# **Data Curation Profile – Human Genomics**

Profile Author	J. Carlson			
Profile Author	N. Brown			
Institution Name	Purdue University			
Contact	J. Carlson, jrcarlso@purdue.edu			
Date of Creation	October 27, 2009			
Date of Last Update				
Version	1.0			
Discipline / Sub- Discipline				
Purpose	Data Curation Profiles are designed to capture requirements for specific data generated by a single scientist or scholar as articulated by the scientist him or herself. They are also intended to enable librarians and others to make informed decisions in working with data of this form, from this research area or sub-discipline. Data Curation Profiles employ a standardized set of fields to enable comparison; however, they are designed to be flexible enough for use in any domain or discipline.			
	A profile is based on the reported needs and preferences for these data. They			
Context	are derived from several kinds of information, including interview and document data, disciplinary materials, and standards documentation.			
Sources of Information	<ul> <li>An initial interview with the scientist conducted in June, 2008</li> <li>A second interview with the scientist conducted in January, 2009.</li> <li>A questionnaire completed by the scientist as a part of the second interview.</li> <li>A published paper explaining the research and the methodology used to gather, process and analyze the data set in question.</li> </ul>			
Scope Note	The scope of individual profiles will vary, based on the author's and participating researcher's background, experiences, and knowledge, as well as the materials available for analysis.			
Editorial Note	Any modifications of this document will be subject to version control, and annotations require a minimum of creator name, data, and identification of related source documents.			
Author's Note	This Human Genomics data curation profile is based on analysis of interviews, a completed worksheet and information gathered from a publication, collected from a researcher working in this research area or sub-discipline. Some sub- sections of the profile were left blank; this occurs when there was no relevant data in the interview or available documents used to construct this profile.			
URL	http://www.datacurationprofiles.org			

# Brief summary of data curation needs

The data set consists of a mySQL database and several text files containing data used for reference purposes.

The data are currently posted on the scientist's personal web site after the results of her work have been published. However, she has stated that she does not have the resources or expertise needed to maintain the data adequately. Once the scientist is confident that the results of her research are sound, she would like to integrate her data into other genomic data repositories.

The scientist believes that there is a need for archiving her data and in providing resources and mechanisms to enable data archiving to researchers at her institution.

Note: The scientist views data management and curation primarily from a technical perspective. Although the scientist is happy that the Libraries are interested in data, she perceives the primarily responsibility for addressing data issues falling to the campus IT department.

# Overview of the research

## **Research area focus**

The scientist describes her work as generating the criteria needed to do experiments in human genomic research. Her research focuses on the discovery of codes that regulate the expression of human genes. This research is done through collecting and analyzing "9-mers", segments of DNA code (A, G, C, T) 9 bases long, from specific regions and making predictions about which ones seem to function as specific codes to turn a gene's function on or off.

As her research is computational rather than experimental in nature, much of her work centers on collecting and analyzing data that's publicly available in human genome browsers.

## Intended audiences

Other researchers who are working in the field of human genomics.

#### **Funding sources**

None, although the scientist may seek funding for her work in this area in the future. As she has not received outside funding for this research, the scientist is not mandated to generate a data management plan or share her data with others outside of her lab.

# Data kinds and stages

#### Data narrative

The first step in the scientist's research process is to generate a reference table of all possible 9mers that may occur in DNA.

Next, two sets of 9-mers were obtained from the National Center of Biotechnology's (NCBI) nucleotide database and UCSC's genomic browser. The first set of 9-mers was selected from DNA sequences containing known promoters with accurately defined transcription initiation sites as identified by other researchers. This data set is smaller but more likely to provide valid results. The second set of 9-mers was selected using proximal promoters that were based on predicted transcription initiation sites. 9-mers are then "cleaned" by using Perl scripts. This process included checking that a beginning of a gene was assigned accurately in the first set, and eliminating any incomplete or ambiguous codes as well as any redundancies in the second set.

Once the data sets were "clean", they were placed into a mySQL database running in Linux. Data collection, access, retrieval, management, and analysis were done using Perl scripts and modules obtained from Bioperl (<u>http://www.bioperl.org</u>). The data were then mathematically filtered using Perl scripts to process the data. Processing consists of comparing 9-mers present in the whole human genome to the 9-mers that are located within genes themselves and how often they occur. 9-mers that occur more frequently in proximity to gene promoters and do not appear frequently in the genome are identified as candidates for analysis.

After processing, the scientist analyzes the 9-mers by generating density plots showing which 9mers are most often found in proximal promoters. The density plots of 9-mers most often found in proximal promoters are then compared to the density of the number of occurrences of each 9mers in human DNA to see if there are any statistically significant differences between the two sets. A statistically significant difference indicates that a particular 9-mer may be involved in gene regulation and warrants further investigation.

The scientist was unable to answer questions regarding the size of the files as she does not manage her files directly.

The categories in the "data stage" column listed in the table below were developed by the authors of this data curation profile. The data specifically designated by the scientist to make publicly available are indicated by the rows shaded in gray.

Data Stages	Output	Typical File Size	Format	Other / Notes	
				Computationally	
				generated all possible 9-	
Defense	A reference table of	5 50140		mers that may appear in	
Reference	all possible 9-mers	5.53MB	.tXt	DNA.	
				Data gathered from	
				genomic browsers	
D	I wo base sets of 9-		1.	according to specific	
Raw	mers		.XIS	criteria.	
				Raw data are checked	
	mySQL database of			and cleaned using peri	
Oleanad	9-mers meeting			scripts and deposited	
Cleaned	scientist's criteria		sqi	Into a mySQL database.	
	9-mers identified				
	according to their			Data and anonanal far	
Drassad	frequency			Data are prepared for	
Processed	frequency		sqi, .txt		
				I ne analyzed statistics	
				Include density plots	
				generated from tools	
Analyzad	Density plate				
Analyzed	Density plots		sqi	genomic data repository	
Assume stations Date					
Augmentative Data					
	Collections of filtered				
Supplementary	data that serve as			Files are publicly	
Data Files for	reference documents			available through	
Reference	for the scientist.		.txt	scientist's web site	

**Note:** The data specifically designated by the scientist to make publicly available are indicated by the rows shaded in gray. Empty cells represent cases in which information was not collected or the scientist could not provide a response.

## Target data for sharing

The mySQL database developed by the scientist was made available from her website, but is currently inaccessible. The scientist stated that she lacks the resources to maintain this database effectively.

To support this research, the scientist developed files listing all possible 9-base elements that may occur in DNA (the reference data set), 9-mers that were found three or more times in proximal promoters of human genes and the predicted correspondence of the collected 9-mers to potential transcription factor binding sites. These files are also publicly available and are currently accessible.

## Use/re-use value of the data

According to the scientist her work thus far has been primarily exploratory in nature. Therefore she has not yet contributed her data to genome browsers such as the one at UCSC. She wants to conduct further evaluations to provide more evidence and help her to feel confident that her predictions are accurate. She does plan to contribute her data (the database and the density plots) to existing genomic browsers in the future. It is unclear if the scientist would be willing to make the density plots available to others before submitting them to genomic browsers.

However, the data that she generates is of value for other researchers engaged in similar areas of study. The work that the scientist has done to prepare genomic data for research in gene regulation could be used by others to do their own calculations and make their own discoveries in this area. The perl scripts used to prepare the data would also be of use to others doing similar research.

## **Contextual narrative**

The scientist's research produces a number of outputs related to the data in multiple formats. Data is collected from genomic browsers in spreadsheets and stored in this format until cleaned. The centerpiece of her data is the mySQL database containing "clean" data sets for her purposes. In addition to the reference data file, the scientist has also created several supplementary text files. These text files contain subsections of the data that have been relevant to the scientist's work, and could be used by others doing similar types of research. These text files include: the 9-mers that were found 3 or more times in proximal promoters of human genes and the predicted correspondence of the collected 9-mers to potential transcription factor binding sites.

Other outputs include the density plots generated through C++, some of which have been reproduced in publications, and the perl scripts used to clean the data, and prepare it for analysis. The genomic data gathered from genomic repositories are described by the scientist as "experimental". As such, the perl scripts are used to mathematically filter the data according to specific criteria to normalize the data in preparation for analysis.

The scientist has published a journal article which explains her methodologies in gathering and working with her data. The scientist has another article in press that focuses on the density plots and a previous book chapter based upon this research.

# Intellectual property context and information

## Data owner(s)

The researcher does not feel that she owns the data; in fact she feels that no one individual or institution really owns the data at all. According to her, the culture of the human genomic field is there is no "one owner" of the research data. The research culture supports, even expects, that data will be shared with others. The field of human genomic research is so new, potentially vast, and ripe for exploration that access to data is vital for the field to grow and develop. Any

restrictions on access to genomic data would stifle progress and slow the development of an important area of research.

#### Stakeholders

The systems manager who designed the perl scripts may be a stakeholder, although the scientist did not characterize him in this manner during discussions.

#### Terms of use (conditions for access and (re)use)

None stated. The data are currently made available from the scientist's personal website without any statements concerning access or re-use from the scientist.

#### Attribution

The ability to cite the dataset in publications is a high priority for the scientist.

# Organization and description of data for ingest (incl. metadata)

#### Overview of data organization and description

There are two different data types within the data set, a database and text files. There are also a set of perl scripts included with the data that were used to process it. The scientist stated that she uploaded the data to the genomic browser at UCSC to use its tools to analyze her data, which may imply that the structure of the database is compatible with the structure of the UCSC's browser. The scientist would like to integrate her data with the UCSC browser at some point in the future when she is more confident with her results. The text files contain the citation to the paper at the top of the file and some explanatory text about the data.

The application of standardized genomics metadata to the dataset is a high priority for the scientist.

#### Formal standards used

Unknown. The metadata fields in this database may be compatible with the genomic browser at UCSC, however this cannot be confirmed as of this writing because public access to the scientist's database is currently offline.

#### Locally developed standards

Unknown.

#### Crosswalks

Unknown. It is unclear at this time if some form of crosswalk would be needed to connect the scientist's data with the UCSC genomic browser.

#### Documentation of data organization/description

The methodology behind data acquisition, processing and analysis is described in a published article. Although this description does not provide detailed information on the organization or description of the data, the scientist believes that it does provide enough information for others to understand and use the data properly.

As the published paper(s) serve as the documentation of the data, the scientist would like to be able to link the data to the publication in some manner.

## Ingest

The ability for the scientist to submit the data to a repository herself is a higher priority for her than the ability to automate the process of data submission into a repository.

# Access

#### Willingness / Motivations to share

The scientist is willing to share the data once the findings from the research have been published. Her primary reluctance to share her data any earlier is that she wants to be sure that the data are usable and have value.

The scientist does currently post her data, and the perl scripts used in her research, publicly on her personal website for others to discover and access. As of this writing however access to the database was not functioning properly.

## Embargo

Once the publication is out, the scientist does not require any embargo of the data.

#### Access control

Once the findings have been published she is willing to share the data with anyone without restriction (although she would like others to cite the publication that uses the data, if they use the data themselves).

#### Secondary (Mirror) site

A secondary data access (mirror) site that would allow continued access to the data if the repository is off-line is a high priority for the scientist.

# Discovery

Although the scientist has made her data publicly available through her website, she does not have a sense of how people discover the data. She does refer people to the data at conferences and refer people to her website in her publications.

The ability for researchers within her field and outside of her field to easily find the data is a high priority for the scientist. The ability for people to find the data through internet search engines is a high priority for the scientist.

The scientist makes her data accessible from her web page in a mySQL database, which could provide insight about the desired search functionality for her data. However, the database is off line as of this writing.

# Tools

The perl scripts used by the scientist to analyze the data are made available alongside of the data.

The mySQL database was developed in a Linux environment.

There are tools available through genomic browsers at NCBI, UCSC and others, to process and analyze this type of data. The ability to connect the data to these tools is a high priority for the scientist.

The scientist selected "I don't know or NA" in response to the question on selecting a level of priority for a data repository to support the use of web services APIs.

# Interoperability

Research in genomics is highly dependent upon access to previous work and data gathered by others in the field. According to the scientist, there is a cultural expectation in the genomics field that data will be made available for analysis by other researchers. Linkages between datasets and tools to process and/or analyze the data are a high priority for the scientist.

According to the scientist, the particular dataset under discussion is too specialized and too experimental in its current state of development to be incorporated with the NCBI or other genomic repository. However, once the scientist is confident that her predictions are accurate she does intend to deposit her data analysis with UCSC. Therefore, the scientist's data may need to be aligned with the structures and standards used at UCSC if this has not been done already.

As the published paper(s) serve as the documentation of the data, the scientist would like to be able to link the data to the publication in some manner.

The scientist responded "I Don't Know or NA" to the question about the ability for the repository to support the use of web services APIs.

# **Measuring impact**

#### **Usage statistics**

The ability to view usage statistics is a high priority for the scientist. The scientist currently employs a usage counter on the database she has made publicly available from her website.

#### Gathering information about users

Gathering information about the users of her data was not addressed by the scientist.

## Data management

## Security / Back-ups

The database is currently backed up on a regular basis (although the scientist was not sure how frequently back-ups occurred). The scientist is aware of the need to plan for data migration but has not taken steps to draft such plans.

## Secondary storage site

A secondary data storage site is a high priority for the scientist. A secondary data storage site at a different geographic location is also a high priority.

# Preservation

#### **Duration of preservation**

The scientist would like her data to be preserved for more than 5 years but less than 10 years, depending on the data's intrinsic value over time.

#### Data provenance

The scientist selected "I don't know or NA" in response to the question on assigning a priority level for documenting any and all changes made to her data over time.

## Data audits

The ability to audit the dataset is a high priority for the scientist.

### **Version control**

The ability of the repository to provide version control for this data set is a high priority for the scientist.

#### **Format migration**

The ability to migrate the dataset into new formats over time is a high priority for the scientist.

**Personnel** – This section is to be used to document roles and responsibilities of the people involved in the stewardship of this data. For this particular profile, information was gathered as a part of a study directed by human subject guidelines and therefore we are not able to populate the fields in this section.

### Primary data contact (data author or designate)

Data steward (ex. library / archive personnel)

**Campus IT contact** 

Other contacts

Notes on personnel