.org/db/das/]
Dazzle Java Library	[http://www.biojava.org/dazzle/]
Test server cluster	[http://skynet.wustl.edu/cgi-bin/das/]
Geodesic	[http://www.biodas.org/geodesic/]
Ensembl DAS	[http://www.ensembl.org/das/]
Drosophila DAS	[http://www.fruitfly.org/cgi-bin/das/]

cized. It is anticipated that word of mouth and publications will be the driving forces in user selection. In addition, search engines can be developed to work with the DAS specification.

Conclusions

The DAS specification is already being used in real-world applications. The July 9 2001 release of the Ensembl database of human genome annotations contains support for DAS, including an integrated DAS viewer and multiple annotation servers (M. Pocock, personal communication, 2001). The WormBase DAS server has recently been supplemented by a third party annotation source of cDNA alignments contributed by The Institute for Genome Research, and a prototype DAS reference server for the Drosophila genome is also available, courtesy of the Berkeley Drosophila Genome Project (B. Marshall, personal communication, 2001). Table 3 lists the URLs where one can learn more about the current state of the art in DAS implementations.

Additional material

Additional file

appendix.pdf - The DAS XML Specification

Summary of the current DAS specification, v 1.01.

[<http://www.biomedcentral.com/content/supplementary/1471-2105-2-7-s1.pdf>]

Acknowledgements

The initial ideas for DAS were developed in conversations with LaDeana Hillier of the Washington University Genome Sequencing Center. This work was primarily supported by NIH National Human Genome Research Institute (NHGRI) grant 2-P01-HG00956 for the *Caenorhabditis elegans* genome project, and by a Howard Hughes Medical Institute (HHMI) Predoctoral Fellowship to RDD. We also gratefully acknowledge additional funding support from HHMI and the NIH NHGRI.

References

1. Smith TF: **The history of genetic sequence databases.***Genomics* 1990, **6**:701-707
2. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999.***Nucleic Acids Research* 1999, **27**:49-54
3. Pruitt KD, Katz KS, Sicotte H, Maglott DR: **Introducing RefSeq and LocusLink: curated human genome resources at the NCBI.***Trends Genet* 2000, **16**:44-47
4. Costanzo MC, Crawford ME, Hirschman JE, Kranz JE, Olsen P, Robertson LS, Skrzypek MS, Braun BR, Hopkins KL, Kondu P, et al: **YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information.***Nucleic Acids Res* 2001, **29**:75-9
5. Eeckman FH, Durbin R: **ACeDB and macace.***Methods Cell Biol* 1995, **48**:583-605
6. Waterson R, Sulston J: **The genome of the *Caenorhabditis elegans*.***Proc. Natl. Acad. Sci* 1995, **92**:10836-10840
7. Shoman LM, Grossman E, Powell K, Jamison C, Schatz BR: **The Worm Community System, release 2.0 (WCSr2).***Methods Cell Biol* 1995, **4**:607-625
8. Skupski MP, Booker M, Farmer A, Harpold M, Huang W, Inman J, Kiphart D, Root S, Schilkey F, Schwertfeger J, et al: **The Genome Sequence DataBase: towards an integrated functional genomics resource.***Nucleic Acids Res* 1999, **27**:35-38
9. Letovsky SI, Cottingham RW, Porter CJ, Li PV: **GDB: the human genome database.***Nucleic Acids Res* 1998, **26**:94-99
10. Cuticchia AJ: **Future vision of the GDB human genome database.***Hum Mutat* 2000, **15**:62-67
11. Stein L, Sternberg P, Durbin R, Thierry Mieg J, Spieth J: **WormBase: network access to the genome and biology of *Caenorhabditis elegans*.***Nucleic Acids Res* 2001, **29**:82-86

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMedcentral will be the most significant development for disseminating the results of biomedical research in our lifetime."

Paul Nurse, Director-General, Imperial Cancer Research Fund

Publish with **BMC** and your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours - you keep the copyright



Submit your manuscript here:

<http://www.biomedcentral.com/manuscript/>

editorial@biomedcentral.com

APPENDIX B

Distinct Transcription Profiles of Primary and Secondary Glioblastoma Subgroups

Distinct Transcription Profiles of Primary and Secondary Glioblastoma Subgroups

Cho-Lea Tso,^{1,2,6} William A. Freije,¹ Allen Day,¹ Zugen Chen,¹ Barry Merriman,¹ Ally Perlina,¹ Yohan Lee,¹ Ederlyn Q. Dia,³ Koji Yoshimoto,³ Paul S. Mischel,^{3,6} Linda M. Liao,^{4,6} Timothy F. Cloughesy,^{5,6} and Stanley F. Nelson^{1,6}

Departments of ¹Human Genetics, ²Medicine/Hematology-Oncology, ³Pathology and Laboratory Medicine, ⁴Neurosurgery, and ⁵Neurology, David Geffen School of Medicine, and ⁶Jonsson Comprehensive Cancer Center, University of California at Los Angeles, Los Angeles, California

Abstract

Glioblastomas are invasive and aggressive tumors of the brain, generally considered to arise from glial cells. A subset of these cancers develops from lower-grade gliomas and can thus be clinically classified as “secondary,” whereas some glioblastomas occur with no prior evidence of a lower-grade tumor and can be clinically classified as “primary.” Substantial genetic differences between these groups of glioblastomas have been identified previously. We used large-scale expression analyses to identify glioblastoma-associated genes (GAG) that are associated with a more malignant phenotype via comparison with lower-grade astrocytomas. We have further defined gene expression differences that distinguish primary and secondary glioblastomas. GAGs distinct to primary or secondary tumors provided information on the heterogeneous properties and apparently distinct oncogenic mechanisms of these tumors. Secondary GAGs primarily include mitotic cell cycle components, suggesting the loss of function in prominent cell cycle regulators, whereas primary GAGs highlight genes typical of a stromal response, suggesting the importance of extracellular signaling. Immunohistochemical staining of glioblastoma tissue arrays confirmed expression differences. These data highlight that the development of gene pathway-targeted therapies may need to be specifically tailored to each subtype of glioblastoma. (Cancer Res 2006; 66(1): 159-67)

Introduction

Human solid tumors undergo multiple genetic evolutionary abnormalities as they evolve from normal cells to early-stage tumors to aggressive cancers (1). Chromosome instability that results in the development of both numerical abnormalities (aneuploidy) and structural abnormalities (chromosomal breakage, deletions, and amplification) is especially striking in many types of solid tumors (1, 2). A series of genome-wide chromosomal imbalance analyses and multiparameter cell-based studies suggest that genomic changes that lead to the loss of tumor suppressor gene function usually occur at early stages, whereas the later stages often involve the accumulation of multiple gain-of-function abnormalities that confer on tumors the potential for malignant transformation (3, 4). It

is possible that the malignant end points that are ultimately reached will prove to be shared in common by many types of tumors (5). The identification and functional assessment of genes altered in the process of malignant transformation is essential for understating the mechanism of cancer development and should facilitate the development of more effective treatments.

Infiltrative astrocytic neoplasms are the most common brain tumors of central nervous system in adults. Glioblastoma multiforme (WHO grade IV) remains a devastating disease, with a median survival of <1 year after diagnosis (6). Glioblastomas are defined by histopathologic features of cellular atypia, mitotic figures, necrotic foci with peripheral cellular pseudopalisading, and microvascular hyperplasia that distinguish it from lower-grade astrocytic tumors (7). Two subgroups of glioblastomas have been established based on clinical experience and have been affiliated with distinct genetic mechanisms of tumorigenesis. Secondary glioblastomas develop slowly through progression from low-grade glial tumors (WHO grade II) or anaplastic glial tumors (WHO grade III) and frequently contain mutations in the *p53* gene (~60%), overexpression of platelet-derived growth factor receptors, and loss of heterozygosity (LOH) at 17p, 19q, and 10q (8, 9). In contrast, primary glioblastomas seem to develop rapidly and manifest high-grade lesion from the outset and are genetically characterized by amplification/overexpression of epidermal growth factor receptor (EGFR; ~60%) and mouse double minute 2 (~50%), PTEN mutations, and loss of all or a portion of chromosome 10 (8, 9).

The objective of this study was to identify gain-of-function genes that are associated with acquisition of malignant features of glioblastomas. In addition, we investigated whether clinically defined primary and secondary glioblastoma subgroups use distinct molecular pathways. DNA microarray experiments were done to establish a transcription database for 101 glial brain tumors for which clinical and pathologic features as well as biopsy material were available. Through a series of comparative analyses against lower-grade astrocytomas, we have identified shared and distinct gene categories of transcripts overexpressed in glioblastoma subgroups that are associated with malignant transformation. The distinct glioblastoma-associated genes (GAG) further led to the discovery of stromal/mesenchymal properties in glioblastoma subgroup.

Materials and Methods

Tumor sample and data collection. The patient tumors and normal samples were collected either from autopsies of glioblastoma patients within 24 hours of death or from patients who underwent surgery at University of California at Los Angeles (UCLA) Medical Center. All samples were collected under protocols approved by the UCLA Institutional Review Board. All histopathologic typing and tumor grading

Note: Supplementary data for this article are available at Cancer Research Online (<http://cancerres.aacrjournals.org/>).

Requests for reprints: Stanley F. Nelson, Department of Human Genetics, David Geffen School of Medicine, University of California at Los Angeles, Room 5506, 695 Young Drive South, Los Angeles, CA 90095. E-mail: sfnelson@mednet.ucla.edu.

©2006 American Association for Cancer Research.

doi:10.1158/0008-5472.CAN-05-0077

was done by one neuropathologist (P.S.M.) according to the WHO criteria (10). The subgrouping of glioblastomas was based on clinical presentation. Secondary glioblastomas were called if there was previous pathologic evidence of lower-grade glioblastomas. All tumors without prior evidence of progression from a lower-grade tumor were clinically classified as primary glioblastomas. All samples were snap frozen in liquid nitrogen and stored at -80°C before being processed for the microarray and reverse transcription-PCR (RT-PCR) analyses. Some of the data presented here are derived from a published microarray study (11). Among the data available from the Freije et al. article (11), 43 clinical grade IV glioblastomas were selected for analysis in this article, for which reliable prior treatment and reliable assessment of primary or secondary glioblastoma status were available. Thirty-eight are clinical primary glioblastomas and 5 are clinical secondary glioblastomas. Additional subsets of glioblastomas were selected to address the question of genetic differences between clinically defined subgroups presented here. As secondary glioblastomas are clearly defined, we primarily sought to expand this group. Thus, an additional set of U133A experiments on clinical grade IV glioblastomas included an additional nine clinical secondary glioblastomas and eight primary glioblastomas. For the comparisons with lower grades done here, nine grade III tumors were included from the Freije et al. study (11). An additional 4 grade II tumors are included here as well as 10 normal brain autopsy samples.

Microarray procedures and data analysis. Total RNA was isolated from tumor samples using a TRIzol reagent (Invitrogen Life Technologies, Carlsbad, CA) and was followed by a cleanup on a RNeasy column (Qiagen, Hilden, Germany). cDNA was generated and cRNA probes were generated using standard protocols (12). Aliquots of each sample were hybridized to U133A oligonucleotide microarray (GeneChip Human Genome U133A, Affymetrix, Santa Clara, CA), which represents $\sim 14,500$ human transcripts. The chips were scanned using the GeneArray scanner (Affymetrix). The CEL files generated by the Affymetrix Microarray Suite (MAS 5.0) were converted into DCP files using the DNA-Chip Analyzer (dChip 1.3, <http://biosun1.harvard.edu/complab/dchip/>). The DCP files were globally normalized, and gene expression values were generated using the dChip implementation of perfect match minus mismatch model-based expression index. All group comparisons were done in dChip.

Gene annotation and tissue expression distribution. Functional annotation of genes was obtained from published literature (PubMed) and the GeneReport of the Source database (<http://source.stanford.edu>). The data for normalized expression distribution for tissue type were obtained from UCLA normal tissue transcription database (<http://www.devmod.org/>) established in our laboratory, which unified data from OMIM, SwissProt, LocusLink, Unigene, Genbank, and Gene Card.

Real-time quantitative and semiquantitative RT-PCR. To verify the microarray data, real-time quantitative RT-PCR was carried out with MJ Opticon PCR Analyzer (MJ Research, Inc., Waltham, MA) using SYBR Green PCR Core Reagents (Applied Biosystems, Foster City, CA). All RNA samples extracted from glioblastoma biopsies were digested with DNase I, which is free of RNase, before reverse transcription (Ambion, Inc., Austin, TX). Total RNA ($2\ \mu\text{g}$) was used as a template for RT-PCR. cDNA synthesis was done for one cycle at 50°C for 30 minutes and 94°C for 2 minutes. The PCR reactions were cycled 30 times [50°C for 2 minutes, 95°C for 10 minutes (94°C for 15 seconds, $58-61^{\circ}\text{C}$ for 1 minute, 72°C for 1 minute) $\times 30$ cycles], and the fluorescence was measured at the end of each cycle to construct amplification curves. The melting curve was determined to verify that the PCR product of appropriate size was created. Quantitation of transcripts was calculated based on a titrated standard curve co-run in the same experiment and calibrated with the expression level of a housekeeping gene (β -actin and glyceraldehyde-3-phosphate dehydrogenase). All determinations were done in duplicate. Primer 3 Input (primer3_www.cgi v 0.2) was used to select primers and nonredundant specific primer sequences was verified using BLAT Search Genome (<http://genome.ucsc.edu>) and National Center for Biotechnology Information BLAST (<http://www.ncbi.nlm.nih.gov/blast/Blast.cgi>). The primer sequences and expected size of amplified PCR products are listed at Supplementary Table S1. The specificity of selected PCR products was confirmed by sequencing.

Tissue microarray and immunohistochemistry. A high-density glioblastoma tissue array was constructed consisting of three representative 0.6-mm cores from formalin-fixed, paraffin-embedded tissue blocks from each of 60 primary glioblastomas, 16 secondary glioblastomas, and 15 normal brain tissues. Sections were stained with polyclonal antibody to YKL-40/cartilage glycoprotein-39 (1:200, Quidel Corp., San Diego, CA) or control antibody for overnight at 4°C . Subsequent immunodetection was done using Vectastain ABC Standard kit (Vector Laboratories, Burlingame, CA) and Vector NovaRED (Vector Laboratories). Staining intensity was scored by a neuropathologist (K.Y.) based on a scale of 0 to 2 in which negatively stained specimens were graded 0, weakly positive samples were graded 1, and strongly positive spots were graded 2 (13). The significance of differences in the incidence of YKL-40/cartilage glycoprotein-39 expression in glioblastoma subgroups and normal brain tissue was calculated using one-tailed, two-proportion Z-test.

Results

Identification of GAGs. We analyzed the expression of $\sim 14,500$ well-characterized human genes (22,283 probe sets) using Affymetrix GeneChip U133A for 102 brain tumors and normal brain tissues consisting 46 grade IV primary glioblastomas, 14 secondary glioblastomas, 13 astrocytomas (4 grade II and 9 grade III), 19 oligodendromas (8 grade II and 11 grade III), and 10 normal brain tissues. To identify gain-of-function genes correlated with the malignant features of glioblastomas, we compared the mean level of normalized transcript levels in each of the two clinically defined glioblastoma groups versus the grade II and III astrocytomas. Probe set signals on the U133A array that were ≥ 2.5 -fold in each glioblastoma group versus the astrocytoma group and with a pairwise t test ($P < 0.05$) were selected. In addition, to avoid inclusion of low-level and unreliable signals, the higher signal needed to exceed 100 and be called present by MAS 5.0 in $>20\%$ of the samples. Genes that were identified using these two filtering criteria were designated as either primary or secondary GAGs. Expression patterns across a set of 16 samples were verified by real-time quantitative RT-PCR analysis of eight selected genes with an average correlation of 0.88 (0.77-0.94).

Shared GAGs reflect common characteristics of hyperproliferation, hypervascularity, and apoptotic resistance in both glioblastoma subgroups. When compared with lower-grade astrocytomas under the defined comparison criteria, 36 GAGs were identified from the secondary glioblastoma group comparison and 73 GAGs were identified from the primary glioblastoma group comparison (data not shown). Because secondary glioblastomas cannot be distinguished from primary glioblastomas histopathologically, we anticipated identifying common genes underlying the phenotypic similarity. Indeed, 15 GAGs were identified in both pairwise comparisons (Fig. 1A; Table 1). These 15 genes share some functional categorization and are involved in mitosis and extracellular response-associated genes. However, although commonly overexpressed in both types of glioblastomas, there were quantitative differences in expression levels between secondary glioblastomas and primary glioblastomas. The secondary glioblastomas showed higher expression in several mitotic cell cycle-associated genes (*RRMP*, *TYMS*, *TOP2A*, *CENPE*, *HEC*, *CDC2*, *TOPK*, and *ANKT*), whereas primary glioblastomas exhibited higher expression of several extracellular response-associated genes (*ADM*, *VEGF*, *FCGBP*, and *COL4A1/COL4A2*). The most highly expressed gene in the secondary glioblastoma subgroup relative to the lower-grade tumors was hepatocyte growth factor receptor (*MET*), which was also induced in the primary glioblastomas but to a lesser degree. Conversely, the most overexpressed gene in the

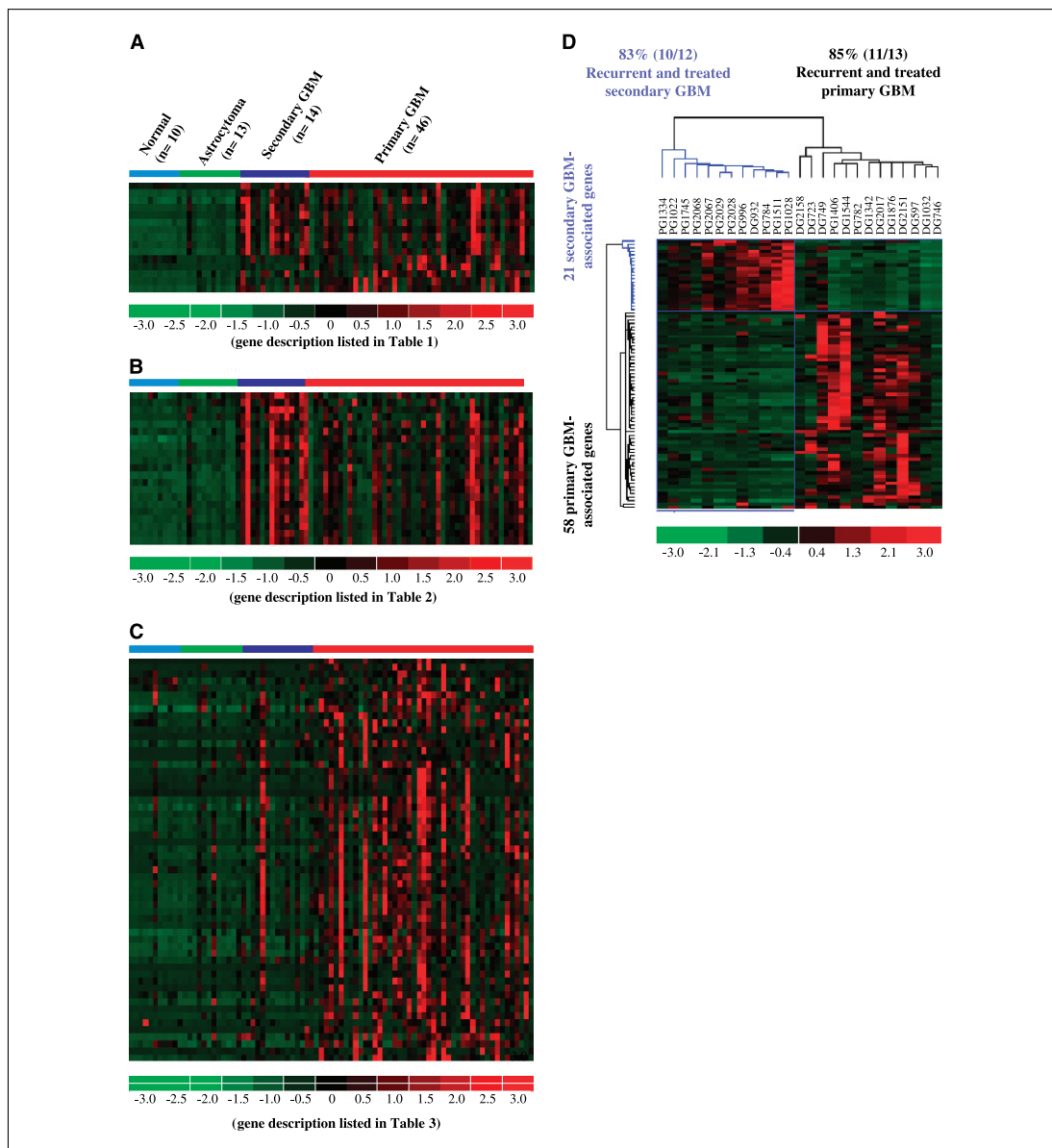


Figure 1. GAGs overexpressed in glioblastomas relative to lower-grade gliomas. All plots show normalized gene expression values converted into a heat map. The \log_2 of the fold difference is indicated by the heat map scale at the *bottom*. Each column is an individual tissue or tumor sample organized into histologic groups defined at the *top*. Each row is a single probe set measurement of transcript abundance for an individual gene. The genes are listed in the same order from *top* to *bottom* as the corresponding tables for each of the four lists. All genes were filtered to select transcripts with ≥ 2.5 -fold expression in the respective glioblastoma (GBM) group relative lower-grade astrocytomas ($P < 0.05$, t test). **A**, shared GAGs overexpressed in glioblastomas: 15 shared GAGs were identified from the intersection of the comparisons of primary glioblastomas versus lower-grade astrocytomas and secondary glioblastomas versus lower-grade astrocytomas. **B**, GAGs overexpressed uniquely in secondary glioblastomas: 21 secondary GAGs were defined as being uniquely detected with a >2.5 -fold overexpression in the secondary glioblastoma group compared with the lower-grade astrocytomas and not overexpressed within the primary glioblastoma group using the same criteria. **C**, GAGs overexpressed uniquely in primary glioblastomas: 58 primary GAGs were defined as overexpressed 2.5-fold relative to lower-grade astrocytomas and not detected in the secondary glioblastomas comparison using the same criteria. **D**, unsupervised sample clustering of primary and secondary glioblastomas that are recurrent and had treatment using 21 secondary glioblastomas (Table 2) and 58 primary GAGs (Table 3).

Table 1. Shared GAGs expressed at higher levels in both primary and secondary glioblastomas compared with astrocytomas

Gene	Symbol	Accession no.	Chromosome	Secondary glioblastomas		Primary glioblastomas	
				Fold	P	Fold	P
<i>Met proto-oncogene</i>	<i>MET</i>	BG170541	7q31	8.48	0.017936	4.21	0.045314
<i>Ribonucleotide reductase M2 polypeptide</i>	<i>RRMP</i>	NM_001034.1	2p25	3.46	0.00015	3.52	0
<i>Thymidylate synthetase</i>	<i>TYMS</i>	NM_001071.1	18p11	2.77	0.00084	2.5	0.000025
<i>Topoisomerase (DNA) IIα, 170 kDa</i>	<i>TOP2A</i>	NM_001067.1	17q21	4.24	0.006921	2.65	0.001064
<i>Centromere protein F, 350/400 kDa</i>	<i>CENPF</i>	NM_005196.1	1q32	6.41	0.001846	2.98	0.000407
<i>Highly expressed in cancer</i>	<i>HEC</i>	NM_006101.1	18p11	3.57	0.003301	2.83	0
<i>Cell division cycle 2, G₁-S and G₂-M</i>	<i>CDC2</i>	AL524035	10q21	4	0.008511	2.5	0.000008
<i>Hypothetical protein</i>	<i>FLJ23468</i>	NM_024629.1	4q35	3.63	0.00146	2.5	0.000023
<i>T-LAK cell-originated protein kinase</i>	<i>TOPK</i>	NM_018492.1	8p21	3.64	0.009279	2.68	0.00019
<i>Nucleolar protein ANKT</i>	<i>ANKT</i>	NM_016359.1	15q14	3.44	0.000624	2.56	0.000004
<i>Adrenomedullin</i>	<i>ADM</i>	NM_001124.1	11p15	3.39	0.020445	6.76	0.00003
<i>Vascular endothelial growth factor</i>	<i>VEGF</i>	M27281.1	6p12	3.12	0.004844	6.1	0.000001
<i>Fc fragment of IgG-binding protein I</i>	<i>FCGBP</i>	NM_003890.1	19q13	2.63	0.036568	3.5	0.000005
<i>Collagen type IV, α1</i>	<i>COL4A1</i>	NM_001845.1	13q34	3	0.002538	3.6	0.000003
<i>Collagen type IV, α2</i>	<i>COL4A2</i>	AK025912.1	13q34	2.65	0.005812	3.2	0.000003

NOTE: Analysis was based on a cutoff of a 2.5-fold increase in relative expression ($P < 0.05$) in glioblastomas compared with astrocytomas.

primary glioblastomas compared with the lower-grade tumors was *ADM*, which was induced in the secondary tumors but to a lesser degree. Both of these genes are well-characterized tumor survival factors, which have been shown to play a critical role in cancer cell division, antiapoptosis, cell migration, and tumor neovascularization (14, 15). Activation of the receptor tyrosine kinase *Met* promotes cell survival by activating phosphatidylinositol 3-kinase signaling cascade (16). Additionally, *Met* sequesters *Fas*, circumventing programmed cell death (17). *ADM*, which is up-regulated during hypoxic insult, promotes the growth and migration of endothelial cells (18), and has also been implicated as a potential immune suppressor substance (15). The detection of up-regulated *VEGF* transcripts likely reflects the hypoxia, which promotes an angiogenic response (19).

Distinct GAGs identified in secondary glioblastomas reflect aggressive cell cycle. Twenty-one distinct GAGs were overexpressed only in the secondary glioblastomas. Remarkably, all 21 genes are associated with mitotic cell cycle (Fig. 1B; Table 2). More specifically, these genes are involved in control of cell cycle (*CKS2*, *CDKN3*, *GAS1*, *CCNB1*, *UBE2C*, and *FOXM1*), DNA synthesis and repair (*ECT2* and *PIR51*), cytokinesis and movements of spindle and chromosomes (*RAMP*, *PRC1*, *TMSNB*, *KIF2A*, *KIF14*, and *KIF20A*), DNA bending (*HMGB2*), kinetochore function (*ZWINT*), chromatid separation and regulation of TP53 (*PTTG1*), and mitotic chromosome condensation (*HCAP-G*). Among these genes, *FOXM1*, which is a transcription factor that regulates the expression of transcription network of genes that are essential for DNA replication and mitosis, showed the highest fold increase (20). We then viewed the gene expression distribution among various types of tissue for these 21 genes using our human tissue transcription database (<http://www.dev.gmod.org/>). The results indicated that most transcripts (19 of 21) were highly expressed in proliferative tissues, fetal livers, and testis (germ cells; see Supplementary Fig. S1), which are indicative of the genes known expression in mitotically active cells. A similar set of overexpressed genes were identified when comparative analyses were done

between secondary glioblastomas and the group of oligodendrocytes (grade II and III; data not shown).

Distinct GAGs identified in primary glioblastomas reflect a tumor cell stromal response. Fifty-eight GAGs that are overexpressed only in primary glioblastomas when compared with astrocytomas reflect the processes of the host-tumor interaction that promote the well-recognized invasive phenotype of glioblastomas (Fig. 1C; Table 3). The annotation of the selected genes reflects the *in situ* stromal response of the cancer cells. The list includes genes that are associated with inflammation, coagulation, immune/complement responses (*SERPINA1/SERPINA3*, *SERPINE1*, *PTX3*, *C5R1*, *FCGR3B*, *CEBPD*, and *TIMPI1*), angiogenesis (*IL-8*, *CA1*, and *CA2*), extracellular matrix (ECM) remodeling (*COL5A1*, *COL6A2*, *MMP-9*, and *C1R*), and status of hypoxia/angiogenesis (*HIP-2*). Moreover, genes that may function as anti-oxidants or promoters for antiapoptotic activities (*CAIII*, *SOD2*, *DPYD*, *NNMT*, and *UPP1*) were identified and are potential predictors for the chemoradiation-resistant phenotypes. The presence of two transforming growth factor- β (TGF- β) target genes (*TGFBI* and *TAGLN*) suggested that TGF- β signaling is involved in malignant progression, whereas two stress-responsive genes (*HO* and *SLC16A3*) reflect inflammatory insults and perhaps a glycolysis shift. The overexpression of monocyte chemotactic factor (*CCL*) corresponded to a group of genes reflected influx of tumor-associated macrophage (*CD14*, *CD163*, *STAB1*, *Z391G*, *LYZ*, and *IFI30*), indicating a more pronounced inflammatory component of primary glioblastomas relative to secondary glioblastomas.

A series of genes that are highly expressed in mesenchymal tissues but not neural or glial cells were identified. These include genes that are typically expressed in tissues like bone, cartilage, tendon, ligament, fat, and muscle (*CHI3L1*, *CHI3L2*, *GPNMB*, *LOX*, *TIA-2*, *COLV/VI*, *BGN*, *MEOX2*, *CAIII*, and *TAGLN*). In particular, *CAIII* and *MEOX2* exceed a 10-fold increase when compared with astrocytomas. *CAIII* functions as an oxygen radical scavenger and hence protects cells from oxidative stress (21), whereas *MEOX2* has a role in mesoderm induction and is an important regulator of vertebrate

limb myogenesis (22). Notably, 11 primary GAGs are located at chromosome 7 (*MET*/7q31, *HIF-2*/7q32, *CAV1*/7q31, *CAV2*/7q31, *SERPINE1*/7q21, *PBEF*/7q22, *GNPMB*/7p15, *UPP1*/7, *MEOX2*/7p22, *EGFR*/7p12, and *SEC61G*/7p11; Tables 1 and 3; ref. 23) and 8 of them were reported to be associated with Akt phosphorylation, anti-apoptosis, hypoxia, angiogenesis, and coagulation, which corresponds to the reported amplification of chromosome 7 in primary glioblastomas (24–28). Detection of increased *IGFBP2* transcripts verified previous reports (29). Analysis of the human tissue transcription database confirmed the preferential expression of these GAGs in multiple stromal/mesenchymal tissue types, including cartilage, cultured chondrocytes, muscle, endothelium (aorta), bone marrow, and monocytes/macrophage (see Supplementary Fig. S3). Similarly, comparable transcription profiles of GAGs were obtained when comparative analysis was done against the group of oligodendromas, which showed a dominant group of overexpressed genes that are associated with mesenchymal cells (data not shown).

Does prior treatment of secondary glioblastomas account for the differences between secondary and primary glioblastomas? To rule out the possibility that the distinct glioblastoma progression-associated genes identified between two subgroups are due to selection pressure (e.g., radiation or chemotherapy), we conducted clustering based analysis of a set of primary glioblastoma ($n = 13$) and secondary glioblastoma ($n = 12$) samples that were recurrent and had been treated before tumor sampling in two different ways. The 25 tumor specimens meeting criteria above were hierarchically clustered using normalized data for all 79 defined type-specific GAGs (21 secondary GAGs and 58 primary GAGs). The predominant subdivision in the tumors is on the basis of primary versus secondary definition: 85% (11 of 13) of the treated primary glioblastomas were clustered by overexpression of

virtually all 58 primary GAGs, whereas 83% (10 of 12) of the treated secondary glioblastomas were clustered by overexpression of 21 secondary GAGs (Fig. 1D). These analyses were restricted to the already defined primary and secondary GAGs and indicate that both tumor groups, regardless of prior treatment, cluster within their clinical grouping based on gene expression of the selected GAGs. Thus, prior treatment is not disrupting this identified gene expression signature of primary and secondary glioblastomas nor is it driving the selection of the genes.

Differential expression of cartilage glycoprotein-39 (*CHI3L1*) in glioblastoma subgroups. Based on our analysis, we consistently observed a significant up-regulation of *CHI3L1* gene expression in primary glioblastomas when compared with low-grade astrocytomas or secondary glioblastomas. To verify this finding, a tissue microarray consisting of tumor cores and matching normal brain counterparts from 60 primary glioblastomas and 16 secondary glioblastomas was constructed and immunohistochemically stained using commercially available antibody to cartilage glycoprotein-39 (YKL-40). *CHI3L1* expression was significantly more frequently detected in the clinically defined primary glioblastoma samples compared with secondary glioblastomas and normal brain. Forty-two percent (25 of 60) of the primary glioblastomas stained positively with average intensities of 1.7 ± 0.46 , whereas 12.5% (2 of 16; $P = 0.0152$) and 6.7% (1 of 15; $P = 0.0054$) were positively stained in secondary glioblastomas and normal brain, respectively. All three positive stains in secondary glioblastomas and a normal brain specimen were weak (Fig. 2).

Discussion

In this study, we used a large-scale gene expression analysis to further characterize clinical subgroups of glioblastomas. We aimed

Table 2. Distinct GAGs expressed at higher levels in secondary glioblastomas compared with astrocytomas

Gene	Symbol	Accession	Chromosome	Fold change	P
<i>Homo sapiens mRNA: cDNA DKFZp564F112</i>		AI049987.1		2.51	0.000299
<i>Growth arrest-specific 1</i>	<i>GAS1</i>	NM_002048.1	9q21	2.58	0.004934
<i>RAD51-interacting protein</i>	<i>PIR51</i>	BE966146	12p13	2.81	0.000739
<i>Thymosin β, identified in neuroblastoma cells</i>	<i>TMSNB</i>	NM_021992.1	Xq21	3.26	0.000834
<i>KIAA0101 gene product</i>		NM_014736.1	15q22	2.62	0.000563
<i>Cyclin-dependent kinase inhibitor 3</i>	<i>CDKN3</i>	AF213033.1	14q22	2.53	0.004978
<i>High-mobility group box 2</i>	<i>HMGB2</i>	BC000903.1	4q31	2.75	0.000438
<i>Ubiquitin-conjugating enzyme E2C</i>	<i>UBE2C</i>	NM_007019.1	20q13	2.92	0.000382
<i>Retinoic acid-regulated nuclear matrix-associated protein</i>	<i>RAMP</i>	NM_016448.1	1	2.94	0.014477
<i>CDC28 protein kinase regulatory subunit 2</i>	<i>CKS2</i>	NM_001827.1	9q22	3.08	0.00485
<i>Epithelial cell transforming sequence 2 oncogene</i>	<i>ECT2</i>	NM_018098.1	3q25	2.97	0.001624
<i>Kinesin family member 20A</i>	<i>KIF20A</i>	NM_005733.1	5q31	3.24	0.000546
<i>Cyclin B1</i>	<i>CCNB1</i>	BE407516	5q12	2.48	0.00095
<i>Pituitary tumor-transforming 1</i>	<i>PTTG1</i>	NM_004219.2	5q35	2.8	0.000862
<i>Chromosome condensation protein G</i>	<i>HCAP-G</i>	NM_022346.1	4p16	3.37	0.003616
<i>ZW10 interactor</i>	<i>ZWINT</i>	NM_007057.1	10q21	2.68	0.006086
<i>asp (abnormal spindle)-like</i>	<i>ASPM</i>	NM_018123.1	1q31	3.24	0.003574
<i>Protein regulator of cytokinesis 1</i>	<i>PRC1</i>	NM_003981.1	15q26	3.45	0.0026
<i>Kinesin family member 14</i>	<i>KIF14</i>	NM_014875.1	1pter-q31.3	2.5	0.006231
<i>Forkhead box M1</i>	<i>FOXM1</i>	NM_021953.1	12p13	3.8	0.000976
<i>Kinesin family member 4A</i>	<i>KIF4A</i>	NM_012310.2	Xq13	2.83	0.00078

NOTE: Analysis was based on a cutoff of a 2.5-fold increase in relative expression ($P < 0.05$) in secondary glioblastomas compared with astrocytomas.

Table 3. Distinct GAGs expressed at higher levels in primary glioblastomas compared with astrocytomas

Gene	Symbol	Accession	Chromosome	Fold change	P
Carbonic anhydrase III, muscle specific	CAIII	NM_005181.2	9q13	11.46	0.007594
Lactotransferrin	LTF	NM_002343.1	3q21	3.8	0.00565
Human clone 137308 mRNA, partial cds.		AU134977		3.02	0.005225
Solute carrier family 2 (facilitated glucose transporter), member 3	SLC2	NM_006931.1	12p13	2.65	0.000054
Hypoxia-inducible protein 2	HIF-2	NM_013332.1	7q32	2.56	0.001092
Guanylate-binding protein 1, IFN-inducible, 67 kDa	GBP1	NM_002053.1	1p22	2.71	0.000028
Chitinase 3-like 2	CHI3L2	U58515.1	1p13	2.68	0.000827
Proteinase inhibitor, clade A (antitrypsin), member 3	SERPINA3	NM_001085.2	14q32	2.64	0.000716
Heat shock 70-kDa protein 6	HSP70B	NM_002155.1	1	3.75	0.000047
Caveolin 2	CV2	NM_001233.1	7q31	2.46	0.000043
Heme oxygenase (decycling) 1	HO	NM_002133.1	22q13	3.2	0.000088
Matrix metalloproteinase-9 (92-kDa type IV collagenase)	MMP-9	NM_004994.1	20q11	4.38	0.002731
Biglycan	BGN	AA845258	Xq28	2.78	0.000053
Collagen type VI, $\alpha 2$	COL6A2	AY029208.1	21q22	2.93	0.019421
Collagen type V, $\alpha 1$	COL5A1	AI983428	9q34	2.46	0.03049
Fc fragment of IgG, low-affinity IIIb, receptor for (CD16)	FCGR3B	NM_000570.1	1q23	2.61	0.000285
Chromosome 8 open reading frame 4	C8orf4	NM_020130.1	8p11	4.05	0.000735
Chemokine (C-C motif) ligand 2	CCL	S69738.1	17q11	2.66	0.002211
Interleukin-8	IL-8	NM_000584.1	4q13	4.83	0.00097
Interleukin-8 COOH-terminal variant (IL8) mRNA, complete cds.		AF043337.1		3	0.01931
Superoxide dismutase 2, mitochondrial	SOD2	W46388	6q25	3.23	0
H. sapiens, clone IMAGE:4711494, mRNA		BF575213		2.53	0
Pre-B-cell colony-enhancing factor	PBEF	NM_005746.1	7q22	3.76	0.000003
Complement component 1, r subcomponent	C1R	AL573058	12p13	3.13	0.000044
CCAAT/enhancer-binding protein, δ	CEBPD	AV655640	8p11	2.54	0.00003
Proteinase inhibitor, clade A (antitrypsin), member 1	SERPINA1	NM_000295.1	14q32	2.62	0.000155
Nicotinamide N-methyltransferase	NNMT	NM_006169.1	11q23	8.71	0.000155
Glycoprotein (transmembrane) nmb	GNPMB	NM_002510.1	7p15	2.71	0.002098
Dihydropyrimidine dehydrogenase	DPYD	NM_000110.2	1p22	2.94	0.000015
Complement component 5 receptor 1 (C5a ligand)	C5R1	NM_001736.1	19q13	3.57	0.000125
S100 calcium-binding protein A8 (calgranulin A)	S100A8	NM_002964.2	1q21	3.84	0.000366
Lysozyme (renal amyloidosis)	LYZ	AV711904	12q14	2.5	0.0012696
IFN- γ -inducible protein 30	IFI30	NM_006332.1	19p13	3.09	0.00002
CD14 antigen	CD14	NM_000591.1	5q31	2.66	0.000031
Ig superfamily protein	Z391G	NM_007268.1	Xq12	2.61	0.000006
CD163 antigen	CD163	NM_004244.1	12p13	3.86	0.000011
Stabilin 1	STAB1	NM_015136.1	3p21	2.63	0.000073
Solute carrier family 16 (monocarboxylic acid transporters)	SLC16A3	NM_004207.1	17q25	3.27	0.000001
Transforming growth factor- β induced, 68 kDa	TGVB1	NM_000358.1	5q31	2.71	0.00091
Fibronectin 1	FN1	BC005858.1	2q34	2.52	0.000002
Epithelial membrane protein 3 S100 calcium-binding protein A11 (Calgizzarin)	EMP3	NM_001425.1	19q13	2.73	0.000175
	S100A11	NM_005620.1	1q21	2.64	0.00007
Tissue inhibitor of metalloproteinase 1 (collagenase inhibitor)	TIMP1	NM_003254.1	Xp11	2.58	0.002723
Caveolin 1, caveolae protein, 22 kDa	CAV1	AU147399	7q31	3.01	0.000025
Lysyl oxidase	LOX	L16895	5q23	5.85	0.000442
Proteinase inhibitor, clade E (plasminogen activator inhibitor type 1)	SERPINE1	NM_000602.1	7q21	3.15	0.001055
Transgelin	TAGLN	NM_003186.2	11q23	3.57	0.000112
Thrombospondin 1	THBS1	NM_003246.1	5q15	3.18	0.015815
Uridine phosphorylase	UPP1	NM_003364.1	7	2.53	0.000003
Chitinase 3-like 1 (cartilage glycoprotein-39)	CHI3L1	M80927.1	1q32	4.17	0.000726
Pentaxin-related gene, rapidly induced by interleukin-1 β	PTX3	NM_002852.1	3q25	5.55	0.000292
Lung type I cell membrane-associated glycoprotein	TIA-2	BF337209	1p36	3.07	0.000054
Short stature homeobox 2	SHOX2	AF022654.1	3q25	2.59	0.000151
Mesenchyme homeobox 2 (growth arrest-specific homeobox)	MEOX2	NM_005924.1	7p22	10.31	0.000219
Insulin-like growth factor-binding protein 2, 36 kDa	IGFBP2	NM_000597.1	2q33	2.49	0.007258
Fatty acid-binding protein 5 (psoriasis-associated)	FABP5	NM_001444.1	8q21	2.86	0.00001
Epidermal growth factor receptor	EGFR	NM_005228.1	7p12	2.66	0.003366
Sec61 γ	SEC61G	NM_014302.1	7p11	2.49	0.000013

NOTE: Analysis was based on a cutoff of a 2.5-fold increase in relative expression ($P < 0.05$) in primary glioblastomas compared with astrocytomas.

to elucidate molecular pathway correlates of observed clinical features of primary and secondary glioblastomas that distinguish them from lower-grade astrocytomas. Our analytic strategy extracted lists of genes that are expressed in glioblastomas but not the lower-grade tumors or normal brain tissue. The comparison to lower-grade astrocytomas was done to attempt to select for genes, which are specific to the end malignant transformation into the highly invasive glioblastomas. The described short lists of genes are descriptive and partially explanatory of known tumor behavior, pathology, and resistance to therapy and provide an insight into how the deregulation of multigene networks leads to tumor malignancy. Moreover, the GAGs identified that are uniquely expressed in primary and secondary glioblastomas provide new leads into diverse mechanisms and properties underlying distinct transformation events or perhaps distinct cells of origin of glioblastoma subgroups. These data extend and complement recent studies using two-dimensional gel analysis, which indicated that clinical and genetic differences in primary and secondary glioblastomas could be recognized at the protein level (30).

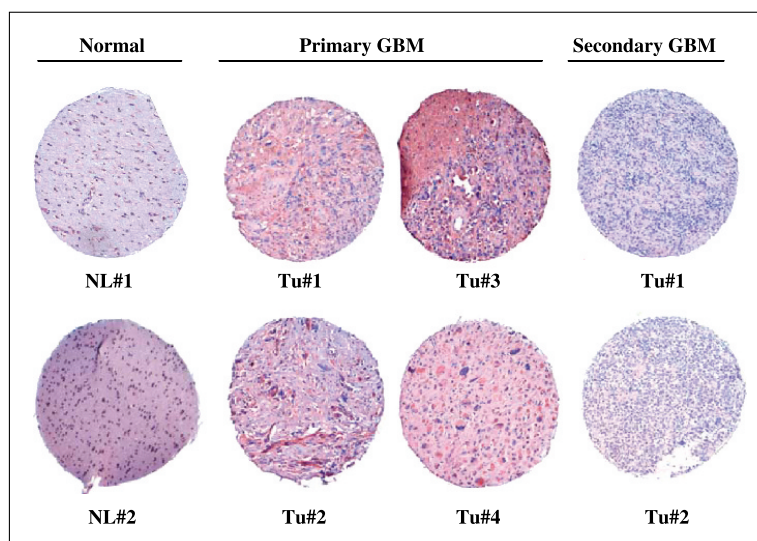
In our study, we used a clinical definition of secondary designation restricted to those tumors with clear prior evidence of a lower-grade tumor. In contrast, the clinical definition of primary glioblastomas is more tenuously based on lack of previous evidence of lower-grade tumor. One would expect that at least a subset of clinically defined primary glioblastomas had a lower-grade initial lesion that progressed asymptotically and would biologically resemble the secondary glioblastoma group. Indeed, there is a strong trend toward the clinical secondary glioblastomas having similar coexpression of secondary GAGs (10 of 14), whereas 24 of 45 of the clinical primary glioblastomas have a similar overexpression of the primary GAGs and instead have a pattern of expression more similar to the overall secondary glioblastoma group. Thus, our data suggest that as many as half of clinically defined primary glioblastomas have the genetic signature of secondary glioblastomas and thus may develop from

lower-grade tumors *in vivo* that are clinically unrecognized. However, this group of primary glioblastomas, which group with the secondary tumors, occur in individuals at a mean age (51 years) range that is typical of primary glioblastomas with the primary glioblastoma signature (54 years) as opposed to the clinically defined secondary glioblastomas (39 years).

Some differences in the frequency of distinct genetic alterations in primary and secondary glioblastomas have been well described (31). These known genomic alterations may partly explain our detection of differential transcription profiles in glioblastoma subgroups. For instance, the large number of genes on chromosome 7, which are up-regulated in primary glioblastomas, may be indicative of chromosome 7 amplification. Further, our data support the notion that mutation or dysfunction of prominent cell cycle regulators is a major mechanism for the malignant transformation in secondary glioblastomas. Deletion of chromosome 17 and/or mutation in p53 have been reported in ~60% of secondary glioblastomas but <10% of primary glioblastomas (32). The p53 tumor suppressor is highly interconnected and mutation of p53 severely disrupts normal cell cycle progression through the modulation of genes that mediate the arrest of cells in the G₁ or G₂ phase (33, 34). p53 mutations, however, are usually found in the low-grade lesions of astrocytomas, indicating that p53 alteration is an early event in astrocytoma progression. A recent report further suggests that retinoblastoma tumor susceptibility gene (*Rb*) may function in the maintenance of chromosome stability by influencing mitotic progression, faithful chromosome segregation, and structural remodeling of mitotic chromosomes (35). LOH in the region containing the *Rb* gene is found in high-grade astrocytomas but not in low-grade astrocytomas, suggesting that disruption of *Rb* is important for the continued malignant transformation to glioblastomas (36).

Primary GAGs strongly reflected a desmoplastic-like phenotype with deposition of abundant collagen. Several markers that implicated the influx of tumor-associated macrophages and lymphocytes were identified. This observation implicates that

Figure 2. Expression of cartilage glycoprotein-39 (YKL-40) in glioblastoma subgroups. Representative immunohistochemical stainings of YKL-40 in clinical glioblastoma subgroups and normal brain. A subset of the images from a tissue array are shown. Each core is 0.06 mm across. Strong positivity of YKL-40 antibody staining was detected in four primary glioblastomas; negative/weak staining was detected in two secondary glioblastomas and two normal brain cores shown.



stromal cells likely participate in promoting such a wound-like phenotype in glioblastoma tumor *in situ*. This link between gene expression signature of fibroblast serum response and cancer progression has been reported (37). Degradation of the ECM by matrix metalloproteinase (MMP) is required in endothelial cell migration, organization, and angiogenesis. *THBS1* promotes tumor invasion of collagens by enhanced MMP-9 production (38) and *IL-8* promotes inflammation, complement response, and coagulation (39). Tumor progression is commonly associated with dysregulation of thrombotic and fibrinolytic processes. The up-regulation of transcripts of inhibitors for proteinase, plasmin, and MMP (*SERPINA1*, *SERPINA3*, *PAI-1*, and *TIMP1*) may function to sustain thrombus formation and prevent fibrinolysis, subsequently inducing coagulative necrosis and hypoxia within pseudopalisades (31, 40). Subsequently, hypoxia induces angiogenesis (*HIF-2*, *MET*, *ADM*, *VEGF*, and *IL-8*) and pseudopalisading cell migration that escape from necrotic zone, thus favoring tumor outgrowth (40).

It seems that the molecular distinction between primary and secondary glioblastomas is not due to a higher frequency of prior treatment in the secondary glioblastoma group but rather reflect genetic differences between the two mechanisms for glioblastoma oncogenesis, which are maintained within the tumors. This conclusion was supported by the comparison of the two glioblastoma subgroups that are both recurrent and had treatment before the biopsy. In addition, in our study, patients with secondary glioblastomas have average younger age (37 ± 9 years) than the patients with primary glioblastomas (51 ± 15 years). We previously published results that included some of the clinical primary glioblastoma samples presented in the analyses presented here. Although most of those samples were clinically primary glioblastomas, there was strong heterogeneity within the glioblastomas. In our previous expression study, which included 63 glioblastomas (11), most glioblastoma samples grouped within two hierarchical clusters; one cluster is defined by overexpression of genes involved in mitosis (HC2A) and the other one is defined by overexpression of ECM components and regulators (HC2B). Thus, the previously defined HC2B group with ECM overexpression is greatly enriched in primary glioblastomas.

In summary, our study explored a complexity of molecular pathways and networks that drives the survival, progression, and invasion of glioblastomas. Several key genes on the list of GAGs corresponded well to previous reports (11, 13, 29, 41–43). Moreover, these data support the concept that the interplay between glioblastoma-derived bone/cartilage-associated factors and tumor-associated stromal cells (fibroblasts, endothelium, and inflammatory cells) plays a key role in the malignant aggressiveness of primary glioblastomas. It has been reported that osteopontin, osteoactivin (*GPVMB*), and osteonectin stimulate tumor invasion and secretion of urokinase-type plasminogen activator through the activation of EGFR, Met, and Akt signaling pathways (44–47). Additional immunostainings of tissue arrays (osteonectin and tenascin C; data not shown) further confirm the mesenchymal properties in primary glioblastomas. Based on our previous study (11) and current observation of a distinct set of genes being expressed in glioblastomas are associated with mesenchymal cells, we have pursued the analysis of primary glioblastoma-derived cell cultures and have indicated that these tumor cell explants possess stem-like properties and can be differentiated into multiple mesenchymal cell lineages,⁷ further highlighting profound differences in glioblastoma subtype and opening up the question regarding the cellular origin of a subset of glioblastomas.

Acknowledgments

Received 1/12/2005; revised 8/15/2005; accepted 9/28/2005.

Grant support: National Cancer Institute grant U01CA88173, Accelerate Brain Cancer Cure, Henry Singleton Brain Tumor Program, Art of the Brain, UCLA DNA Microarray Facility, National Institute of Neurological Diseases and Stroke, National Institute of Mental Health Microarray Consortium grant U24NS43562, Women's Reproductive Health Research Center grant 5K12HD001281 (W.A. Freije), and Integrated Graduate Education and Research Traineeship grant (A. Day).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

We thank the patients who participated in this study.

⁷ Tso et al, unpublished data.

References

- Shackney SE, Shankey TV. Common patterns of genetic evolution in human solid tumors [review]. *Cytometry* 1997;29:1–27.
- Cai WW, Mao JH, Chow CW, et al. Genome-wide detection of chromosomal imbalances in tumors using BAC microarrays. *Nat Biotechnol* 2002;20:393–6.
- Zhang LH, Qin LX, Ma ZC, et al. Allelic imbalance regions on chromosomes 8p, 17p and 19p related to metastasis of hepatocellular carcinoma: comparison between matched primary and metastatic lesions in 22 patients by genome-wide microsatellite analysis. *Cancer Res Clin Oncol* 2003;129:279–86.
- Smith CA, Pollice AA, Gu LP, et al. Correlations among p53, Her-2/neu, and ras overexpression and aneuploidy by multiparameter flow cytometry in human breast cancer: evidence for a common phenotypic evolutionary pattern in infiltrating ductal carcinomas. *Clin Cancer Res* 2000;6:112–26.
- Morales CP, Souza RF, Spechler SJ. The hallmarks of cancer. *Cell* 2000;100:57–70.
- Scott JN, Rewcastle NB, Brasher PM, et al. Long-term glioblastoma multiforme survivors: a population-based study. *Can J Neurol Sci* 1998;25:197–201.
- Prat DJ, Van Meir EG. Vaso-occlusive and prothrombotic mechanisms associated with tumor hypoxia, necrosis, and accelerated growth in glioblastoma [review]. *Lab Invest* 2004;84:397–405.
- Kleihues P, Ohgaki H. Primary and secondary glioblastomas: from concept to clinical diagnosis [review]. *Neuro-oncol* 1999;1:44–51.
- Maruno M, Ninomiya H, Ghulam Muhammad AK, et al. Whole-genome analysis of human astrocytic tumors by comparative genomic hybridization. *Brain Tumor Pathol* 2000;7:21–7.
- Kleihues P, Louis DN, Scheithauer BW, et al. The WHO classification of tumors of the nervous system [review]. *J Neuropathol Exp Neurol* 2002;61:215–25.
- Freije WA, Castro-Vargas FE, Fang Z, et al. Gene expression profiling of gliomas strongly predicts survival. *Cancer Res* 2004;64:6503–10.
- Shai R, Shi T, Kremen TJ, et al. Gene expression profiling identifies molecular subtypes of gliomas. *Oncogene* 2003;22:4918–23.
- Choe G, Horvath S, Cloughesy TF, et al. Analysis of the phosphatidylinositol 3'-kinase signaling pathway in glioblastoma patients *in vivo*. *Cancer Res* 2003;63:2742–6.
- To CT, Tsao MS. The roles of hepatocyte growth factor/scatter factor and met receptor in human cancers [review]. *Oncol Rep* 1998;5:1013–24.
- Cuttitta F, Pio R, Garayoa M, et al. Adrenomedullin functions as an important tumor survival factor in human carcinogenesis [review]. *Microsc Res Tech* 2002;57:110–9.
- Maulik G, Madhiwala P, Brooks S, et al. Activated c-Met signals through PI3K with dramatic effects on cytoskeletal functions in small cell lung cancer. *J Cell Mol Med* 2002;6:539–53.
- Wang X, DeFrances MC, Dai Y, et al. Mechanism of cell survival: sequestration of Fas by the HGF receptor Met. *Mol Cell* 2002;9:411–21.
- Miyashita K, Itoh H, Sawada N, et al. Adrenomedullin provokes endothelial Akt activation and promotes vascular regeneration both *in vitro* and *in vivo*. *FEBS Lett* 2003;544:86–92.
- Terman BI, Stoletoy KV. VEGF and tumor angiogenesis. *Einstein Quart. J Biol Med* 2001;18:59–66.
- Wang X, Kiyokawa H, Dennewitz MB, Costa RH. The Forkhead Box m1b transcription factor is essential for hepatocyte DNA replication and mitosis during mouse liver regeneration. *Proc Natl Acad Sci U S A* 2002;99:16881–6.
- Raisanen SR, Lehenkari P, Tasanen M, et al. Carbonic anhydrase III protects cells from hydrogen peroxide-induced apoptosis. *FASEB J* 1999;13:513–22.
- Mankoo BS, Collins NS, Ashby P, et al. Mox2 is a component of the genetic hierarchy controlling limb muscle development. *Nature* 1999;400:69–73.

23. A. Artan S, Oner U, et al. The importance of genomic copy number changes in the prognosis of glioblastoma multiforme. *Neurosurg Rev* 2004;27:58–64.
24. Li L, Ren CH, Tahir SA, Ren C, Thompson TC. Caveolin-1 maintains activated Akt in prostate cancer cells through scaffolding domain binding site interactions with and inhibition of serine/threonine protein phosphatases PP1 and PP2A. *Mol Cell Biol* 2003;23:9389–404.
25. Feldkamp MM, Lala P, Lau N, Roncari L, Guha A. Expression of activated epidermal growth factor receptors, Ras-guanosine triphosphate, and mitogen-activated protein kinase in human glioblastoma multiforme specimens. *Neurosurgery* 1999;45:1442–53.
26. Xiao GH, Jeffers M, Bellacosa A, et al. Anti-apoptotic signaling by hepatocyte growth factor/Met via the phosphatidylinositol 3-kinase/Akt and mitogen-activated protein kinase pathways. *Proc Natl Acad Sci U S A* 2001;98:247–52.
27. Zhang Q, Wu Y, Chau CH, et al. Crosstalk of hypoxia-mediated signaling pathways in upregulating plasminogen activator inhibitor-1 expression in keloid fibroblasts. *J Cell Physiol* 2004;199:89–97.
28. Hu CJ, Wang IY, Chodosh LA, Keith B, Simon MC. Differential roles of hypoxia-inducible factor 1 α (HIF-1 α) and HIF-2 α in hypoxic gene regulation. *Mol Cell Biol* 2003;23:9361–74.
29. Sallinen SL, Sallinen PK, Haapasalo HK, et al. Identification of differentially expressed genes in human gliomas by DNA microarray and tissue chip techniques. *Cancer Res* 2000;60:6617–22.
30. Furuta M, Weil RJ, Vortmeyer AO, et al. Protein patterns and proteins that identify subtypes of glioblastoma multiforme. *Oncogene* 2004;23:6806–14.
31. von Deimling A, von Ammon K, Schoenfeld D, et al. Subsets of glioblastoma multiforme defined by molecular genetic analysis. *Brain Pathol* 1993;3:19–26.
32. Watanabe K, Tachibana O, Sata K, et al. Overexpression of the EGF receptor and p53 mutations are mutually exclusive in the evolution of primary and secondary glioblastomas. *Brain Pathol* 1996;6:217–23.
33. Vogelstein B, Lane D, Levine AJ. Surfing the p53 network. *Nature* 2002;408:307–10.
34. Zhao R, Gish K, Murphy M, et al. Analysis of p53-regulated gene expression patterns using oligonucleotide arrays. *Genes Dev* 2003;14:981–93.
35. Zheng L, Lee WH. Retinoblastoma tumor suppressor and genome stability [review]. *Adv Cancer Res* 2002;85:13–50.
36. Henson JW, Schnitker BL, Correa KM, et al. The retinoblastoma gene is involved in malignant progression of astrocytomas. *Ann Neurol* 1994;36:714–21.
37. Chang HY, Sneddon JB, Alizadeh AA, et al. Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. *PLoS Biol* 2004;2:E7.
38. Wang TN, Albo D, Tuszynski GP. Fibroblasts promote breast cancer cell invasion by upregulating tumor matrix metalloproteinase-9 production. *Surgery* 2002;132:220–5.
39. Morikawa S, Takabe W, Mataka C, et al. The effect of statins on mRNA levels of genes related to inflammation, coagulation, and vascular constriction in HUVEC. Human umbilical vein endothelial cells. *J Atheroscler Thromb* 2002;9:178–83.
40. Brat DJ, Van Meir EG. Vaso-occlusive and prothrombotic mechanisms associated with tumor hypoxia, necrosis, and accelerated growth in glioblastoma. *Lab Invest* 2004;84:397–405.
41. Tanwar MK, Gilbert MR, Holland EC. Gene expression microarray analysis reveals YKL-40 to be a potential serum marker for malignant character in human glioma. *Cancer Res* 2002;62:4364–8.
42. Rickman DS, Bobek MP, Misk DE, et al. Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis. *Cancer Res* 2001;61:6885–91.
43. Markert JM, Fuller CM, Gillespie GY, et al. Differential gene expression profiling in human brain tumors. *Physiol Genomics* 2001;5:21–33.
44. Das R, Mahabeshwar GH, Kundu GC. Osteopontin induces AP-1-mediated secretion of urokinase-type plasminogen activator through c-Src-dependent epidermal growth factor receptor transactivation in breast cancer cells. *J Biol Chem* 2004;279:11051–64.
45. Tuck AB, Hota C, Wilson SM, Chambers AF. Osteopontin-induced migration of human mammary epithelial cells involves activation of EGF receptor and multiple signal transduction pathways. *Oncogene* 2003;22:1198–205.
46. Das R, Mahabeshwar GH, Kundu GC. Osteopontin stimulates cell motility and nuclear factor κ B-mediated secretion of urokinase type plasminogen activator through phosphatidylinositol 3-kinase/Akt signaling pathways in breast cancer cells. *J Biol Chem* 2003;278:28593–606.
47. Rich JN, Shi Q, Hjelmeland M, et al. Bone-related genes expressed in advanced malignancies induce invasion and metastasis in a genetically defined human cancer model. *J Biol Chem* 2003;278:15951–7.

APPENDIX C

**Cartilage-selective genes identified in genome-scale
analysis of non-cartilage and cartilage gene expression.**

Cartilage-selective genes identified in genome-scale analysis of non-cartilage and cartilage gene expression

Vincent A Funari¹, Allen Day², Deborah Krakow^{1,2,3}, Zachary A Cohn¹, Zugen Chen², Stanley F Nelson^{2,4} and Daniel H Cohn^{*1,2,4}

Address: ¹Medical Genetics Institute, Cedars-Sinai Medical Center, SSB-3, 8700 Beverly Blvd, Los Angeles, CA 90048, USA, ²Departments of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095, USA, ³Departments of Obstetrics and Gynecology, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095, USA and ⁴Departments of Pediatrics, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095, USA

Email: Vincent A Funari - vincent.funari@cshs.org; Allen Day - allenday@ucla.edu; Deborah Krakow - deborah.krakow@cshs.org; Zachary A Cohn - zuzuelf@yahoo.com; Zugen Chen - zchen@ucla.edu; Stanley F Nelson - snelson@ucla.edu; Daniel H Cohn* - dan.cohn@cshs.org

* Corresponding author

Published: 12 June 2007

BMC Genomics 2007, **8**:165 doi:10.1186/1471-2164-8-165

This article is available from: <http://www.biomedcentral.com/1471-2164/8/165>

© 2007 Funari et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received: 16 February 2007

Accepted: 12 June 2007

Abstract

Background: Cartilage plays a fundamental role in the development of the human skeleton. Early in embryogenesis, mesenchymal cells condense and differentiate into chondrocytes to shape the early skeleton. Subsequently, the cartilage anlagen differentiate to form the growth plates, which are responsible for linear bone growth, and the articular chondrocytes, which facilitate joint function. However, despite the multiplicity of roles of cartilage during human fetal life, surprisingly little is known about its transcriptome. To address this, a whole genome microarray expression profile was generated using RNA isolated from 18–22 week human distal femur fetal cartilage and compared with a database of control normal human tissues aggregated at UCLA, termed Celsius.

Results: 161 cartilage-selective genes were identified, defined as genes significantly expressed in cartilage with low expression and little variation across a panel of 34 non-cartilage tissues. Among these 161 genes were cartilage-specific genes such as cartilage collagen genes and 25 genes which have been associated with skeletal phenotypes in humans and/or mice. Many of the other cartilage-selective genes do not have established roles in cartilage or are novel, unannotated genes. Quantitative RT-PCR confirmed the unique pattern of gene expression observed by microarray analysis.

Conclusion: Defining the gene expression pattern for cartilage has identified new genes that may contribute to human skeletogenesis as well as provided further candidate genes for skeletal dysplasias. The data suggest that fetal cartilage is a complex and transcriptionally active tissue and demonstrate that the set of genes selectively expressed in the tissue has been greatly underestimated.

Background

Skeletogenesis begins with condensation of mesenchymal chondroprogenitor cells to form the cartilage anlagen that pattern the early skeleton. Subsequently, for bones that grow by endochondral ossification, the chondrocytes differentiate further to establish the growth plates. At the joint surfaces, development of articular cartilage facilitates and maintains joint movement during fetal life. These multi-step processes require the coordinated expression of many genes, including genes encoding extracellular matrix proteins and morphogens, as well as proliferative, angiogenic, and apoptotic signals [1]. Most of our knowledge of the function of the genes involved has been derived from developmental studies in model systems and cell lines [2] as well as from the identification of disease genes in skeletal disorders.

Whole genome analysis of chondrocyte gene expression has the potential to reveal novel genes and gene expression programs which define the tissue. Although the complete set of genes expressed in human cartilage has not yet been described, analysis of human cartilage cDNA libraries has provided an initial *in vivo* picture of the cartilage transcriptome [3-6]. These investigations have also identified expression of both known and novel genes. Comparative microarray studies in rat cartilage [7] and several chondrocyte cell lines [8,9] have provided a larger set of genes of potential importance in chondrocytes, including genes specific to the stages of chondrocyte differentiation. Wang et al. (2004) identified 92 genes with two-fold variation in expression between hypertrophic and proliferative growth plate chondrocytes. In this *in vivo* study, significant gene expression changes were principally associated with cell cycle, transcription, extracellular matrix structure, receptor and transporter functions. In microarray studies of mouse micromass cultures [8], 212 genes exhibited at least a ten-fold difference in gene expression as the cultures differentiated. Thus global characterization of gene expression is beginning to describe the identities of key regulatory molecules and their targets in chondrocytes.

Disrupting genes involved in the organization and maturation of the growth plate and/or the stability of articular cartilage results in inherited skeletal disorders that range from perinatal lethal phenotypes to mild disorders with early-onset osteoarthropathy as their major feature [10,11]. Of the approximately 370 clinically distinguishable skeletal dysplasias [12], mutations in 115 genes have been associated with about 150 disorders. Many of these disease genes are expressed in a cartilage-selective pattern, and therefore identifying additional genes uniquely expressed in cartilage should yield new skeletal dysplasia candidate genes.

To identify a larger set of genes uniquely expressed in chondrocytes, this study describes a genome-scale gene expression profile for 18–22 week human fetal cartilage. There were 161 genes which appeared to be selectively expressed in fetal cartilage, comprising a variety of novel genes that may contribute to skeletal development. The data suggest a complex pattern of cartilage gene expression and indicate that the number of genes selectively expressed in cartilage has been greatly underestimated.

Results

Identification of cartilage-selective genes

To define a set of genes preferentially or uniquely expressed in normal human fetal cartilage, cartilage probeset intensities were compared with probeset intensities across a variety of normal tissues. A two-step process was employed for gene identification, consisting of a training step and a validation step (see the additional data file 1, for a flow chart of an overview of the analysis). The tissue-selectivity of a representative sampling of the identified genes was confirmed by quantitative RT-PCR.

The training dataset consisted of five independent cartilage samples and 41 non-cartilage samples, all analyzed using Affymetrix U133 2.0 Plus arrays. The average correlation coefficient among the cartilage samples (R^2) was 0.96. To identify unbiased relationships within the data, and to test the robustness of the normalization and tissue-specificity, an unsupervised approach [13], in which the genes and tissues were grouped based only on expression patterns, was employed. Probesets with the greatest variation across all tissues and whose expression in any two arrays differed by at least two standard deviations from their mean expression across the entire set of samples were selected. This selection yielded 9483 probesets.

Two-way hierarchical clustering based on similarity of expression of these 9483 probesets within the samples was performed (Figure 2). Samples from the same tissues clustered together, indicating that the normalization was sufficiently robust to allow tissue-selective expression patterns to be identified. Even with these relatively non-stringent selection criteria, the results showed a surprisingly large number of genes with a fetal cartilage-selective expression pattern. At least 89 probesets representing 64 genes with coordinately higher expression in cartilage relative to non-cartilage tissues appeared to drive the clustering of the two groups (Figure 2B). These probesets formed a gene expression node in the dendrogram which shared an overall expression correlation of 0.99. The genes represented by these probesets included some well established cartilage-selective genes, including aggrecan (AGC1), type \times collagen (COL10A1), and matrilin 3 (MATN3), among others. Thus, a comparative approach with microarrays can identify genes whose expression is cartilage-selective.

To define a ranked list of genes significantly expressed in cartilage, a supervised analysis [13], comparing cartilage versus non-cartilage gene expression, was employed. This consisted of a two-class analysis with a modified t-test (SAM) (See additional data file 2, for the complete results of this analysis). There were 2634 probesets representing 1720 genes with at least three-fold differential expression when comparing cartilage and non-cartilage tissues, with a false discovery rate of zero. Of these, 2446 of the 2634 probesets demonstrated higher expression in cartilage with respect to non-cartilage tissues, while the remaining 188 probesets were expressed at significantly higher levels in the other tissues. As observed for the hierarchical clustering, probesets representing well-known cartilage markers, including *COL2A1*, *AGC1*, *COMP*, *COL9A3*, and *MMP3* were among the top genes listed. In addition, lubricin (*PRG4*), an articular cartilage-specific marker, was also identified, confirming the ability to identify genes specific to fetal articular cartilage. Indeed, among the top 35 probesets more highly expressed in cartilage, only four probesets, representing unannotated genes, were derived from genes not previously known to be expressed in cartilage.

In silico validation

Three array platforms were used to validate the 2446 probesets identified in the supervised analysis and generate a robust list of cartilage-selective genes (Table 1). A majority of these probesets (2245 probesets (> 92%)) were identified in 124 U133A and 74 U133B arrays using the Celis database (see Materials and Methods), and represented expression from 34 normal tissues. A small proportion of the probesets (201/2446) are not found on the Affymetrix™ Human Genome U133A/B Arrays, so these probesets were identified in the analysis of 26 U133 Plus 2.0 arrays, representing eight non-cartilage tissues. A summary of the validation and the tissue distribution are available as additional files.

Of the three platforms, the U133A dataset was the most robust with regard to the number of arrays, biological replicates, diversity of tissues, probes identified, and gene annotation. From this platform, 1363 of the 2446 probesets identified in the SAM analysis as expressed at a higher level in cartilage were obtained. Two hundred seventy-four of the 1363 probesets (274/1363), representing 237 genes, exhibited at least five-fold higher expression when compared to non-cartilage tissues and were ranked by cartilage-specificity using an analog of coefficient of variation (CV) (see Methods). Of these, 56 probesets, representing 49 genes, were identified with a CV < 50% in non-cartilage samples, constituting the cartilage-selective gene set from this platform (Table 1, left). Twenty of these genes have mutations that have been associated with skel-

etal phenotypes in humans and/or mice, representing 44% of the probes selected from this platform.

Eight hundred eighty-two of the 2446 probesets were identified from the U133B validation set. Two hundred fourteen of these probesets, representing 158 genes, were well measured in cartilage with at least five-fold higher expression in cartilage relative to non-cartilage tissues. Of these, 77 probesets had a CV less than 50% in non-cartilage samples, representing 71 cartilage-selective genes (Table 1, center), including 3 genes also identified using the U133A platform (*COL11A1*, *EDIL3*, and *PDPN*).

A subset of the cartilage-selective genes was represented only on the Human Genome U133 Plus 2.0 arrays and were selected from the analysis of 28 non-cartilage samples. In total 201/2446 probesets were not represented in the U133A/B array set. Of these 201 probesets, 96 probesets, representing 85 genes, had a five-fold higher expression in cartilage than non-cartilage samples. By including the CV selection criterion, 52 probesets, representing 50 cartilage-selective genes were identified and added to the complete tally (Table 1, right), including 6 genes also identified using the U133A and U133B arrays (*IRAK2*, *NRP2*, *WTAP*, *PITPNC1*, *AKR1C2*, and *PTK2*).

In summary, 480 genes demonstrating enriched or specific expression in cartilage were selected from the com-

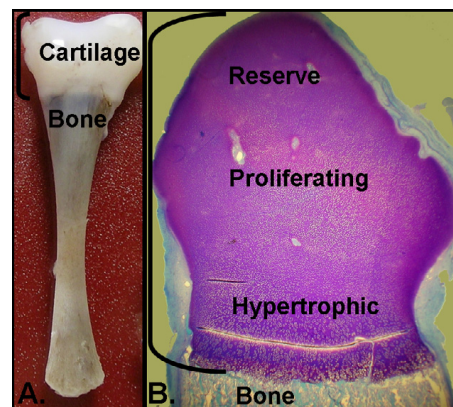


Figure 1
Human fetal cartilage section dissected for RNA expression profiling. (A) Normal distal femur cartilage from an 18–22 week fetus. Brackets define the cartilage that was dissected and used for RNA profiling. (B) Toluidine blue stained longitudinal section of the distal head of the femur magnified 8.5 fold. The dissected portion included chondrocytes from the articular, reserve, proliferating, and hypertrophic zones.

Table 1: Cartilage-selective genes validated *in silico* on U133A (left), U133B (center), and U133 Plus 2.0 (right) platforms.

U133A SYMBOL	Probe	CV	Class	Mouse	Human	SAM	Fold	U133B SYMBOL	Probe	CV	Class	Mouse	Human	SAM	Fold	U133 2.0 SYMBOL	Probe	CV	Class	SAM	Fold
AGC1	207692_s_at	9.0	ST		x	14	65.6	COL11A1	229271_x_at	11.0	ST	x	x	3	54.2	FLJ16008	156868_at	6.4		44	8.2
AGC1	217161_x_at	9.2	ST		x	31	15.5	230895_at	230895_at	11.3				4	67.0	UBE3C	1560739_a_at	8.5	EZ	305	5.1
MATN1	206905_s_at	9.6	ST			22	75.3	EDIL3	233668_at	11.8	SI			221	6.0	1563414_at	1563414_at	8.8		77	5.1
COL10A1	217428_s_at	9.9	ST	x	x	284	5.9	C10orf49	236800_at	15.5				13	76.0	IRAK2	1553740_a_at	11.0	SI	107	6.3
MATN3	206091_at	10.1	ST		x	32	6.7	IRAK2	231779_at	15.6	SI			158	25.1	COL25A1	1555253_at	11.4	ST	516	5.1
MMP13	205959_at	10.2	ME	x	x	6	28.9	SLC4A5	234976_x_at	16.3	EZ			1036	6.4	KIAA0701	1554292_a_at	11.6		57	5.9
MATN4	207123_s_at	10.4	ST			59	7.8	LOC399959	239672_at	16.4				633	10.2	SULT1C2	1553321_a_at	11.7	ME	105	7.5
AGC1	205679_x_at	10.6	ST		x	18	28.6	IBSP	236028_at	16.4	SI			628	19.0	ITGB1	1561042_at	11.7	SI	542	7.0
COL11A2	216993_s_at	11.5	ST	x	x	116	7.3	EDNRA	243555_at	22.0	SI	x		213	5.6	RP4-756G23.1	1557123_a_at	13.1		214	6.2
NOS2A	210037_s_at	11.6	SI			20	18.7	HSUP1	229899_s_at	22.2				1994	5.1	NRP2	1555468_at	13.7	SI	45	10.2
LECT1	206309_at	13.3	SI	x		34	52.3	COL9A2	232542_at	22.5	ST		x	1929	5.0	MSI2	1552364_s_at	14.2	SI	178	5.2
HAS2	206432_at	14.2	ME			91	5.4	CMAH	229604_at	23.0	EZ			272	8.0	PTGFR	1555097_a_at	17.0	SI	470	8.0
WISP1	206796_at	15.6	SI			15	7.1	SNX5	232666_at	23.1	SI			108	5.3	BCL10	1557257_at	17.5	EZ	1014	6.1
HAPLN1	205523_at	18.5	ST	x		8	18.8	EDG2	232716_at	23.2	SI	x		712	5.4	WTAP	1560274_at	17.5	SI	228	6.3
COL9A1	222008_at	19.4	ST	x	x	19	185.1	TNFRSF18	224553_s_at	23.9	SI			124	9.4	RUNX1	1557527_at	17.8	SI	297	13.0
CIQTNF3	220988_s_at	20.3				291	5.4	USP12	236975_at	23.9	EZ			1776	6.3	LACTB	1552485_at	18.6	EZ	659	6.5
COL11A2	213870_at	20.5	ST	x	x	11	136.9	RP6-213H19.1	224407_s_at	25.3	SI			313	5.8	WVVP2	1552737_s_at	21.1	EZ	144	10.9
COL10A1	205941_s_at	20.5	ST	x	x	684	38.5	BIC	229437_at	26.9	SI			571	5.5	PITPNC1	1568949_at	21.5	SI	258	5.5
NGFB	206814_at	21.0	SI			78	7.1	KIAA1718	225142_at	27.8				651	11.5	ARIH1	1558710_at	22.2	EZ	1264	5.2
PDPN	204879_at	21.5	ST	x		103	8.8	RBI1CC1	237626_at	27.9	SI			1489	5.5	ADAMTS9	1554697_at	22.7	ME	769	7.4
222348_at	222348_at	21.8				227	8.4	VTAP	244219_at	28.1	SI			371	6.1	SYNJ2	1555009_a_at	23.7	SI	277	5.9
SLC28A3	220475_at	21.9	SI			296	5.1	RHOQ	239258_at	28.2	SI			784	5.4	SRGAP1	1554473_at	23.8	SI	809	5.1
EIF2C2	213310_at	22.4	EZ			1071	9.1	RPS6KA3	241460_at	28.3	SI	x	x	529	9.1	MGC17337	1552277_a_at	24.2		690	6.0
SOX5	207336_at	23.5	SI	x		1250	8.2	SEMA6D	233801_s_at	29.1	SI			245	5.3	ISS5841_at	1555841_at	24.5		1606	6.8
DSPG3	206439_at	24.0	ME			5	15.0	FN1	235629_at	29.6	ST			622	10.5	CD44	1565868_at	24.8	SI	1025	6.2
COL11A1	37892_at	25.7	ST	x	x	12	82.4	ULBP2	238542_at	29.7	SI			101	6.1	SLC41A2	1562208_a_at	24.9	ST	407	8.7
COL11A1	204320_at	26.0	ST	x	x	73	63.6	229221_at	229221_at	30.1				176	7.6	RHOF	1554539_a_at	26.7	SI	704	6.4
SOD2	215078_at	26.2	EZ			76	10.7	PTK2	241453_at	30.2	SI			355	7.7	ARF1	1565651_at	27.7	SI	768	5.6
HSPC159	219998_at	27.6	EZ			126	8.1	CHST11	226368_at	30.6	ME			310	5.3	ISS2288_at	1552288_at	28.0		127	17.1
BMP2	205289_at	29.3	SI			591	17.4	VASN	225867_at	31.9	SI			593	5.9	B3GNT5	1554835_a_at	29.5	ME	331	5.5
CSPG4	214297_at	29.5	ST			17	39.5	SCUBE3	230290_at	33.3	SI			123	6.4	B3GNT7	1555963_x_at	30.4	ME	339	18.3
CHST3	32094_at	30.0	EZ		x	1560	6.3	LOC338758	238893_at	33.4				998	5.2	SETD5	1569106_s_at	31.2		839	6.0
RELB	205205_at	30.1	SI			390	7.7	LOXL4	227145_at	33.8	EZ			260	7.5	OSMR	1554008_at	31.9	SI	85	50.3
CYTL1	219837_s_at	30.7	SI			151	17.0	SLC25A37	242335_at	33.9	SI			765	12.3	ZFYVE16	1554638_at	32.2		722	5.2
FZD10	219764_at	30.8	SI			51	7.3	228910_at	228910_at	34.0				630	5.4	AKR1C2	1562102_at	32.8	EZ	559	5.7
BDKRB1	207510_at	31.6	SI			239	5.1	PITPNC1	239808_at	34.3	SI			1129	11.8	WTAP	1558783_at	33.2	SI	946	8.0
ITGA10	206766_at	32.6	SI	x		41	29.4	FNDC3B	222693_at	34.3	SI			1394	10.2	LRRC8C	1558517_s_at	33.4		731	5.3
RNF24	210706_s_at	32.6				774	5.1	PDPN	226658_at	35.2	ST			125	15.7	ISS2289_a_at	1552289_a_at	34.4		164	28.5
NUPL1	204435_at	33.0	SI			1680	5.2	LOC201181	241383_at	35.2				173	14.5	SFXN3	1559993_at	35.7	SI	160	7.8
RNF24	204669_s_at	33.0				1349	6.0	FNDC3B	244022_at	37.0	SI			456	7.3	LRP11	1561180_at	37.8	SI	393	10.3

Table 1: Cartilage-selective genes validated *in silico* on U133A (left), U133B (center), and U133 Plus 2.0 (right) platforms. (Continued)

AKRIC2	217626_at	33.6	EZ			595	12.3	CHST11	226372_at	37.9	ME		554	5.9	B3GNT7	1555962_at	38.2	ME	396	12.5
LOC283824	213725_x_at	35.5				1433	6.5	225611_at	225611_at	38.2			1117	7.4	ZNF146	1569312_at	38.3		374	10.5
PDLIM4	214175_x_at	35.6	SI			212	5.6	NRP2	232701_at	38.6	SI		58	6.5	ZCCHC7	1556543_at	41.5		1906	6.2
FOSL1	204420_at	37.8	SI		x	458	15.2	TNFRSF10D	227345_at	38.7	SI		363	6.5	ATF1	1565269_s_at	42.0	SI	533	6.7
HAPLN1	205524_s_at	38.4	ST	x		16	137.6	FNDCC3B	232472_at	39.4	SI		327	5.9	HIG2	1554452_a_at	43.0	EZ	119	44.6
MIA	206560_s_at	41.0	ST	x		36	8.1	229242_at	229242_at	39.7			330	6.3	KLF7	1555420_a_at	43.3	SI	1056	5.7
MMP12	204580_at	41.3	ME			29	8.2	ANKRD28	229307_at	39.7	SI		820	7.6	BCL2L1	1558143_a_at	43.8	EZ	683	6.2
TNMD	220065_at	41.9	ST	x		113	5.0	TRPS1	234351_x_at	39.7	SI	x	841	5.6	TGIF	1566901_at	43.8	SI	380	6.1
RLF	204243_at	43.0	SI			478	7.8	GLIS3	230258_at	40.2	SI		872	5.4	MCOLN2	1555465_at	43.9		520	8.6
EDIL3	207379_at	44.2	SI			84	5.4	SCUBE3	228407_at	40.2	SI		515	5.9	ChGn	1569387_at	44.7	ME	843	5.4
THBS3	209561_at	45.0	SI	x		716	7.2	SCYL1BP1	226337_at	40.5			1714	6.9	PTK2	1559529_at	45.1	SI	562	6.2
MMP3	205828_at	46.6	EZ	x		1	73.5	YME1L1	232216_at	40.9	ME		2251	5.5	FAM62B	1555830_s_at	49.5		837	6.6
RELA	209878_s_at	47.1	SI			971	7.3	KIAA0999	242920_at	41.0			1731	8.4						
LIF	205266_at	47.9	SI			120	15.7	236289_at	236289_at	41.2			445	5.4						
BMP6	206176_at	48.3	SI	x		60	38.9	GALNTL2	236361_at	42.1	ME		290	15.7						
ETNK1	219017_at	49.0	EZ			1400	5.1	PET112L	228441_s_at	42.2	EZ		333	9.8						
								ZNF697	227080_at	42.2			1119	5.5						
								FNDCC3B	222692_s_at	42.3	SI		747	9.4						
								GPC6	223730_at	42.4	SI		511	9.5						
								COL27A1	225292_at	42.7	ST		135	5.4						
								NRP2	229225_at	43.1	SI		96	9.8						
								UFM1	242669_at	44.1	EZ		624	6.6						
								ASAM	228082_at	44.3	SI		68	11.7						
								KCNT2	234103_at	44.4	EZ		299	5.3						
								ERO1L	222646_s_at	44.6	EZ		1358	6.8						
								MAST4	225613_at	44.6	SI		642	7.9						
								228314_at	228314_at	44.7			1501	6.5						
								C8orf72	232668_at	45.0			400	15.6						
								RPS6	238156_at	45.7	EZ		1922	5.1						
								SQSTM1	244804_at	46.6	EZ	x	1440	5.6						
								230204_at	230204_at	46.8			26	78.1						
								TBX15	230438_at	46.9	SI		932	5.9						
								244533_at	244533_at	47.0			752	5.5						
								235821_at	235821_at	47.2			35	9.9						
								ZNF160	239954_at	47.5			1897	7.1						
								SERPINE2	236599_at	47.9	EZ		80	12.1						
								236685_at	236685_at	49.5			1555	5.5						

Genes were ranked by selectivity (CV) and classified into four functional classes Structural Protein (ST), Enzyme (EZ), Signaling (SI), and Extracellular Matrix Enzyme (ME). Genes associated with skeletal phenotypes in mice and/or human or genes are denoted in Mouse and Human columns, respectively. SAM defines the rank order of each gene identified by two-class SAM analysis of cartilage versus non-cartilage tissues. The average cartilage expression divided by the median of non-cartilage expression (Fold) is listed for each gene.

Table 2: Non-redundant set of 161 cartilage-selective genes organized by chromosomal location.

Chromosomal Location	Symbol
chr1p11.1	TBX15
chr1p12	ZNF697
chr1p13.1	NGFB
chr1p21	COL11A1
chr1p22	BCL10
chr1p22	MCOLN2
chr1p22.2	228314_at
chr1p22.2	LRRC8C
chr1p31.1	PTGFR
chr1p32	RLF
chr1p33-p32	COL9A2
chr1p35	MATN1
chr1p36.21	PDPN
chr1p36.3	TNFRSF18
chr1q21	ITGA10
chr1q21	THBS3
chr1q24.2	SCYL1BP1
chr1q31.3	KCNT2
chr1q41	244533_at
chr1q42	ARF1
chr1q42.13	222348_at
chr2p13	SLC4A5
chr2p14	HSPC159
chr2p21	RHOQ
chr2p24-p23	MATN3
chr2q11.1-q11.2	SULT1C2
chr2q13	236289_at
chr2q13	BCL2L11
chr2q14.3	FLJ16008
chr2q32	KLF7
chr2q33.3	NRP2
chr2q33-q35	SERPINE2
chr2q34	FN1
chr2q37.1	B3GNT7
chr3p14.3-p14.2	ADAMTS9
chr3p24.3	ANKRD28
chr3p24.3	GALNTL2
chr3p25.3	IRAK2
chr3p25.3	SETD5
chr3q26.31	FNDC3B
chr3q28	B3GNT5
chr4p16-p15	CYTL1
chr4q21-q25	IBSP
chr4q22.1-q23	229221_at
chr4q25	COL25A1
chr4q27-q28	PET112L
chr4q31.23	EDNRA
chr4q35.1	1563414_at
chr5p13.1	OSMR
chr5p13.3	C1QTNF3
chr5p15.2-q14.3	ZFYVE16
chr5q12.3	225611_at
chr5q12.3	MAST4
chr5q14	EDIL3
chr5q14.3	230204_at
chr5q14.3	230895_at
chr5q14.3	HAPLN1
chr5q31.1	PDLM4
chr5q35	SQSTM1
chr6p21.3	COL11A2
chr6p21.3	SCUBE3
chr6p21.32	CMAH
chr6q24.2	236685_at

parison of cartilage and non-cartilage tissues with data derived from the U133A (n = 237), U133B (n = 158) and U133 Plus 2.0 (n = 85) platforms. Of these, a non-redundant set of 161 genes (Table 2), including 11 uncharacterized genes and 16 genes represented by unannotated probesets, were classified as cartilage-selective. These data greatly expand the number of genes known to be selectively expressed in cartilage and emphasize the unique pattern of gene expression that determines its properties.

qRT-PCR validation

Quantitative RT-PCR was used to independently assess the tissue-selectivity of the genes identified in the microarray analysis. For each of the three microarray platforms, the probesets with a CV less than 50% were divided into 10% intervals (0–10% CV, 10–20% etc.) (Table 1), and one gene from the middle of each interval was selected for analysis by qRT-PCR.

All of the thirteen of genes analyzed demonstrated higher expression in cartilage than in the seven non-cartilage tissues studied (Table 3). Also, with one exception (*OSMR*), the selection threshold of at least five-fold higher expression in cartilage tissues as compared with the average expression among all non-cartilage tissues imposed for the microarray analysis, was observed. For most of the genes studied by qRT-PCR, there was little expression in the seven non-cartilage tissues (median Ct = 33.2), indicating that including the coefficient of variation in the ranking algorithm preferentially identifies genes selectively expressed in cartilage. Also, there was an inverse correlation between the gene rank and the standard deviation in expression level among non-cartilage tissues, indicating that genes with a higher rank were more selectively expressed in cartilage. Finally, there was a trend of decreasing cartilage selectivity moving from the U133A to U133B to U133 2.0 qRT-PCR validations, likely reflecting the decreasing robustness of comparison datasets in the respective platforms. Overall the qRT-PCR experiments replicated and validated findings derived from the comparative microarray data.

Discussion

Using genome-scale microarrays, gene expression in human fetal cartilage was compared with a robust set of other normal tissues. Hierarchical clustering showed remarkable similarity among the 18–22 week fetal cartilage expression profiles and demonstrated that a subset of the cartilage transcriptome is composed of a unique gene set not generally expressed in the other tissues studied. Using SAM, 2446 probesets measured preferential expression of 1712 genes with at least three-fold higher expression in cartilage as compared with other tissues. 1028 (42%) of these probesets matched genes identified in a cartilage growth plate cDNA library [4] validating their

Table 2: Non-redundant set of 161 cartilage-selective genes organized by chromosomal location. (Continued)

chr6p24-p23	BMP6
chr6q12-q14	COL9A1
chr6q21-q22	COL10A1
chr6q25	ULBP2
chr6q25.1	LRP11
chr6q25.3	SOD2
chr6q25.3	SYNJ2
chr6q25-q27	WTAP
chr7q32.1	HIG2
chr7q34	KIAA1718
chr7q36.3	FAM62B
chr7q36.3	UBE3C
chr8p21	TNFRSF10D
chr8p21.2	SLC25A37
chr8p21.3	ChGn
chr8p22-q21.13	RB1CC1
chr8q12.1	C8orf72
chr8q24	EIF2C2
chr8q24.12	HAS2
chr8q24.12	TRPS1
chr8q24.1-q24.3	VWIP1
chr8q24.22	235821_at
chr8q24-qter	PTK2
chr9p13.2	ZCCHC7
chr9p21	RPS6
chr9p24.2	GLIS3
chr9q22.2	SLC28A3
chr9q31.1	1555841_at
chr9q31.1	MGC17337
chr9q31.3	EDG2
chr9q32	229242_at
chr9q32	COL27A1
chr10p11.2	ITGB1
chr10p13	C10orf49
chr10p14	YME1L1
chr10p15-p14	AKR1C2
chr10q22.1	CHST3
chr10q24	LOXL4
chr10q24.31	SFXN3
chr11p11.2	228910_at
chr11p13	CD44
chr11q13	FOSL1
chr11q13	RELA
chr11q22.3	MMP12
chr11q22.3	MMP13
chr11q22.3	MMP3
chr11q23.3	KIAA0999
chr11q24.1	ASAM
chr11q24.1	LOC399959
chr12p12.1	ETNK1
chr12p12.1	SOX5
chr12q	CHST11
chr12q13	ATF1
chr12q14.2	SRGAP1
chr12q21	DSPG3
chr12q21.33	LOC338758
chr12q23.1	KIAA0701
chr12q23.3	SLC41A2
chr12q24.31	RHOF
chr12q24.33	FZD10
chr13q12.13	NUPL1
chr13q12.13	USP12
chr13q13.3	UFMI
chr13q14-q21	LECT1
chr13q32	GPC6
chr14q22.1	ERO1L
chr14q32.1-q32.2	BDKRB1

expression in cartilage via an independent dataset. The identification of genes known to have restricted patterns of expression in cartilage confirmed the presence of RNA derived from the reserve (*GREM1*), hypertrophic (*BMP6*, *COL10A1*), and terminally differentiated (*MMP13*) chondrocytes, in addition to genes expressed throughout all zones of the growth plate. This analysis suggested that there is differential transcriptional regulation of many genes in fetal cartilage and that the data could be used to identify genes selectively expressed in the tissue.

Tissue-selective genes have been previously defined as genes with enriched expression in a particular tissue [14] and characterized with algorithms dependent on the degree of differential expression relative to other tissues, including *t*-test [15], SAM [16], fold change [14,17,18], and enrichment scores [14]. While these approaches successfully identify tissue-selective genes, the reliance on fold change reduces the significance of many selectively expressed genes with low fold change. To compensate for this and identify cartilage-selective genes expressed at lower levels, the approach presented here placed increased significance on the preferentially expressed genes that showed the least variation of expression in non-cartilage tissues. This was made possible by the use of publicly-released reference gene expression data performed on the same platform and led to the reliable identification of genes with lower fold changes, but high cartilage selectivity. The impact of the use of coefficient of variation on the ranked gene list is apparent in Tables 1 and 2. In the U133A dataset, nine of the top 25 genes were ranked higher than 100 in significance in the SAM ranking. The average fold change of the probes for these nine genes was 10.7, while the average fold change of the probes for the other 14 of the top 25 genes was 42.2. One of these probes, *COL10A1* was among the top four cartilage-selective genes using the CV algorithm but ranked at 284 by SAM (Table 1). In the U133B dataset, which contains a higher percentage of unannotated genes, 4 of the top 50 probes had a SAM ranking below 100, and the average SAM ranking was 576. Overall, to identify only the most cartilage-selective genes, a threshold of 50% coefficient of variation was used across all three platforms, yielding 161 cartilage-selective genes. A subset of 13 of the 161 cartilage-selective genes was studied by quantitative RT-PCR in cartilage and eight non-cartilage tissues to independently assess tissue selectivity. The data confirmed the cartilage-selectivity of genes with less than 50% CV, validating the selection procedure and suggesting that the gene expression patterns determined by microarray analysis are representative.

The coefficient of variation selection approach could, in theory, equally select for three different patterns of expression: cartilage-specific genes; genes with a consistent level

Table 2: Non-redundant set of 161 cartilage-selective genes organized by chromosomal location. (Continued)

chr15q21.1	SEMA6D
chr15q22.1	LACTB
chr15q24	ARIHI
chr15q24.2	CSPG4
chr15q26.1	AGC1
chr16p13.12	LOC283824
chr16p13.3	VASN
chr16q22.1	WWP2
chr17q11.2-q12	NOS2A
chr17q21.2	LOC201181
chr17q22	MSI2
chr17q24.2	PITPNC1
chr18p11.3	TGIF
chr19p13.11	I552288_at
chr19p13.11	I552289_a_at
chr19q13.1	ZNF146
chr19q13.32	RELB
chr19q13.32-q13.33	MIA
chr19q13.41	ZNF160
chr20p11	SNX5
chr20p12	BMP2
chr20p13-p12.1	RNF24
chr20q13.13	HSUP1
chr20q13.1-q13.2	MATN4
chr21q21.3	BIC
chr21q22.3	RUNX1
chr22q12.2	LIF
chr22q13.2	RP4-756G23.1
chrXp22.2-p22.1	RPS6KA3
chrXq21.33-q23	TNMD
chrXq26.2	RP6-213H19.1

of baseline expression in non-cartilage tissues; and genes with significant but equal expression in all tissues (e.g. housekeeping genes). In this data analysis, however, the most highly ranked genes consistently demonstrated little or no expression in non-cartilage tissues. The data thus demonstrate that incorporating coefficient of variation preferentially selected for genes not significantly expressed in non-cartilage tissues, yielding genes likely to have important and perhaps unique roles in cartilage.

Regardless of expression level, a cartilage-selective expression pattern suggests that the product of each identified gene may have a functional role in the development of the skeleton. Concordant with this hypothesis, mutations in 25 of the 161 selected genes have been associated with skeletal phenotypes in humans and/or mice. Included among them were the products of the well characterized genes encoding aggrecan and the cartilage-specific collagens, gene products known to have a prominent role in skeletal development and endochondral ossification. By this measure, the remaining genes may be candidate genes for skeletal dysplasias in which the disease gene has yet to be identified. As new skeletal dysplasia loci are defined, coincidence between a locus and a cartilage-selective gene may promote rapid identification of the disease gene. Knockout of the orthologous genes in mice would also

facilitate exploring the role of each gene in skeletal development.

Classification of the biological roles of the products of the cartilage-selective gene set reveals genes with diverse functions including structural proteins of the cartilage extracellular matrix, enzymes that modify them, and 41 gene products with unannotated function. There were 65 genes that are components of signaling pathways, and only 43% of these were identified by sequence analysis of a comparable fetal cartilage cDNA library [4]. Among the genes were elements of the nitric oxide, VEGF, TNF/RANK, and gp130 pathways, all of which have known roles in the growth plate [19-23]. Mutations in the genes encoding some of the molecules in these pathways, including *RPS6SKA3*, *LIFR*, *TNFRSF11A* and *IKBKG*, have been associated with human skeletal dysplasias [12], again suggesting that the remaining genes may also serve critical roles in endochondral ossification.

Multiple genes encoding members of the LIF/gp130 signaling pathway met the definition of cartilage-selective genes. LIF is a cytokine that is expressed in terminally-differentiated growth plate chondrocytes [24] and signals through the gp130/LIFR complex. Homozygosity for loss of function mutations in the LIF receptor produces the recessively inherited skeletal dysplasia, Stuve-Wiedemann syndrome [25]. In addition to their skeletal features, these patients have cardiovascular, pulmonary, gastrointestinal, neurologic and metabolic abnormalities, likely attributable to the role that LIFR plays in embryonic or fetal development. Genes on the cartilage-selective gene list upstream of the receptor include *RELA* and *RELB*, NF-KB survival transcription factors that increase transcription of LIF [26], as well as the *LIF* gene itself. Through the LIFR/gp130 complex, LIF can regulate both the JAK/STAT and ERK MAP kinase pathways. Pathway components downstream of the receptor include ATF1, part of the ATF1/CREB transcription factor complex that participates in ERK MAP kinase signaling [27,28]. The ATF1/CREB complex is also regulated by phosphorylation by the product of the *RPS6SKA3* gene [29,30], another gene in the MAPK/ERK pathway that is associated with a skeletal phenotype. The gene encoding RPS, a phosphorylation target of *RPS6SKA3* [30], was also cartilage-selective, but the role of this protein in growth plate differentiation has yet to be determined. Finally, the gene encoding FOSL1, a FOS-like transcription factor activated by the ERK/MAPK pathway which binds cJUN to form a transcription complex [31,32], was among the cartilage-selective genes identified. Thus comparative microarray analysis has identified multiple components of a regulatory pathway that can be explored to further evaluate their importance in growth plate differentiation and endochondral ossification.

Table 3: Summary of qRT-PCR amplification of cartilage-selective genes in fetal cartilage and seven non-cartilage tissues.

CV Interval	Symbol	Neg	Ct		Fold
			NC	C	
U133A					
0–10%	AGC1	2	33	23	371
10–20%	NOS2A1	3	34	23	1158
20–30%	DSPG3	6	35	25	661
30–40%	BDKRB1	2	34	28	49
40–50%	MMP3	3	33	20	4988
U133B					
10–20%	C10orf49	7	35	23	Unique
20–30%	KIAA1718		30	27	8
30–40%	LOC20118	4	34	29	20
40–50%	ASAM		31	26	29
U133 2.0					
10–20%	KIAA0701		30	27	5
20–30%	MGC17337		30	27	5
30–40%	OSMR		28	27	3
40–50%	HIG2		30	25	21

Representative genes are listed from each validation platform in order of 10% CV interval. (Neg) Number of non-cartilage tissues in which amplification was not detected. Average Ct values for each gene were calculated for both cartilage (C) and non-cartilage (NC) tissues. Where no amplification was observed the maximum Ct value (i.e. 35) was used for calculations. Fold difference (Fold) is calculated from the difference in cartilage and non-cartilage Ct values.

While this study has provided a deep set of genes that exhibit a cartilage-selective expression pattern, there are some limitations to the analysis. First, the study focused on total cartilage RNA, including all types of growth plate chondrocytes, as well as articular cartilage. As a result, it cannot be determined if the selected genes are expressed in all types of chondrocytes or only a subset of cells. In this context, nine of the cartilage selective genes have been shown to be more highly expressed in hypertrophic cells relative to proliferating chondrocytes in the rat and/or mouse [7,8]. Second, the cartilage samples were derived from a single anatomic site and a narrow window of fetal development, so it is unclear to what extent the observed gene expression pattern can be generalized. Third, neither all possible non-cartilage tissues nor each type of cell within each tissue were studied, so cartilage-selectivity could be affected if additional fetal and/or adult tissues that express the identified genes were found. This may be particularly important for other connective tissues such as bone, tendon and ligament which contain cells known to express some of the cartilage-selective genes identified here.

Not all genes selectively expressed in developing cartilage will necessarily be identified using this approach. For

instance, the *COL2A1* gene fell just below the rigorous 50% CV standard set to define cartilage-selectivity. The underlying reasons for this are complex. Probe performance as well as the known expression of *COL2A1* in fetal liver and heart, are likely to have had an effect, as both factors could have contributed to the variation in expression in non-cartilage tissues. In addition, the approach presented here treated the three expression platforms, U133A, U133B and U133 2.0 equally from the viewpoint of the threshold for cartilage-selectivity. Because the comparative dataset of normal tissues was both broader and deeper for the 133A platform, additional genes from this platform, albeit with greater than 50% CV, could be considered to be tissue-selective (e.g. *COL2A1*). Thus, a platform independent threshold would likely yield additional genes of interest within the U133A dataset. Finally, tissue-specific genes were identified using only microarrays and a single generalized algorithm. Additional genes selectively expressed in cartilage could be identified by less stringent criteria or other methods.

Conclusion

Genome-scale comparative expression analysis using human fetal cartilage and a broad set of normal human tissues has identified 161 cartilage-selective genes, including 27 uncharacterized genes. The data identify novel gene products that may provide essential roles in normal skeletogenesis and suggest new candidates for the over 100 inherited skeletal disorders in which the disease gene has not been identified. The results demonstrate that fetal cartilage is a complex and transcriptionally active tissue, and that the set of genes selectively expressed in cartilage has been greatly underestimated.

Methods

A flow chart outlining methods and results as well as other supplemental information is provided in additional data file 1.

Cartilage specimen collection and processing

Seven independent 18–22 week normal human fetal cartilage samples were studied under an Institutional Review Board approved protocol. Cartilage from the distal femur was dissected to remove bone and any adherent non-cartilage tissue (Figure 1). RNA was isolated and purified as previously described [4] and the quality and quantity of RNA were confirmed using an Agilent 2100 bioanalyzer and a Nanodrop ND-1000 spectrophotometer, respectively. Probe labeling, microarray hybridization, washing and scanning were carried out as detailed in Affymetrix protocols [33]. Five samples were used to probe Affymetrix™ U133 Plus 2.0 microarrays; and two samples were used to probe the Affymetrix™ Human Genome U133A/B set. Annotations were from version 11/15/06. The data are publicly available in the GEO database series

[GEO:GSE6565]. An additional sample was fixed in formalin, sectioned and stained with toluidine blue.

Non-cartilage microarray data

This project made use of the Celsius database [34,35], which is a database of publicly available microarray datasets from Gene Expression Omnibus, Array Express, and individual databases. Only CEL files are entered into the database, permitting reprocessing using identical algorithms to enable experimental comparisons. Only data from Affymetrix™ Human Genome U133A/B and Plus 2.0 platforms that contained clear annotation that they were derived from normal human tissues were selected for this analysis.

Microarray analysis

Data normalization and transformation

Raw data were normalized using the RMA algorithm with default parameters, available as part of the Bioconductor R library [36,37]. In brief, each CEL file was processed separately with an invariant pool of 50 arrays from a matching platform. Higher signal intensities observed in a subset of U133A non-cartilage samples from one provider [38] were additionally normalized by subtracting the median from other non-cartilage samples. After normalization, the training dataset was \log_2 transformed prior to analysis.

Median derived analog of CV for cartilage-selectivity ranking

In highly expressed cartilage genes, the degree of cartilage selectivity was defined as a median derived analog of CV (average deviation/median) applied to expression of these genes in non-cartilage tissues. The analog of CV was used to allow for greater tolerance of gene expression in some cartilage containing tissues without affecting the cartilage-selective assessment. A CV of 50% was empirically determined as a mathematically acceptable threshold for cartilage selectivity for probes across all three validation datasets (U133A, U133B, and U133 Plus 2.0).

Training

For the unsupervised analysis, probe intensities in all tissues were subtracted by the median probe intensity in cartilage, so all expression was defined relative to cartilage (i.e. the median of cartilage expression was set at zero) for each selected probe. Probesets with the greatest variation across all tissues and whose expression in at least two samples differed by two standard deviations from their mean expression across the entire set of samples were selected. Two-way hierarchical clustering was performed using Pearson's correlation to group genes and arrays based on the similarity of their expression patterns [39]. For the supervised analysis, the significance analysis of microarrays (SAM) two class method [40] was applied. 100 hun-

dred permutations were used and at least three-fold variation between cartilage and non-cartilage expression was required.

Validation

The probeset expression profiles for the 2,446 probesets identified in the training set (see Results) were acquired from 224 arrays representing 34 different tissues on three different platforms: Affymetrix U133A, U133B and U133 Plus 2.0. The data from the first platform, U133A, consisted of 1363 probeset profiles from 124 arrays, representing two normal fetal cartilage and 122 normal non-cartilage (32 tissues) samples (see additional data file 3, for the tissue distribution of samples used in this analysis). Arrays represented in this dataset were mostly from two large normal tissue expression profiling projects [38]. The second dataset, U133B, consisted of 882 expression profiles identified on 72 U133B microarrays from two normal fetal cartilage and 74 normal non-cartilage samples. These samples were primarily from the UCLA normal tissue microarray project (Chen, Day and Nelson, unpublished). The U133 2.0 dataset consisted of 201 expression profiles obtained from 26 Human Genome U133 Plus 2.0 arrays representing expression from five normal fetal cartilage and 21 non-cartilage (eight different tissue types) samples. The five cartilage arrays for the U133 2.0 platform were technical replicates of the arrays used in the training dataset. The non-cartilage arrays were a subset of the training dataset set aside for this validation only. From the 2446 probesets selected, probesets that exhibited at least a five-fold difference between the average cartilage intensity and the median signal intensity of all non-cartilage tissues, were selected. Cartilage-specificity was then determined using a median-derived analog of coefficient of variation (CV) as described above. Probesets with less than 50% CV were defined as reflecting cartilage-selective expression.

qRT-PCR

One microgram of RNA from seven tissues (brain, prostate, kidney, liver, heart, thyroid, and testis) in the FirstChoice® Human Total RNA survey panel (Ambion) was reverse transcribed using a high-capacity cDNA archive kit (ABI) and random primers. For cartilage, RNA from three independent cartilage samples was pooled and reverse transcribed. Amplification reactions were performed in triplicate using 100 ng of each cDNA. Thirty-five cycles of amplification were carried out in an ABI 7300 using the validated QuantiTect Gene Expression Assays and SYBR Green PCR kit (Qiagen). To assess specificity, amplification products were subjected to melting curve analysis and gel electrophoresis. The $2^{-[\Delta\Delta Ct]}$ method was employed to calculate relative amplification. This was performed using an average of endogenous references (18S, GAPDH, and HPRT1) to improve normalization

across the panel of tissues used [41]. For genes where no amplification was detected in a tissue, a Ct value of 35 was assigned, reflecting the maximum number of cycles carried out.

Abbreviations

CV: coefficient of variation; SAM: significance analysis of microarrays

Authors' contributions

VF designed and carried out the bioinformatics analysis and qRT-PCR, and drafted the manuscript. AD participated in data acquisition and normalization. DK obtained, dissected and isolated the cartilage RNA and critically revised the manuscript. ZAC participated in the bioinformatics analysis. ZC participated in generation of microarray data. SN participated in the design and coordination of the study and critically revised the manuscript. DC conceived the study, participated in its design and coordination, performed microarray analysis and critically revised the manuscript. All authors read and approved the final manuscript.

Additional material

Additional File 1

A flow chart illustrating a summary of the analysis. An unsupervised (black arrows) and supervised analysis (blue arrows) were performed with gene expression from 46 U133 2.0 Affymetrix arrays. An independent validation set comprised of 224 Affymetrix arrays (dashed arrows) was also used to test the 1713 genes for the most robust fetal cartilage selective genes.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-8-165-S1.pdf>]

Additional File 2

Supervised analysis and summary of in silico validation. Ranked list of cartilage genes more highly expressed by at least three fold in cartilage than non-cartilage tissues in the training dataset (five fetal cartilage samples compared to 41 normal non-cartilage samples). Ranked order is based on expression profiles obtained from 46 U133 Plus 2.0 arrays analyzed with SAM 2 class analysis with 100 permutations and with a False Discovery Rate (FDR) of 0. (B) Each probeset was independently evaluated in the validation datasets five fold higher expression using independent samples and three independent platforms as outlined in methods. Present indicates probe is identified in validation platform; Enriched indicates gene is expressed five fold higher in cartilage than non-cartilage tissues; Cartilage selectivity indicates at least five fold higher expression in cartilage than non-cartilage with a CV score of < 50% in non-cartilage samples. (C) "X" denotes gene was identified in a fetal cartilage cDNA library.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-8-165-S2.pdf>]

Additional File 3

Tissue distribution training and validation sets. (A) 31 Non-cartilage tissues and 124 arrays were used for the validation of cartilage selective genes identified on the U133A chip. Two fetal cartilage samples were compared against 122 non-cartilage samples. (B) 27 non-cartilage Tissues and 74 arrays were used for the validation of cartilage selective genes identified on the U133B chip. Two fetal cartilage samples were compared against 72 non-cartilage arrays (C) Eight non-cartilage tissues and 26 arrays used for the validation of cartilage selective genes identified on the U133B chip. Five fetal cartilage samples were compared against 72 non-cartilage arrays.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-8-165-S3.pdf>]

Acknowledgements

The authors thank Louis Fridkis and Brian O'Connor for their contributions to the CELSIUS database. The authors also thank the UCLA microarray core for assistance with the generation and analysis of the microarray data. This work was supported in part by grants from the NIH (HD22657 and RR00425 to DHC and DK) and (HL072367 and U24NS052108 to SFN) and DK was supported by the Joseph Drown Foundation. AD was supported by a grant from the NSF UCLA-IGERT (DGE-9987641). DHC and DK are recipients of Winnick Family Clinical Scholars Awards.

References

1. Goldring MB, Tsuchimochi K, Ijiri K: **The control of chondrogenesis.** *J Cell Biochem* 2006, **97**(1):33-44.
2. Pacifici M, Koyama E, Iwamoto M, Gentili C: **Development of articular cartilage: what do we know about it and how may it occur?** *Connect Tissue Res* 2000, **41**(3):175-184.
3. Pogue R, Sebald E, King L, Kronstadt E, Krakow D, Cohn DH: **A transcriptional profile of human fetal cartilage.** *Matrix Biol* 2004, **23**(5):299-307.
4. Krakow D, Sebald ET, Pogue R, Rimoin LP, King L, Cohn DH: **Analysis of clones from a human cartilage cDNA library provides insight into chondrocyte gene expression and identifies novel candidate genes for the osteochondrodysplasias.** *Mol Genet Metab* 2003, **79**(1):34-42.
5. Zhang H, Marshall KW, Tang H, Hwang DM, Lee M, Liew CC: **Profiling genes expressed in human fetal cartilage using 13,155 expressed sequence tags.** *Osteoarthritis Cartilage* 2003, **11**(5):309-319.
6. Tagariello A, Schlaubitz S, Hankeln T, Mohrmann G, Stelzer C, Schweizer A, Hermanns P, Lee B, Schmidt ER, Winterpacht A, Zabel B: **Expression profiling of human fetal growth plate cartilage by EST sequencing.** *Matrix Biol* 2005, **24**(8):530-538.
7. Wang Y, Middleton F, Horton JA, Reichel L, Farnum CE, Damron TA: **Microarray analysis of proliferative and hypertrophic growth plate zones identifies differentiation markers and signal pathways.** *Bone* 2004, **35**(6):1273-1293.
8. James CG, Appleton CT, Ulici V, Underhill TM, Beier F: **Microarray analyses of gene expression during chondrocyte differentiation identifies novel regulators of hypertrophy.** *Mol Biol Cell* 2005, **16**(11):5316-5333.
9. Stokes DG, Liu G, Coimbra IB, Piera-Velazquez S, Crowl RM, Jimenez SA: **Assessment of the gene expression profile of differentiated and dedifferentiated human fetal chondrocytes by microarray analysis.** *Arthritis Rheum* 2002, **46**(2):404-419.
10. Lohmander LS: **What can we do about osteoarthritis?** *Arthritis Res* 2000, **2**(2):95-100.
11. Reginato AM, Olsen BR: **The role of structural genes in the pathogenesis of osteoarthritic disorders.** *Arthritis Res* 2002, **4**(6):337-345.
12. Superti-Furga A, Unger S: **Nosology and classification of genetic skeletal disorders: 2006 revision.** *Am J Med Genet A* 2006.

13. Brazma A, Vilo J: **Gene expression data analysis.** *FEBS Lett* 2000, **480(1)**:17-24.
14. Liang S, Li Y, Be X, Howes S, Liu W: **Detecting and profiling tissue-selective genes.** *Physiol Genomics* 2006, **26(2)**:158-162.
15. Hsiao LL, Dangond F, Yoshida T, Hong R, Jensen RV, Misra J, Dillon W, Lee KF, Clark KE, Haverty P, Weng Z, Mutter GL, Frosch MP, Macdonald ME, Milford EL, Crum CP, Bueno R, Pratt RE, Mahadevappa M, Warrington JA, Stephanopoulos G, Stephanopoulos G, Gullans SR: **A compendium of gene expression in normal human tissues.** *Physiol Genomics* 2001, **7(2)**:97-104.
16. Shyamsundar R, Kim YH, Higgins JP, Montgomery K, Jordan M, Sethuraman A, van de Rijn M, Botstein D, Brown PO, Pollack JR: **A DNA microarray survey of gene expression in normal human tissues.** *Genome Biol* 2005, **6(3)**:R22.
17. Zhao SH, Recknor J, Lunney JK, Nettleton D, Kuhar D, Orley S, Tugle CK: **Validation of a first-generation long-oligonucleotide microarray for transcriptional profiling in the pig.** *Genomics* 2005, **86(5)**:618-625.
18. Saito-Hisaminato A, Katagiri T, Kakiuchi S, Nakamura T, Tsunoda T, Nakamura Y: **Genome-wide profiling of gene expression in 29 normal human tissues with a cDNA microarray.** *DNA Res* 2002, **9(2)**:35-45.
19. Teixeira CC, Ischiropoulos H, Leboy PS, Adams SL, Shapiro IM: **Nitric oxide-nitric oxide synthase regulates key maturational events during chondrocyte terminal differentiation.** *Bone* 2005, **37(1)**:37-45.
20. Sims NA, Jenkins BJ, Quinn JM, Nakamura A, Glatt M, Gillespie MT, Ernst M, Martin TJ: **Glycoprotein 130 regulates bone turnover and bone size by distinct downstream signaling pathways.** *J Clin Invest* 2004, **113(3)**:379-389.
21. Zelzer E, Mamluk R, Ferrara N, Johnson RS, Schipani E, Olsen BR: **VEGFA is necessary for chondrocyte survival during bone development.** *Development* 2004, **131(9)**:2161-2171.
22. Kanegae Y, Tavares AT, Izpisua Belmonte JC, Verma IM: **Role of Rel/NF-kappaB transcription factors during the outgrowth of the vertebrate limb.** *Nature* 1998, **392(6676)**:611-614.
23. Wu S, De Luca F: **Inhibition of the Proteasomal Function in Chondrocytes Down-Regulates Growth Plate Chondrogenesis and Longitudinal Bone Growth.** *Endocrinology* 2006, **147(8)**:3761-3768.
24. Grimaud E, Blanchard F, Charrier C, Gouin F, Redini F, Heymann D: **Leukaemia inhibitory factor (lif) is expressed in hypertrophic chondrocytes and vascular sprouts during osteogenesis.** *Cytokine* 2002, **20(5)**:224-230.
25. Dagoneau N, Scheffer D, Huber C, Al-Gazali LI, Di Rocco M, Godard A, Martinovic J, Raas-Rothschild A, Sigaudy S, Unger S, Nicole S, Fontaine B, Taupin JL, Moreau JF, Superti-Furga A, Le Merrer M, Bonaventure J, Munnich A, Legeai-Mallet L, Cormier-Daire V: **Null leukemia inhibitory factor receptor (LIFR) mutations in Stuve-Wiedemann/Schwartz-Jampel type 2 syndrome.** *Am J Hum Genet* 2004, **74(2)**:298-305.
26. Leung TH, Hoffmann A, Baltimore D: **One nucleotide in a kappaB site can determine cofactor specificity for NF-kappaB dimers.** *Cell* 2004, **118(4)**:453-464.
27. Belmonte N, Phillips BW, Massiera F, Villageois P, Wdziekonski B, Saint-Marc P, Nichols J, Aubert J, Saeki K, Yuo A, Narumiya S, Ailhaud G, Dani C: **Activation of extracellular signal-regulated kinases and CREB/ATF-1 mediate the expression of CCAAT/enhancer binding proteins beta and -delta in preadipocytes.** *Mol Endocrinol* 2001, **15(11)**:2037-2049.
28. Gliki G, Abu-Ghazaleh R, Jezequel S, Wheeler-Jones C, Zachary I: **Vascular endothelial growth factor-induced prostacyclin production is mediated by a protein kinase C (PKC)-dependent activation of extracellular signal-regulated protein kinases 1 and 2 involving PKC-delta and by mobilization of intracellular Ca2+.** *Biochem J* 2001, **353(Pt 3)**:503-512.
29. Gupta P, Prywes R: **ATF1 phosphorylation by the ERK MAPK pathway is required for epidermal growth factor-induced c-jun expression.** *J Biol Chem* 2002, **277(52)**:50550-50556.
30. Wang Y, Prywes R: **Activation of the c-fos enhancer by the erk MAP kinase pathway through two sequence elements: the c-fos AP-1 and p62TCF sites.** *Oncogene* 2000, **19(11)**:1379-1385.
31. Owens JM, Matsuo K, Nicholson GC, Wagner EF, Chambers TJ: **Fra-1 potentiates osteoclastic differentiation in osteoclast-macrophage precursor cell lines.** *J Cell Physiol* 1999, **179(2)**:170-178.
32. Eferl R, Hoebertz A, Schilling AF, Rath M, Karreth F, Kenner L, Amling M, Wagner EF: **The Fos-related antigen Fra-1 is an activator of bone matrix formation.** *Embo J* 2004, **23(14)**:2789-2799.
33. **Affymetrix Technical Support Protocols** [<http://www.affymetrix.com/>]
34. Day A, Carlson MRJ, Dong J, O'Connor BD, Nelson SF: **Celsius: A community resource for Affymetrix microarray data.** *Genome Biology* in press.
35. **Celsius Microarray Database** [<http://genome.ucla.edu/projects/celsius/>]
36. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, **19(2)**:185-193.
37. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smyth C, Smyth G, Tierney L, Yang JY, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5(10)**:R80.
38. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci U S A* 2004, **101(16)**:6062-6067.
39. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization.** *Mol Biol Cell* 1998, **9(12)**:3273-3297.
40. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci U S A* 2001, **98(9)**:5116-5121.
41. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paep A, Speleman F: **Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes.** *Genome Biol* 2002, **3(7)**:RESEARCH0034.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

