This is a preprint of an article accepted for publication in Journal of Documentation. Huang, H. (in press, 2014). Domain knowledge and data quality perceptions in genome curation work. Journal of Documentation.

Domain Knowledge and Data Quality Perceptions in Genome Curation Work

Hong Huang

School of Information, University of South Florida, Tampa, Florida, 33620.

Telephone: (813) 974-3520; Fax: (813) 974-6840; E-mail: honghuang@usf.edu

Abstract

Purpose- This article aims at understanding genomics scientists' perceptions in data quality assurances based on their domain knowledge.
Design/methodology/approach- The study used a survey method to collect responses from 149 genomics scientists grouped by domain knowledge. They ranked the top-five quality criteria based on hypothetical curation scenarios. The results were compared using Chi-Square analysis.
Findings- Scientists with domain knowledge of biology, bioinformatics, and computation did not reach a consensus in ranking data quality criteria. Findings showed that biologists cared more about curated data that can be concise and traceable. They were also concerned about skills dealing with information overloading. Computational scientists on the other hand value making curation understandable. They paid more attention to the specific skills for data wrangling.
Originality/value- This study takes a new approach in comparing the data quality perceptions for scientists across different domains of knowledge. Few studies have been able to synthesize models to interpret data quality perception across domains. The findings may help develop data quality assurance policies and training seminars and maximize the efficiency of genome data management.

Keywords Genome curation, Interdisciplinary, Domain knowledge, Data quality dimensions, Data quality skills, Scientist behaviors

Introduction

The proliferation of heterogeneous genomic data types represents the diverse concepts of biology (Sanderson, 2011; Wu *et al.*, 2010; Yang *et al.*, 2011). Genome curation is the process of digitizing and integrating disparate pieces of genomic data and their related literatures to

facilitate the sharing of genomic knowledge (Reed *et al.*, 2006). The genome curation process can be facilitated by using standardized terminologies and metadata schemas (MacMullen and Denn, 2005; Pagani *et al.*, 2012; Willis *et al.*, 2012). There are well established terminologies (e.g., gene ontology) and metadata standards in biosciences for describing data-types, protocols used in experiments, and gene ontology for molecular functions (Leonelli *et al.*, 2011; Mayor and Robinson, 2014). It is a complex process that requires multidisciplinary knowledge, pertinent work experience, and skills relevant to the effective execution of multi-faceted curation operations (Burkhardt *et al.*, 2006). Thus, genomic research has become a data rich domain requiring not only effective methods to process, interpret, and reuse genomic data (Salimi and Vita, 2006; Samuel *et al.*, 2008), but extensive knowledge of the fields of biology, bioinformatics, and computational science.

Scientists working on genome curation require domain knowledge in areas such as biology, bioinformatics and computational science. Scientists conducting genome curation generally possess either PhDs or Masters degrees in biology, bioinformatics, computer science, or other related disciplines (Burge *et al.*, 2012). Wet-lab research experience in biochemistry and molecular biology contributes meaningfully to their collective ability to determine and select the desired information resources that can help their curation or annotation work (Burge *et al.*, 2012). In certain cases, subject expertise or domain knowledge is essential to ensure acceptable upstream phases of genomic data management and planning (Bentley, 2006). Genomics scientists with a wide array of experience, participate in comprehensive training and workshops in order to improve their curation skills (Sanderson, 2011; Shimoyama *et al.*, 2009). They also consult the curation manual regularly to ensure that they follow curation standards in identifying data elements, assigning nomenclature, and annotating genomic-related data with biological information (Samuel and Klumke, 2008).

It has been found that genomics scientists have shared certain requirements for data quality, leading to the development of a general data quality model for genome curation (Huang *et al.*, 2012).Within genome curation, the context for both information use and information operation is complicated. As a result of varieties of domain knowledge exists among genomic scientists, the respective complexities of domain knowledge and work experience might consequently affect scientists' decision making. The relationship between domain knowledge types for genome curation and data quality assurance activities remains unknown. Scientists from different domains and backgrounds could make conflicting data quality decisions when assigning the same genome curation tasks, resulting in the current gap in understanding of the curation problems associated with data quality assurance when different domain knowledge is exchanged among biology, bioinformatics, and computational science.

The purpose of this study is to understand the relationship between different types of domain knowledge and scientists' data quality requirements. Specifically, the respective performances of three different user groups, who possess domain knowledge in the fields of computational science, bioinformatics, and biology, will be examined in order to identify their perceptions of data quality requirements. The findings could benefit the development of domain sensitive data quality and skill models for genomic research communities, yielding both improved resource integration and more cost-effective collaborative solutions.

Literature Review

Scientists conducting genome curation work have been trained in their disciplinary knowledge (e.g., biology, computer science) at the post-graduate degree level or higher. Biological research has progressed to an intensive data process and evaluation using multiple data mining tools. The data-driven approach has become a common research practice for scientists (Reed *et al.*, 2006; Goth, 2012). Data curation and manipulation tools need to be customized by scientists to fit into a specific biological context (Lathe *et al.*, 2008; Huang *et al.*, 2011; Pruitt *et al.*, 2012). Biologists also need data analysis support from computational scientists to process the massive data sets produced through their research. The task is not easy because the traditions and cultures of these domains are not the same (Wooley and Lin, 2005). Genomics scientists need much closer scrutiny to explicate the characteristics of domain knowledge in both biology and computer science. It is through such scrutiny that they can adopt effective practices for data quality assurance and data exchange among distinct disciplines.

Domain Knowledge in Genome Curation

Genomics research has grown and changed rapidly. Genomic data curation originally started as sequence analysis only (Reed *et al.*, 2006). It has since incorporated a wide variety of data processes and analysis such as genome-wide association studies, microarrays, protein-protein interactions, and literature text-mining (Cole and Bawden, 1996; Bartlett and Toms, 2005; Ioannidis and Khoury 2011; Lathe *et al.*, 2008; Sanderson, 2011; Shachak and Fine, 2008). Curating genomic data is a highly interdisciplinary process requiring scientists to have diverse skills.

Domain knowledge can be defined as the degree of familiarity with a particular domain or subject area (Allen, 1991; Ju, 2007; Wildemuth, 2004). It encompasses declarative knowledge (*knowing what*), procedural knowledge (*knowing how*), and conditional knowledge (*knowing when and where*), (Alexander, 1992; Hjørland and Albrechtsen, 1995). Domain and discipline knowledge seem to "fall along a continuum that is defined by both external and internal factors" (Alexander, 1992, p36). Relationships within a domain, the rules of that domain, and its historical context all need to be considered to embrace the complete and meaningful domain

knowledge of a discipline (Hjørland and Albrechtsen, 1995). Genome science is an interdisciplinary field that requires collaborative work with both biologists and computational scientists. Wooley and Lin (2005) distinguish biologists from computational scientists in their research goals and working practices (see table 1).

Table 1. Examples of working objectives and practices in biology and computational science (Wooley and Lin, 2005).

	Biology	Computational science
Working objectives	• Understand the mechanism of development for living organisms, and then use that understanding to determine examples of application areas for biological data. • Seek signals in the noise of their experimental data. • Provide solutions to individual and specific problems.	• Identify the unknown patterns within massive biological data sets. • Search for boundary conditions and constraints. • Develop universal solutions to solve many problems.
Working practices	• Research is driven by experiment and observation. • Question the mathematical soundness of their approach by providing exceptions to their cases. • Limited freedom to establish rules. • Use categorical statements informally.	• Research is driven by analytical methods and techniques. • May underestimate the complexity of the biological problems, oversimplify biological models and give out universal statements that fall short of expected exceptions. • Open to the establishment of their own rules for developing algorithms. • Take categorical statements literally.

Computational science develops algorithms and software tools to support data retrieval, organization and analysis (Fenstermacher, 2005). However, there are distinct sets of rules for

data configuration and operations between biology and computer science (Wooley and Lin, 2005). Biologists are particularly interested in seeking "signal in the noise of their experimental data" (Wooley and Lin, 2005, pp367). Since biological research is driven by experiment and observation, its goals consist of finding solutions to individual and specific problems. In contrast, computational scientists are trained to "search for boundary conditions and constraints" (Wooley and Lin, 2005, pp 367). Computational science research is driven by analytical methods and techniques, and its research goals are the development of solutions that can solve many problems. Computational scientists who work with biological data are trained to "take categorical statements literally, whereas biologists use them informally" (Wooley and Lin, 2005, pp367).

Because of the constraints imposed by nature, biology has limited freedom to establish rules. These constraints are consistent with the rules applied to the biological phenomena. In contrast, computer science is open to the establishment of one's own rules provided that doing so allows sense to be made of the algorithm (Wooley and Lin, 2005). Biologists might focus on understanding the mechanism of development for living organisms, and then use that understanding to determine examples of application areas for biological data (Wooley and Lin, 2005). In contrast, computational scientists are *data* scientists. They are more engaged in attempts to identify the unknown patterns within massive data sets (Wooley and Lin, 2005). As programmers, computational scientists could easily underestimate the complexity of the biological problems, and therefore both oversimplify biological models as well as give out universal statements that fall short of expected exceptions (Wooley and Lin, 2005). However biologists, particularly those untrained in quantitative sciences, always question the mathematical soundness of their approach by providing exceptions to their cases (Wooley and Lin, 2005). During the genome curation process, both the biologist and the computational scientists collaborate with each other. During this process however, they may experience conflicts and disagreements in defining curation roles and thus yield contested interpretation of curation data.

Previous research indicates that scientists' domain knowledge affects their information seeking behavior and their interactions with information systems and software tools (Brown, 2003; Hemminger *et al.*, 2007; Vibert *et al.*, 2009; Wu *et al.*, 2012). It can be assumed that biologists have a high level of declarative knowledge of biology, whereas computational scientists have a high level of procedural knowledge of computer systems. Although both biologists and computational scientists might be expected to know how to use computer programs or curation tools, computational scientists probably enjoy a broader knowledge of tools and programs. However, when a biologist interacts with a new curation tool, s/he holds the advantage over a computational scientist of knowing the particular semantics (the words or terminologies about biological concepts) used in that program (Bartlett and Toms, 2005; Chilana *et al.*, 2009). In other tasks such as accurately predicting the options available in a generic help

menu or in the functions of menu interface designed for automatic genome annotation systems, the biologist may be at a disadvantage compared to the computational scientist (Chilana *et al.*, 2009; Shachak and Fine, 2008).

Domain knowledge affects scientists' decisions in the determination of data processing strategies, data-quality assurance activities, analytic tools selection, and result evaluation (Chilana *et al.*, 2009; Ju, 2007; Vibert *et al.*, 2007, 2009; Wu *et al.*, 2012). The scientists with biology domain knowledge could easily find the exceptions or special cases (Wooley and Lin, 2005) for which annotation tools and guidelines might not yet be available. Similarly, computational scientists can benefit from the wet-lab experiences of biologists to develop both complex software tools and standardized workflows (Chilana *et al.*, 2009). Scientists need to remain open to explore new research opportunities in a typical domain as an "outsider", and develop strategies for exploring and translating information from unfamiliar domains to manage their interdisciplinary information work (Palmer and Neumann, 2002). Development of A comprehensive data curation model can help yield the high-quality curation products that both biology and computational science require. Such a data model encourages two domains to work closely with each other thereby reducing domain crossing barriers while merging knowledge across disciplinary boundaries (Haythornthwaite, 2006; Klein, 1996).

Data Quality and Domain Knowledge

There are different working domains and scholarly contexts through which data quality can be both operationalized and defined. It has been argued that data quality as a concept is contextual and must be evaluated within the context (Strong *et al.*, 1997; Stvilia *et al.*, 2007). An aspect of a DQ concept is defined as a DQ dimension (Huang *et al.*, 2012; Stvilia *et al.*, 2007; Wang and Strong, 1996). Several studies have assessed specific DQ dimensions in different domains. One study explored progress in the accuracy assessments of automated genome curation tasks (Brent, 2008), whereas another examined in an online interactive community, patterns in credibility (Lankes, 2008). Wang and Strong (1996, p 6) provided a definition for quality, describing it as "fitness for use." This indicated the importance of defining data quality within context of use (Strong *et al.*, 1997; Stvilia *et al.*, 2007). The need to comprehend the extent to which user satisfaction is realized has the potential to characterize data quality within a particular context (Evans and Lindsay, 2005; Huang *et al.*, 2012).

According to research, sets of DQ dimensions that have been determined to be important, include those pertaining to gene-ontology curation behaviors (MacMullen, 2006), online scholarly information (Rieh, 2002), and consumer health information (Frické and Fallis, 2004; Stvilia *et al.*, 2009). Genomics scientists suggest that trust-related dimensions such as Unbiased and Believability are important in genome curation when they indirectly assess the quality of curation data (Huang *et al.*, 2012). Data quality aspects related to trust help scientists gauge the degree of confidence they can have. Studies have shown that domain knowledge could facilitate

researchers in evaluating the trustworthiness of reference sources (Vibert *et al.*, 2012). Data standards, metadata schemas, and curated databases were developed to facilitate the accessibility of disparate genomic data sets (Barrett *et al.*, 2012; Willis *et al.*, 2012). DQ models were developed to describe and capture the overall value structure and the context for DQ for a genome curation community (Huang *et al.*, 2012), a Wikipedia community (Stvilia *et al.*, 2007) and online health information consumers (Stvilia *et al.*, 2009).

Lee and Strong (2003) have argued that three knowledge modes are related to data quality dimensions. According to Lee and Strong (2003), the *declarative*, or *knowing-what*, may be defined as understanding the activities through which the data production processes are realized. *Procedure*, or *knowing how*, is defined as understanding procedures needed to respond to known DQ difficulties and obstacles (Lee and Strong, 2003). *Knowing-why* is defined as contextual knowledge that can formulate the questions to understand related purposes and the ability to analyze underlying principles (Lee and Strong, 2003). During the data process, it has been found that the prioritization of DQ dimensions differs among users with varying knowledge modes (Lee and Strong, 2003). The genome curation community in fact requires a set of DQ skills to guarantee data quality itself. Genome curation work requires excellent written and verbal communication skills to facilitate the acquisition and description of genomics data. Knowledge in biology and/or bioinformatics also helps to evaluate quality control of experimental data. Genome curation work is data-driven; much of the scientists' time is spent on data wrangling or "munging", ie., dealing with the large scale of genomic data for data preprocessing, integration, and data cleaning and validation (Heer and Kandel, 2012; Reed *et al.*, 2006). Through a survey of DQ professionals holding a series of professional employment positions, Chung *et al.* in 2002 created a practical educational framework and described three useful DQ categories, each one pertaining to a particular set of capabilities of DQ skills, specifically *technical*, *adaptive* and *interpretive*.

Domain experts obtain domain-specific knowledge, work-related experience, and trainings. This experience or knowledge can also support data-quality related activities and allow domain experts to make greater use of data-quality information than those without related knowledge (Fisher *et al.*, 2003). Users with experience or domain knowledge might be sensitive in detecting both errors and missing data (Klein *et al.*, 1997; Sanbonmatsu *et al.*, 1992), adaptive in using contextual and relevant information (Sanbonmatsu *et al.*, 1992; Payne *et al.*, 1993), and proficient in organizing information (Mackay and Elam, 1992). Domain knowledge could guide users to effectively test the validity of their discovered knowledge (Owrang and Grupe, 1996). Domain knowledge could also improve the performance of information seeking (Marchionini, 1993; Tabatabai and Shore, 2005; Vibert *et al.*, 2007). Users with excellent domain knowledge may have greater accessibility to desired information, more flexibility to handle relevant information, and better contribution to knowledge representation (Rouet *et al.*, 1997; Vibert *et al.*, 2009). The processing of extensive knowledge of information sources in their disciplines aids

domain experts in the evaluation of both the usefulness and trustworthiness of documents (Vibert *et al.*, 2009).

Differences in knowledge and experience across domains also create barriers to a consensus in work activities or processes in an interdisciplinary collaborative work environment (Wooley and Lin, 2005). Paradigms in a particular domain can be referred to as concrete problem solutions, procedures of experiments, and theoretical models shared by the scientists in a community (Kuhn, 1974; Eysenck, 1991). However, counting on paradigms to formalize scientific thinking might possibly limit the development and evolution of a discipline (Watt, 2000). In addition, prior experience or knowledge is not always a positive (Fisher *et al.*, 2003). For example, experience or work knowledge might affect users' perceptions and expectations for data quality (Klein *et al.*, 1997), and may cut off the decision process unacceptably early (Dukerich and Nichols, 1991). Sometimes, users with sufficient knowledge might show less attention to related information (e.g., data quality information) than those who do not have such knowledge (Yates *et al.*, 1991). They might also be more inclined to perform tasks less accurately than users without prior experience (Gilliland *et al.*, 1994). Genome data curation is performed by scientists with different domain knowledge and skills. Domain knowledge differences in genomics scientists could influence the beliefs and expectations of data quality assurance activities for genome-curation specific annotation tasks and activities.

Research Questions

This was an exploratory study. It sought to understand the relationship between perception of DQ dimensions and skills and domain knowledge among genomics scientists. Specifically, the study investigated the following two research questions:

RQ1: *How do genomics scientists with different domain knowledge of genomic curation processes prioritize DQ dimensions?* This question is explored through comparing survey rankings of DQ dimensions among biologists, bioinformaticians, and computational scientists in genome curation.

RQ2: *How do genomics scientists with different domain knowledge of genomic curation processes prioritize DQ skills?* This question is investigated by comparing the survey rankings of DQ skills among biologists, bioinformaticians, and computational scientists in genome curation.

Methods

The study collected and analyzed survey data. The survey questions were collected and modified from the previous DQ dimensions and skills items found in the literature (Chung *et al.*, 2002; Wang and Strong, 1996; Lee *et al.*, 2006). Survey participants were genomics scientists who had published journal articles related to genome annotation, curation, and genomic research. Participants were given two scenarios that represented and conceptualized genome curation activities. These scenarios were developed by using scenario-based task analysis (Carroll, 1997;

Diaper, 2004; Go and Caroll, 2004; Huang *et al.*, 2012). Participants were provided the same set of written requirements for genome curation that can be used for understanding user perception (see appendix 1). Scientists thus can perceive the data quality requirements provided by a common set of curation tasks as scenarios. The first scenario asked scientists to pick *the top five* DQ Dimensions, from a total 17 DQ dimensions; the second scenario asked for the ranking of *the top five* DQ Skills, from a total of 17 DQ skills (Table 2 and Table 3). In addition, the subjects were asked to open-ended comments on the clarity and comprehensibility of the survey questions, as well as additional concerns about data quality or skills in genome curation. The 149 survey respondents were further grouped by their domain knowledge, specifically biology, computational science, and bioinformatics. Scientists who selected trainings in both Biology and Computer Science related disciplines were grouped as "Bioinformatics" (n =38). Bioinformaticians have knowledge proficiency in both biology and computer science domains. Additionally, those who chose computer science and related disciplines were grouped as "Computational Science" (n =24). Last, scientists with biology training and wet-lab experience were grouped as "Biology" (n=87). For curation experience, the majority (90% of the participants) of the scientists in this study had one year or more work experience in genome curation, and 40% had more than 5 years' experience. With regard to age, 88% of the participants were between 30~40 years old.

Table 2. List of data quality dimensions for top-five rankings and their categories.

Groups	***Data-quality dimensions***
Accuracy	*Accuracy:* Sequence records are correct and free of error
	Unbiased: Sequence records are unbiased and objective
	Believability: Sequence records are regarded as credible and believable
Accessibility	*Accessibility:* Sequence records are easily and quickly retrievable for access
	Traceability: The derivation history of the sequence records is documented and traceable
	Appropriate amount of information: The volume of the sequence records is appropriate for this scenario
Usefulness	*Interpretability:* Sequence records are in appropriate languages, symbols, and units, and the definitions are clear for interpretation
	Understandability: Sequence records are easily understandable
	Ease of manipulation: Sequence records are easy to manipulate and make it easy to carry out various tasks described in this scenario
	Consistency: Sequence records are presented in a consistent format
	Value-added: Sequence records contain additional annotations from the tasks in this scenario and these annotations are beneficial and add value
Relevancy	*Relevancy:* Sequence records contain information relevant to the scenario
	Concise representation: Sequence records are concisely represented

	Completeness: Annotated sequence records are not missing and are fully annotated according to the steps described in this scenario. *Up-to-date:* Sequence records are sufficiently up-to-date for this scenario *Reputation:* Sequence records are highly regarded and reputable in terms of their source or content

* Lists of data-quality dimensions and their groupings based on previously reported data quality dimensions and skills models (Chung *et al.*, 2002; Wang and Strong, 1996; Huang *et al.*, 2012)

Table 3. List of data-quality skills for top-five rankings and their categories.

Groups	***Data-quality skills***
Adaptive skills	*User requirement:* Ability to translate subjective user requirements for data quality into objective technical specification (such as use of Quality Function Deployment) *Data entry improvement:* Skills and ability to analyze and improve the data entry process in order to maintain data quality *Organization policies:* Ability to establish and maintain organizational policies and rules for data quality management *Change process:* Ability to manage the change process/transitions resulting from the data quality management project *Data quality cost/benefit:* Skills and ability to conduct cost/benefit analysis of data quality management *Information overload:* Understanding the information overload that managers often face and ability to reduce information overload
Interpretative skills	*Data error detection:* Ability to detect and correct errors in databases *Software tools:* Experience and ability to use diverse commercially available data quality software packages
DQ literacy skills	*Data quality dimensions:* Quality dimensions are concepts/"virtues" that define data quality. Data quality dimension skills are the ability to define and describe diverse dimensions of data quality (such as relevancy, believability, accessibility, ease of understanding) *Data quality measurement:* Data quality measurement is an operationalization of a data quality dimension. Data quality measurement skills are the ability of assessing the variation along the dimension. *Data quality implication:* Understanding pervasiveness of data quality problems and their potential impacts
Technical skills	*Data quality audit:* Ability to conduct data quality auditing (formal review, examination, and verification of data quality) *Statistical techniques:* Ability to apply statistical techniques to manage and control data quality *Data mining skills:* Data mining and knowledge discovery skills for analyzing data in a data warehouse *Data warehouse setup:* Ability to integrate multiple databases into an integrated data warehouse

	Analytic models: Ability to apply diverse analytic models (such as regression model and multidimensional model) for data analysis
	Structural Query Language (SQL): Skills and ability to apply SQL to estimate the accuracy of data

* Lists of data-quality skills and their groupings based on previously reported data quality dimensions and skills models (Chung *et al.*, 2002; Wang and Strong, 1996; Huang *et al.*, 2012)

Distribution and collection of the survey was conducted online through the Qualtrics software (http://www.qualtrics.com). The survey data was analyzed with STATA 11 software (College Station, Texas, USA) to perform descriptive statistics and Chi-Square analysis. The author computed the percentiles for the occurrences for each of the 17 DQ dimensions and DQ skills being ranked by the users as *the top five* DQ dimensions or skills. Next, the computed percentiles of each DQ dimension and skill were ranked from the largest to the smallest, and then the cumulative percentage for each DQ dimension and skill were also calculated (see Appendix 2 and 3). The cumulative percentage for each DQ dimension or skill was calculated as follows:

$$Yi = \sum_{i=1}^{k} Xi \Big/ \sum_{j=1}^{N} Xj$$

X represents the percentile value for a DQ dimension or DQ skill for the number of top-five ranking occurrences divided by the total top five ranking occurrences. *Yi* is defined as the cumulative percentage for the *ith* ranking of DQ dimensions or skills accumulated from the percentiles from the first DQ dimension or skill ranking to the *ith*. *i* takes values from 1 to *k*. *k* is the number of DQ dimensions or skills accumulated from the *1st* to *ith* rank. The value of *jth* is from 1 to *N*. *N* is the total number of DQ dimensions (*N* = 17) or skills (*N* = 17). For that reason, the value of *Y* for the last accumulated ranking (*17th*) for DQ dimensions or skills is 100% (see appendix 1 and 2). Only those DQ dimensions or skills with cumulative rankings less than 90% were kept as those greater than 90% only count for a trivial portion— specifically, less than 10% of total occurrences in top five ranking chosen by the users— and can be ignored.

The selected DQ dimensions and skills within the top 90% accumulated ranking lists were further grouped into categories based on previous reported data quality dimensions and skills models (Chung *et al.*, 2002; Wang and Strong, 1996; Huang *et al.*, 2012) as represented in Table 2 and 3. Finally, the aggregated percentage was computed for each category, for both DQ dimensions and skills models, by adding up the percentile of each dimension or skill in a category. The researchers then ranked these categories in decreasing order based on their aggregated percentages (Fig 2-3).

Findings

Chi-Square analysis of the genome curation survey results for the top-five DQ dimensions and skills selections and rankings found differences in priorities of specific DQ skills and dimensions. Some of these differences were statistically significant. Those DQ dimensions and skills that were affected by domain knowledge were identified. Furthermore, there are specific DQ trade-offs for a typical group of DQ dimensions and skills found in different user groups, particularly among computational scientists and biologists. DQ trade-offs occurred when the DQ expectations of scientists did not match the actual needs in the domain. Data curation models or policies can in fact be defined more specifically to meet the domain dependent needs, suggesting that new curation procedures and data standards need to be developed in order to accommodate different requirements among users.

The descriptive statistical analysis of the survey data for the occurrences of each DQ dimension revealed the top-five most important DQ dimensions for each group of scientists, ranked from highest to lowest. Table 4 and 5 showed the descriptive summary of the rankings for all the DQ dimensions and skills in different domain experts. Particularly, the statistical significant ones and their Chi-square values were bold/italic, and cells of the top five rankings were also highlighted for each group. The five most important dimensions for computational scientists were: Accuracy, Accessibility, Completeness, Understandability, and Appropriate amount of information. According to biologists, the five most important DQ dimensions were: Accuracy, Accessibility, Completeness, Believability, and Up-to-date. Last, bioinformaticians ranked the top-five DQ dimensions as: Accuracy, Accessibility, Completeness, Believability, and Interpretability. It is worth noting that for all three groups Accuracy, Accessibility and Completeness were among the most important DQ dimensions. Interestingly, computational scientists did not rank Believability as one of the top five, but both biologists and bioinformaticians did. In contrast, computational scientists ranked Understandability as of particular importance. Biologists were interested in Believability and Currency ("Up-to-date") and bioinformaticians cared more about Interpretability.

Table 4. Rankings of DQ dimensions based on the domain knowledge

Attribute	Computation (n=24)		Biology (n=87)		Bioinformatics (n=38)		χ^2	χ^2	χ^2
	Mean rank	Ranked by	Mean rank	Ranked by	Mean rank	Ranked by	(Comp vs Bioinf)	(Comp vs Biol)	(Biol vs Bioinf)
Accessibility	1.7	17(70.8%)	2.6	58(66.7%)	1.9	22(57.9%)	1.055	0.149	0.883
Accuracy	2.5	17(70.8%)	1.9	64(73.6%)	1.7	28(73.7%)	0.060	0.071	0.0002
Appropriate amount of	2.0	7(29.2%)	2.3	26(29.9%)	3.3	10(26.3%)	0.060	0.005	0.164

information									
Believability	2.8	7(29.2%)	2.9	34(39.1%)	2.5	17(44.7%)	1.503	0.794	0.350
Completeness	4.0	10(41.7%)	3.0	43(49.4%)	2.9	20(52.6%)	0.708	0.454	0.109
Concise representation	3.7	5(20.8%)	4.7	4(4.6%)	4.3	3(7.9%)	2.191	***6.655***	0.544
Consistent representation	2.5	7(29.2%)	2.8	30(34.5%)	3.7	14(36.8%)	0.387	0.239	0.065
Ease of manipulation	2.7	5(20.8%)	3.3	23(26.4%)	4.0	9(23.7%)	0.068	0.313	0.105
Interpretability	4.3	7(29.2%)	3.5	14(16.1%)	4.1	16(42.1%)	1.055	2.096	***9.813***
Relevance	4.0	3(12.5%)	3.8	7(8.0%)	3.0	4(10.5%)	0.057	0.455	0.203
Reputation	2.0	2(8.3%)	3.3	6(6.9%)	3.5	4(10.5%)	0.081	0.058	0.473
Security	0.0	0(0.0%)	3.0	7(8.0%)	4.4	2(5.3%)	1.305	2.061	0.307
Traceability	2.0	2(8.0%)	3.8	24(27.6%)	4.1	8(21.1%)	1.759	***3.887***	0.593
Unbiased	3.3	7(29.2%)	3.1	21(24.1%)	3.3	10(26.3%)	0.060	0.252	0.067
Understandability	3.8	10(41.7%)	4.1	13(14.9%)	3.8	7(18.4%)	***3.994***	***8.178***	0.238
Up-to-date	3.8	7(29.2%)	4.1	33(37.9%)	4.0	9(23.7%)	0.231	0.627	2.406
Value added	5.0	2(8.3%)	4.0	7(8.0%)	4.0	5(13.2%)	0.342	0.0021	0.796

Note. Bold/Italics: Chi-Square scores were statistically significant (p<0.05). Top five DQ skills for each group have the cell highlighted.

Chi-Square analysis (Table 4) found several significant differences in data quality perceptions among scientists with different domain knowledge. Compared to biologists and bioinformaticians, computational scientists held a higher expectation in Understandability and a stronger need for Concise representation. Bioinformaticians expressed a particular interest in Interpretability. Unlike computational scientists, both biologists and bioinformaticians ranked Believability as one of the five most important dimensions. Biologists also ranked Traceability higher than other two groups (Table 4).

As for DQ skills (Table 5), all three user groups shared the belief that Data error detection, Data mining skills, DQ quality measurement, and Statistical techniques were very important DQ skills for genome curation work. Biologists have a stronger need for two DQ literacy skills: DQ measurement, DQ implication. Bioinformaticians care about DQ literacy skills specifically DQ measurement, and DQ dimensions (Table 5). While there are some shared preferences between groups, the results also indicated as well, that the ranking of skills varied. Computational scientists ranked from highest to lowest, what they felt to be the most important DQ skills as Data-error detection, DQ measurement, Statistical techniques, Data mining skills, and DQ implication. Among biologists, the top five DQ skills were ranked from highest to lowest as Data-error detection, DQ measurement, Data mining skills, Statistical techniques, and DQ implication. And bioinformaticians ranked the top five most important to least as DQ error detection, Data mining skills, Statistical techniques, DQ dimensions, and DQ measurement.

Among these groups, Data quality error detection was found to be the most important skill when performing annotation work within the genome annotation context. When looking at

the ranking patterns among biologists, computational scientists, and bioinformatics, the importance rankings include data quality literacy skills as well as Interpretative skills. Interestingly, importance rankings as demonstrated in Table 5, indicate a strong demand by computational scientists for Statistics techniques.

Chi-Square analysis results also suggest that there is a stronger preference for Data warehouse setup and Information overloading skills for biologists than computational scientistsand computational biologists care more about Structure Query language (SQL) than the other two groups. It is worth noting however, that bioinformaticians, as indicated in Table 5, have higher expectations regarding Data mining skills than do biologists.

Table 5. Rankings of DQ skills based on the domain knowledge

Attribute	Computation (n=24)		Bioinformatics (n=38)		Biology (n=87)		χ^2	χ^2	χ^2
	Mean rank	Ranked by	Mean rank	Ranked by	Mean rank	Ranked by	(Comp vs Bioinf)	(Comp vs Biol)	(Bioinf vs Biol)
Analytic models	2.5	3(12.5%)	3.7	6(15.8%)	4.2	19(21.8%)	0.128	1.032	0.605
Change process	3.0	2(8.3%)	4.5	2(5.3%)	4.1	11(12.6%)	0.230	0.338	1.546
Data mining skills	3.3	10(41.7%)	3.3	23(60.5%)	2.5	34(39.1%)	2.102	0.053	***4.904***
Data-entry improvement	3.5	7(29.2%)	3.4	7(18.4%)	3.0	27(31%)	0.972	0.031	2.125
Data-error detection	1.7	15(62.5%)	2.8	30(78.9%)	2.6	58(66.7%)	1.999	0.145	1.914
Data-quality audit	3.4	9(37.5%)	3.3	7(18.4%)	3.6	24(27.6%)	2.797	0.885	1.191
Data-quality cost/benefit	0.0	0(0.0%)	3.7	3(7.9%)	4.3	6(6.9%)	2.000	1.750	0.039
Data-quality dimensions	2.0	7(29.2%)	1.7	15(39.5%)	2.0	29(33.3%)	0.683	0.149	0.437
Data-quality implication	2.6	9(37.5%)	2.4	8(21.1%)	2.7	30(34.5%)	1.999	0.075	2.255
Data-quality measurement	3.2	10(41.7%)	2.4	14(36.8%)	1.9	34(39.1%)	0.144	0.053	0.056
Data-ware house set-up	0.0	0(0.0%)	3.7	9(23.7%)	3.4	16(18.4%)	***6.649***	***5.157***	0.463
Information overload	1.7	6(25%)	2.5	2(5.3%)	4.6	7(8.0%)	***5.009***	***5.229***	0.307
Organization policies	3.7	5(20.8%)	2.7	10(26.3%)	3.2	16(18.4%)	0.241	0.073	1.008
Software tools	4.6	9(37.5%)	3.7	12(31.6%)	3.5	23(26.4%)	0.230	1.122	0.347
Statistical techniques	3.3	10(41.7%)	3.3	17(44.7%)	3.7	30(34.5%)	0.056	0.421	1.185
Structure query language	1.0	2(8.3%)	4.0	1(2.6%)	0.0	0(0.0%)	1.039	***7.383***	2.308
User requirement	4.0	7(29.2%)	2.8	5(13.2%)	3.4	23(26.4%)	2.415	0.071	2.683

Note. Bold/Italics: Chi-Square scores were statistically significant (p<0.05). Top five DQ skills for each group have the cell highlighted.

In regard to domain knowledge-based differences evidenced in the rankings of four DQ dimension categories (Figure 2), all three user groups regarded the Accuracy group as the

primary DQ concerns in genome curation work. Rankings also indicated that biologists care more about the data accessibility issues than the other two groups. Both computational scientists and bioinformaticians care more about usefulness of current curation than the biologists.

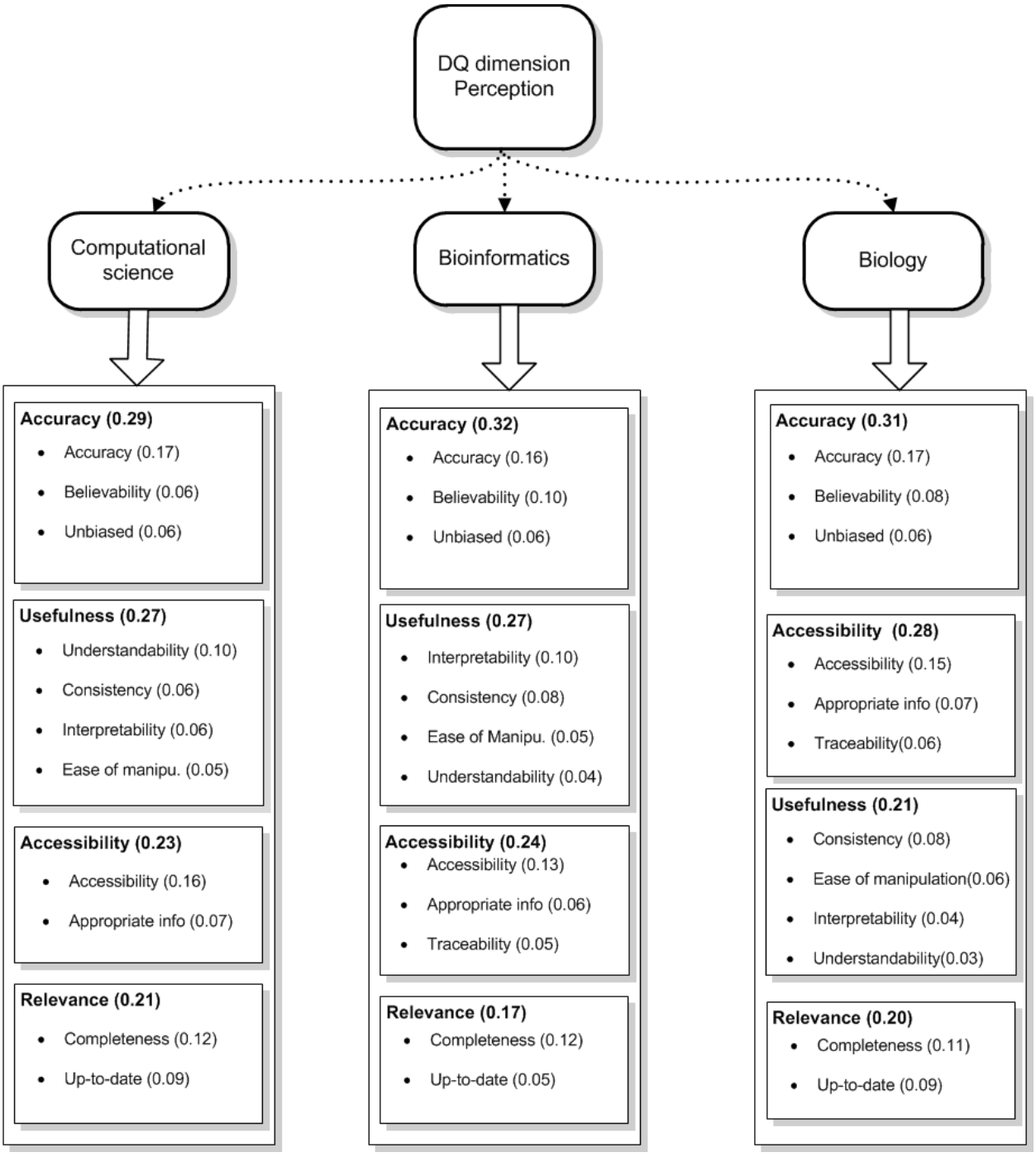

Figure 2. Domain knowledge based DQ dimension priorities. Only those with cumulative rankings of less than 90% were kept (see Appendix 1).

The study also compared the rankings of the four DQ skills categories by participants with different domain knowledge based on previous data skills models (Table 3). The results are shown in Figure 3. The findings indicated that computational scientists consider Adaptive skills more important for genome curation work than did the other two user groups. All three user groups however, regarded Technical skills as important for dealing with genome curation.

This is a preprint of an article accepted for publication in Journal of Documentation. Huang, H. (in press, 2014). Domain knowledge and data quality perceptions in genome curation work. Journal of Documentation.

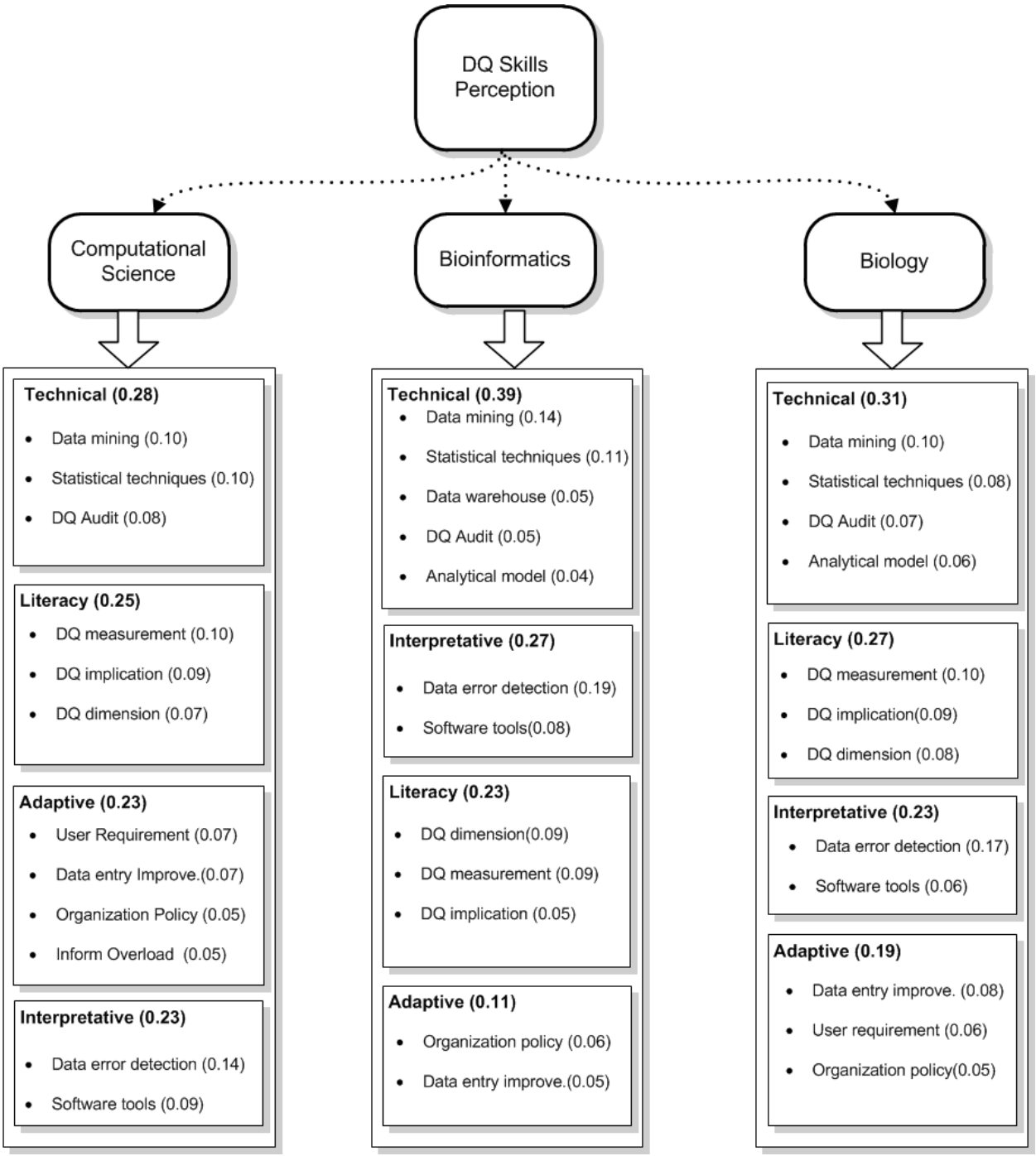

Figure 3. Domain knowledge based DQ skills grouping priorities. Only those with cumulative rankings of less than 90% were kept (see Appendix 2).

Discussion

This study determined that scientists with different domain knowledge prioritize DQ dimensions and/or skills differently. The first research question focused on DQ dimension perception gaps among users with different domain knowledge. Significant perception differences were found among all three groups in the categories of Relevant Information (Concise representation) and Useful Information (Interpretability, Understandability) (Figure 1). Believability was indicated among the top-five DQ dimensions for both biologists and bioinformaticians, but not for computational scientists. Users with different domain knowledge also assigned different priorities among the DQ skills requirements. These differences are observed among the technical skills; specifically Data mining skills, Data warehouse set-up, and SQL. Because users with different domain knowledge held specific sets of prioritized DQ dimensions and skills requirements, the contextualized data quality models were defined based on the domain knowledge of the users.

Domain Knowledge and DQ Dimensions Perception

Knowing-what knowledge required scientists to define the genome related biological research questions/curation goals. This requires theoretical biological knowledge to understand what to do with genomic data. *Knowing-how*, formally known as *procedure knowledge*, refers to the ability to carry out a task through sequential procedures, such as running the sequential tasks for genome sequences analysis. Procedural knowledge required scientists to develop automatic annotation tools and procedures for the support of genome curation work. Having obtained this knowledge, scientists may then focus on the development of a practical solution to the curation problem related to genome curation work. *Knowing why* knowledge is defined as the understanding of the reasons and principles underlying the work practice (Lee and Strong, 2003). In genome curation work, scientists who hold knowledge of both biology and computer science are at a greater advantage of understanding the purpose of the curation work in both genome curation activities and procedures, a reason why genomics scientists require cross-domain knowledge/skills both in biology and computational science for curation work. Biologists are trained by means of wet-lab experiments, making an extra effort to understand the context of data derived by an unfamiliar technique. Learning new lab techniques takes many years to master, making high-quality data appraisal difficult. Biologists are most likely interested in interpretation of the curation data through their knowledge of biology, in the help of computational scientists or programmers in data interpretation, and in explanation of the curation requirements for software development. Biologists examine authoritative data sources and evaluate their annotation. They therefore care more about the Believability of the data sources. Biologists also pay attention to the exceptions/special cases of biological knowledge (Wooley and Lin, 2005). It is important to develop genome curation systems that enable the trace function to fully capture the update of curated biological knowledge for data access and preservation (Shimoyama *et al.*, 2009). Computational scientists or programmers however, obtain training in

both the procedures and data mining protocols of the genome curation process and related issues in data management. They focus more on developing a technical and practical solution to a biological problem in data curation.

Depending on the curation problems and selected approaches, scientists might experience a mismatch of their understanding in a single aspect of data quality for their curation needs. For instance, accessibility barriers might be perceived differently among users. Some scientists might think certain genomic data simply physically unavailable rather than inaccessible. However, other scientists might interpret the barriers as being technical, based on the following reasons: 1) the coded data may be barely interpreted; 2) data may be represented in different formats which are unrecognized; 3) a large volume of data is in fact hard to locate (Strong *et al.*, 1997). Arguably, biologists analyze poor quality data every day, which may make a plausible argument for allowing 100% access to all data, even the poor data, because this is important to the domain. 'Big data' models may tolerate lower data quality in favor of massive increases in data quantity.

Similarly, there is a distinction between being mutually understood and logically sound data interpretation. Bioinformatics scientists from the domains of biology and computer science, care more about the interpretation of the curation data to the extent that data is recorded in appropriate languages, symbols, units, and the degree to which definitions and classifications are clear. Data and information can be mutually understandable within a user group, but may not be interpretable outside that group because of unfamiliarity of specific language, scientific symbols, and data formatting structure. Genomics scientists with knowledge of bioinformatics ranked Interpretability significantly higher than did the two other groups. It is also presumed that scientists occupying both biology and computer science domains, do in fact command sufficient knowledge of both fields to "assess the integrity of the data and to grasp their meaning" (Borgam, 2012, p1072). Computational scientists were found to care more about Understandability of the curation records than other two groups. They might focus on offering help to design user analysis tools for better use/reuse of curated data. Computational scientists usually require more insightful biological knowledge, background readings and reference materials to ensure their data curation that makes good biological sense and is understandable, both in intermediate and final curated records/outputs.

The curated data sources could be lab reports, field notes, archival records and other information objects. Genomics scientists have to use various sources of information to digitize and integrate the disparate pieces of genomic data. The represented curation should be concise and well-organized, as "one-point access" of a richly curated repository (Chilana *et al.*, 2008, p. 76). Computational scientists ranked Concise representation highly, and believed the improvement of the genome curation and its data representation in a concise and coherent fashion could improve understandability of data, and therefore reduce the burden of the flood of information being processed. Scientists with domain knowledge of computer science or bioinformatics could aid the development of data formats and metadata standards to support both

external data linkage, and heterogeneous data referencing. Survey data suggested that the usefulness of curated data could improve the support of user-friendly browsing, retrieving and data manipulation in an online collaborative environment. Similar findings were also observed in the following open-ended survey comments:

> "Moving between concise and detailed representations may be helpful."
> "Having Graphics [is] nice [to browse]."
> "These accessions should have been linked to the page."
> "Well described but data not structured; therefore it will be difficult to parse in automatic ways."

In addition, currency ("Up-to-date") was ranked highly by biologists. It might be concluded that curated genomic data should be frequently updated and reassessed because of the rapid changing nature of biological knowledge (Huang *et al.*, 2012). Curated information should be the most current information, with interoperation from different database platforms, as the examples below from the survey comments show:

> "Cross-compatibility with other public database, and the up-to-date relevant linkage to external databases […]"
> "This curation record has the most recent detail as both protein and CDS sequences are available with accessible hyperlinks."

Domain Knowledge and DQ Skills Perceptions

Curating genomic data requires highly-developed interdisciplinary skills, including a capacity for critical thinking and problem solving, and for cross-disciplinary thinking. Most of the scientists in this study are scholars with PhDs (81 %), obtaining educational training or research experience in either biology or other related fields. It also requires skills in information, communication and technology. Biological experts have a high level of proficiency in domain knowledge—biology. They are more confident in making judgments, evaluation or comments for curation program outputs (Chilana *et al.*, 2009). They are good at interpreting curation results, but need to consult computational scientists or programmers to obtain complicated programming tools for data mining, the switching between different database platforms, and the locating of relevant curation resources. Computational scientists offer technical support and translate curation problems into actionable programming tools. They need to work closely with biologists to ensure that their curation program outputs are consistent with the original biological problem (Chilana *et al.*, 2009). The survey data suggests a trend in genome curation work for the engagement of more scientists with both computer science and biology domain knowledge. Bioinformaticians with knowledge of both domains have advantages integrating biological knowledge into applicable solutions of curation. Computational scientists regard the use of curation of data mining and database tools (e.g., Data warehousing, SQL) as important, since

their jobs involve data wrangling, integration and retrieval in large-scale databases (Heer and Kandel, 2012).

All three user groups value highly the DQ technical skills. This finding indicated that curating genomic data requires a great number of data mining and statistical analysis tools to support data curation related tasks. The DQ Adaptive skills mattered more to computational scientists than the other two groups (Figure 2). Adaptive skills are those that allow computational scientists or programmers to actively interact with other users, which helps facilitate understanding of users' requirements, and translate the curation problems into practical solutions.

Data quality literacy reflects the ability of users to understand data quality related concepts, knowledge and skills. Particularly, data quality literacy skills such as DQ dimensions and DQ implication were ranked highly among biologists. This finding suggests that grasping the necessary knowledge of data quality concepts, the related assessment methods, and their ways to identify potential data quality problems are prerequisites for scientists to secure high quality curation work. Computational scientists and biologists might have sufficient skills when operating with their own domains, but they might be also interested in the cross-disciplinary skills required for scientific data management and data quality assurance. Such skills, as well as other annotation and data mining skills, could facilitate the curation activities, data quality assurance, and data provenance services in genome curation work.

The trends regarding quality assurance and knowledge creation activities (e.g., data annotation) ultimately evolves into higher expectations for bioinformatics literacy, including data quality literacy on the part of users. According to a recent survey (Burge *et al.*, 2012), the biocuration community believes that a genome curator, having both research experience and a strong biological or computational background, would benefit their work tremendously. The differences among users' perceptions need to be benchmarked, collected, and communicated. The empirically-based community feedback is needed to design appropriate strategies for improvement in curation quality.

The findings of this research can help develop curation domain specific data quality models. Computational scientists ranked Usefulness higher than Accessibility, whereas biologists did the opposite (Figure 1). This may also suggest the presence of trade-offs among different data quality dimensions is related to the users' domain knowledge. Literature suggests that data quality activities are not free, it requires the user's priorities, including participation in a possible trade-offs based on the different dimensions of quality optimization (Ballou and Pazer, 1995; Stvilia et al, 2007). The identified data quality trade-offs are reasonable. This is provided that the overall data quality is of sufficient and good enough for its use in the research practice. For example, Accessibility is expensive when dealing with a high volume of the data. These data quality trade-offs can be justified when organized and curated data is difficult to obtain and access with given time restrictions. Biologists required sufficient computational skills or knowledge to access and retrieve the data they want, but they may have to accept and tolerate

raw and unstructured curated data in exchange for having timely access to important information. Similarly, computational scientists possess advanced skills in genome related data wrangling (Heer and Kendal, 2012), they focus on making their collected data more usable by adding more curated information.

Genomics scientists, like scholars from other scientific disciplines, require sufficient data curation and process skills to conduct tremendous data manipulation work. This study collected empirical data through a survey of members in a particular scientific community. It reports members' perceived priorities for data quality criteria and identified related DQ skills in the context of genomic data curation work (see Figures 2, 3). The findings of this study can be used in the development of genomic data curation procedures, policies and training modules. These curation artifacts could be used by the current curation team and by future institutional end-users and participants, who may themselves not possess extensive trainings in data curation and data management.

Conclusion

The way scientists solve problems in genome curation today is probably not the way scholars and practitioners did so a decade ago. Since technology is growing, our knowledge and abilities are also increasing, and our analytical methods are changing as well. Genome curation work is a collaborative process executed through a dynamic complex interaction among those scientists who hold diverse domain knowledge and work experience. It requires scientists to read tremendous amounts of research literature, and to obtain solid domain knowledge. It also requires scientists to be flexible and adaptive to deal with different scales of genomic related data, to make sound judgments regarding the annotated information in the genome context, and to ensure the capture of all related information within the data model.

Scientists' domain knowledge and experience in genome curation work eventually impacts their priorities for the data quality criteria. Overall, scientists must process enormous amounts of distributed data through many different tools developed to aid them in knowledge discovery. This work will allow for richer knowledge representation and manipulation. This study also has some limitations. The data was collected by survey, rather than direct observation to collect the opinions of the scientists regarding data quality skills and dimensions requirements used to develop the data quality models for genome curation. The data are therefore only approximations of the respondents' actual value models for quality and for data quality skills used in practice. Future research collection of additional empirical data through observations and interviews can help determine the community's data curation and quality assurance practices. What is more, the importance of these concepts was recorded by survey participants at the time of survey completion; the follow-up interview provides an opportunity to validate where modifications are necessary. It should also be pointed out that the data quality skills used as the survey instrument were in fact based on previous studies (Chung *et al.*, 2002). As new data

management technologies evolve (e.g., computing with Graphics Processing Units (GPUs) and "cloud" technologies), these items and related constructs may require a revisit to update the priorities of the community regarding data quality assurances skills.

Genomics research is data-intensive. Some significant differences were observed in scientists' perception of data quality requirements in genome curation work which required calibration of their knowledge across different domains. This study found that given a common curation task with the same data-quality information, genomics researchers with diverse domain knowledge make different decisions regarding data-quality trade-offs. Through this study, the identification of the variations of the DQ models based on domain knowledge can help better understand the function of data quality in context of domain knowledge. It can also help identify related curation tools and supports for the genomics research community, and to develop curation policies, procedures, training modules and strategies, and problem-solving paths tailored to the curation work. Future studies could involve the collection of additional data and the development of operational models of these trade-offs, allowing them being used in practice to optimize quality assurance activities.

References

Allen, B. (1991), "Topic knowledge and online catalog search formulation", *Library Quarterly*, Vol. 61 No.2, pp.188–213.

Bartlett, J.C. and Toms, E.G. (2005), "Developing a protocol for bioinformatics analysis: An integrated information behavior and task analysis approach", *Journal of the American Society for Information Science and Technology*, Vol. 56 No.5, pp. 469–482.

Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I.,and Ostell, J. (2012), "BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata", *Nucleic acids research*, Vol. 40 No.D1, pp. D57-D63.

Bentley, D.R. (2006), "Whole-genome re-sequencing", *Current Opinion in Genetics & Development*, Vol. 16, pp. 545–552.

Brent, M.R. (2008), "Steady progress and recent breakthroughs in the accuracy of automated genome annotation", *Nature Reviews Genetics*, Vol. 9, pp. 62–73.

Brown, C.M. (2003), "The changing face of scientific discourse: Analysis of genomic and proteomic database usage and acceptance", *Journal of the American Society for Information Science & Technology*, Vol. 54 No. 10, pp. 926-938.

Burge, S., Attwood, T.K., Bateman, A., Bateman, A., Berardini, T.Z., Cherry, M., and Gaudet, P. (2012), "Biocurators and biocuration: surveying the 21st century challenges", *Database*, bar059.

Burkhardt, K., Schneider, B. and Ory, J. (2006), "A biocurator perspective: annotation at the Research Collaboratory for Structural Bioinformatics Protein Data Bank", *PLoS Computational Biology*, Vol. 2, e99.

Carroll, J.M. (1997), "Scenario-based design", In Helander, M. and Landauer T.K. (Eds.), *Handbook of human–computer interaction,* Amsterdam, North Holland, pp. 383–406.

Chilana, P.K., Palmer, C.L. and Ko, A.J. (2009), "Comparing bioinformatics software development by computer scientists and biologists: An exploratory study", *Software Engineering for Computational Science and Engineering, SECSE '09. ICSE Workshop*, Vol.72 No.79, pp. 23–24.

Chung, W., Fisher, C. and Wang R. (2002), "What skills matter in data quality?", In *the 7th International Conference on Information Quality (ICIQ-02),* Boston, MA.

Cole, N. J. and Bawden, D. (1996), "Bioinformatics in the pharmaceutical industry", *Journal of Documentation*, Vol. 52 No. 1, pp. 51-68.

Diaper, D. (2004), "Understanding task analysis in human computer interaction", In Diaper D. and Stanton N. (Eds.), *The handbook of task analysis for human–computer interaction,* Erlbaum, Mahwah, NJ, pp. 117–133.

Dukerich, J. M. and Nichols, M. L. (1991), "Causal information search in managerial decision making", *Organizational Behavior and Human Decision Processes*, Vol. 50 No.1, pp. 106-122.

Evans, J. R. and Lindsay, W.M. (2005), *The management and control of quality*. Thomson Learning, Cincinnati, OH, pp. 132-136.

Eysenck, H. J. (1991). "Dimensions of personality: 16, 5 or 3?—Criteria for a taxonomic paradigm", *Personality and individual differences*, Vol.12 No. 8, pp. 773–790.Fenstermacher, D. (2005), "Introduction to bioinformatics", *Journal of the American Society for Information Science and Technology*, Vol. 56 No.5, pp. 440–446.

Fisher, C. W., Chengalur-Smith, I. and Ballou, D. P. (2003), "The impact of experience and time on the use of data quality information in decision making", *Information Systems Research*, Vol. 14 No.2, pp. 170-188.

Frické, M. and Fallis, D. (2004), "Indicators of accuracy for answers to ready reference questions on the internet", *Journal of the American Society for Information Science and Technology,* Vol. 55 No.3, pp. 238–245.

Gilliland, S. W., Wood, L. and Schmitt, N. (1994), "The effects of alternative labels on decision behavior: the case of corporate site selection decisions", *Organizational behavior and human decision processes*, Vol. 58 No. 3, pp. 406-427.

Go, K. and Carroll, J. (2004), "Scenario-based task analysis", In Diaper D. and Stanton N. (Eds.), *The handbook of task analysis for human-computer interaction*, Erlbaum, Mahwah, NJ, pp. 117–133.

Goth, G. (2012), "Preserving digital data", *Communications of the ACM*, Vol. 55 No. 4, pp. 11-13.

Haythornthwaite, C. (2006), "Learning and knowledge networks in interdisciplinary collaborations", *Journal of the American Society for Information Science and Technology*, Vol. 57 No.8, pp. 1079-1092.

Heer, J. and Kandel, S. (2012), "Interactive analysis of big data", *XRDS: Crossroads, The ACM Magazine for Students*, Vol. 19 No.1, pp. 50-54.

Hemminger, B. M., Saelim, B., Sullivan, P. F. and Vision, T. J. (2007), "Comparison of full-text searching to metadata searching for genes in two biomedical literature cohorts", *Journal of the American Society for Information Science and Technology*, Vol. 58 No.14, pp. 2341-2352.

Hjørland, B. and Albrechtsen, H. (1995), "Toward a new horizon in information science: domain-analysis", *Journal of the American Society for Information Science and Technology*, Vol. 46 No.6, pp. 400-425.

Huang, H., Andrews, J. and Tang, J. (2012), "Citation characterization and impact normalization in bioinformatics journals", *Journal of the American Society of Information Science and Technology*, Vol. 63 No.3, pp. 490-497.

Huang, H., Lu, J., Hunter, W. and Liang, S. (2011), "Using IBM Content Manager for genomic data annotation and quality assurance tasks", *IBM Journal of Research and Development.* Vol. 55 No. 6, pp. 13.

Huang, H., Stvilia, B., Jörgensen, C. and Bass, H. (2012), "Prioritization of data quality dimensions and skills requirements in genome annotation work", *Journal of the American Society for Information Science and Technology*, Vol. 63 No.1, pp. 195-207.

Ioannidis, J.P. and Khoury, M.J. (2011), "Improving validation practices in 'omics' research", Science, Vol. 334 No. 6060, pp. 1230–1232.

Ju, B. (2007), "Does domain knowledge matter: Mapping users' expertise to their information interactions", *Journal of the American Society for Information Science and Technology*, Vol. 58, pp. 2007–2020.

Klein, B. D., Goodhue, D. L. and Davis, G. B. (1997), "Can humans detect errors in data? Impact of base rates, incentives, and goals", *MIS Quarterly*, Vol. 21 No. 2, pp.169-194.

Kuhn, T. S. (1974). "Second thoughts paradigms", In Suppe, F. (Ed.), *The structure of science theories*, pp. 459–482. London: University of Illinois Press.

Lathe, W., Williams, J., Mangan, M. and Karolchik, D. (2008), "Genomic data resources: challenges and promises". *Nature Education,* Vol. 13.

Lee, Y. and Strong, D. (2003), "Knowing – why about data processes and data quality", *Journal of Management Information Systems,* Vol. 20 No. 3, pp. 13–39.

Leonelli, S., Diehl, A.D., Christie, K.R., Harris, M.A. and Lomax, J. (2011). "How the gene ontology evolves", *BMC Bioinformatics*, Vol. 12, p325.Mackay, J. M. and Elam, J. J. (1992), "A comparative study of how experts and novices use a decision aid to solve problems in complex knowledge domains", *Information Systems Research*, Vol. 3 No. 2, pp.150-172.

MacMullen, W. (2006), *Contextual analysis of variation and quality in human-curated gene ontology annotations*, PhD dissertation, University of North Carolina.

MacMullen, W.J. and Denn, S.O. (2005), "Information problems in molecular biology and bioinformatics", *Journal of the American Society for Information Science and Technology*, Vol. 56 No.5, pp. 447-456.

Mao, J. Y. and Benbasat, I. (2000), "The use of explanations in knowledge-based systems: Cognitive perspectives and a process-tracing analysis", *Journal of Management Information Systems,* Vol. 17 No. 2, pp. 153-180.

Marchionini, G., Dwiggins, S., Katz, A. and Lin, X. (1993), "Informationseeking in full-text end-user-oriented search systems: The roles of domain and search expertise", *Library & Information Science Research,* Vol.15, pp. 35–69.

Mayor, C. and Robinson, L. (2013). "Ontological realism, concepts and classification in molecular biology: development and application of the gene ontology", Journal of Documentation, Vol. 70 No. 1, pp.173–193.

Owrang O, M. M., and Grupe, F. H. (1996), "Using domain knowledge to guide database knowledge discovery". *Expert Systems With Applications*, Vol. 10, No.2, pp. 173-180.

Pagani, I., Liolios, K., Jansson, J., Chen, I. M. A., Smirnova, T., Nosrat, B.,and Kyrpides, N. C. (2012), "The Genomes OnLine Database (GOLD) v. 4: status of genomic and

metagenomic projects and their associated metadata", *Nucleic Acids Research*, Vol. 40 No. D1, pp. D571-D579.

Payne, J. W., Bettman, J. R. and Johnson, E. J. (1993), *The adaptive decision maker*. University Press, Cambridge.

Palmer, C. L. and Neumann, L. J. (2002), "The information work of interdisciplinary humanities scholars: exploration and translation", *The Library Quarterly*, Vol. 72 No. 1, pp. 85-117.

Pruitt, K.D., Tatusova, T., Brown G.R., and Maglott, D.R. (2012), "NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy", *Nucleic Acids Research*, Vol. 40 No.D1, pp. D130-D135.

Reed, J. L., Famili, I., Thiele, I. and Palsson, B. O. (2006), "Towards multidimensional genome annotation", *Nature Reviews Genetics*, Vol. 7 No. 2, pp. 130-141.

Rieh, S. (2002), "Judgment of information quality and cognitive authority in the Web", *Journal of the American Society for Information Science and Technology*, Vol. 53 No. 2, pp. 145–161.

Rouet, J.-F., Favart, M., Britt, M.A. and Perfetti, C.A. (1997), "Studying and using multiple documents in history: Effects of discipline expertise", *Cognition and Instruction*, Vol. 15, pp. 85–106.

Salimi, N. and Vita, R. (2006), "The biocurator: connecting and enhancing scientific data", *PLoS Computational Biology*, Vol. 2 No. e125.

Salzberg S. (2007), "Genome re-annotation: A wiki solution?", *Genome Biology*, Vol. 8, pp. 102–102.

Samuel, V., Gussman, A., and Klumke, W. (2008), "Toward an online repository of standard operating procedures (SOPs) for (meta)genomic annotation", *OMICS: A Journal of Integrative Biology*, Vol. 12 No. 2, pp. 137–141.

Sanbonmatsu, D. M., Kardes, F. R., and Herr, P. M. (1992), "The role of prior knowledge and missing information in multiattribute evaluation", *Organizational Behavior and Human Decision Processes*, Vol. 51 No. 1, pp. 76-91.

Sanderson, K. (2011), "Bioinformatics: curation generation", Nature, Vol. 470, pp. 295–296.

Shachak, A. and Fine, S. (2008), "The effect of training on biologists acceptance of bioinformatics tools: A field experiment", *Journal of the American Society for Information Science and Technology*, Vol. 59 No. 5, pp. 719-730.

Shimoyama, M., Hayman, G. T., Laulederkind, S. J. F., Nigam, R., Lowry, T. F., … Dwinell, M. R. (2009), "The rat genome database curators: who, what, where, why", *PLoS Computational Biology*, Vol. 5 No. e1000582.

Strong, D., Lee Y. and Wang R. (1997), "Data quality in context", *Communication of the ACM*, Vol. 40 No.5, pp. 103–110.

Stvilia, B., Gasser, L., Twidale, M., and Smith L. (2007), "A framework for information quality assessment", *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 12, pp. 1720–1733.

Stvilia, B. and Gasser, L. (2008), "An activity theoretic model for information quality change", *First Monday*, Vol. 13, No. 4.

Stvilia, B., Twidale, M., Smith, L. C. and Gasser, L. (2008), "Information quality work organization in Wikipedia", *Journal of the American Society for Information Science and Technology*, Vol. 59 No. 6, pp. 983–1001.

Tabatabai, D. and Shore, B.M. (2005). "How experts and novices search the Web", *Library & Information Science Research*, Vol. 27, pp. 222–248.

Vibert, N., Rouet, J.-F., Ros, C., Ramond, M. and Deshoullières, B. (2007), "The use of online electronic information resources in scientific research: The case of neuroscience", *Library & Information Science Research*, Vol. 29, pp. 508–532.

Vibert, N., Ros, C., Bigot, L. L., Ramond, M., Gatefin, J. and Rouet, J. F. (2009), "Effects of domain knowledge on reference search with the PubMed database: An experimental study", *Journal of the American Society for Information Science and Technology*, Vol. 60 No.7, pp. 1423-1447.

Wang, R. and Strong, D. (1996), "Beyond accuracy: What data quality means to data consumers", *Journal of Management Information Systems*, Vol. 12 No. 4, pp. 5–35.

Watt, W. B. (2000). "Avoiding paradigm-based limits to knowledge of evolution", In Clegg *et al.* (Ed.), *Evolutionary Biology*, pp. 73–96. Springer US.

Willis, C., Greenberg, J., and White, H. (2012). "Analysis and synthesis of metadata goals for scientific data", *Journal of the American Society for Information Science and Technology*, Vol. 63 No.8, pp. 1505-1520.

Wildemuth, B.M. (2004), "The effects of domain knowledge on search tactic formulation", *Journal of the American Society for Information Science and Technology*, Vol. 55 No. 3, pp. 246–258.

Wooley, J. C., and Lin, H. S. (Eds.). (2005), *Catalyzing inquiry at the interface of computing and biology*. National Academies Press.

Wu, L. L., Huang, M. H. and Chen, C. Y. (2012), "Citation patterns of the pre-web and web-prevalent environments: The moderating effects of domain knowledge", *Journal of the*

American Society for Information Science and Technology, Vol. 63 No. 11, pp. 2182-2194.

Wu, J., Zhang, Y., Zhang, H., Huang, H., Folta, K. and Lu, J. (2010), "Whole genome wide expression profiles of *Vitis amurensis* grape responding to downy mildew by using Solexa sequencing technology", *BMC plant biology*, Vol. 10 No. 1, pp. 234.

Yang, X., Ye, Y., Wang, G., Huang, H., Yu, D. and Liang, S. (2011), "VeryGene: linking tissue-specific genes to diseases, drugs, and beyond for knowledge discovery", *Physiological genomics*, Vol. 43 No. 8, pp. 457-460.

Yates, J. F., McDaniel, L. S. and Brown, E. S. (1991), "Probabilistic forecasts of stock prices and earnings: The hazards of nascent expertise", *Organizational Behavior and Human Decision Processes*, Vol. 49 No. 1, pp. 60-79.

Appendix 1. Two genome curation scenarios.

Scenarios
Scenario 1: **Production, curation, and submission of Expressed Sequence Tags (ESTs) data** In this scenario, you will generate primary sequence data. For this purpose, you will process, curate, annotate, and submit sequence data as annotated sequence records in a public database. Specifically, you will produce a cDNA library, and obtain 1,000 random sequence reads (ESTs) from that cDNA library. The library contains clones from a model organism for which a genome sequence is publicly available. As part of preparing these annotated records, you will be taking steps which include annotation and data quality assurance steps to: • process the raw data to remove vector or low quality sequences, • annotate the sequences with regards to the genome location, • predict gene products using routine bioinformatic tools such as BLAST alignments, open reading frames (ORFs) predictions, and comparison of predicted proteins to protein motif databases, • produce additional annotation to link these predicted gene products to gene ontology, molecular networks, or biochemical pathways, • submit these ESTs and associated annotations to two different databases, GenBank and your species specific database. *The phrase "sequence records" refers to both the primary DNA sequences themselves and all the associated annotations.
Scenario 2: **Whole genome data curation in a model organism** In this scenario, you will generate genome annotation records for a particular model organism. You will use the full spectrum of genome annotation approaches including: predicted gene and protein

annotation, sequences comparisons and alignments, genome variations analysis, the organization and annotation of molecular networks and biochemical pathways. You will employ these approaches using specialized databases, bioinformatics software, and literature mining to:

1. Create sequence records for release to the public.
   a. Curate, annotate genome sequence data features from the sequence data by identifying the gene features (e.g., promoters, gene length, terminators) and genomic properties (e.g., motifs, repeats) from the sequence data.
   b. Create explicit comments to the sequence data organized along a schema that needs to be specified (e.g., gene name, gene function, enzyme identifier, bibliographic reference, experimentally identified feature, ESTs, etc.)
   c. Compare, correct, reannotate, or externally link the sequence data to the data available in other databases or scientific literature.

2. Conduct data quality control by corresponding with collaborators regarding missing or inaccurate information.

3. Assist in problem identification and recommend enhancements to the procedures in genome annotation work.

*These two scenarios were adopted from Huang *et al.*, (2012).

Appendix 2. Cumulated percentiles of the ranked DQ dimensions in domain knowledge

	Bioinfo			Biology			Both	
	%	Cumulated %		%	Cumulated %		%	Cumulated %
Accessibility	14.7	14.7	Accuracy	15.5	15.5	Accuracy	14.9	14.9
Accuracy	14.7	29.3	Accessibility	14	29.5	Accessibility	11.7	26.6
Completeness	8.6	37.9	Completeness	10.4	40	Completeness	10.6	37.2
Understandability	8.6	46.6	Believability	8.2	48.2	Believability	9.0	46.3
Appro amount of info	6.0	52.6	Up-to-date	8.0	56.2	Interpretability	8.5	54.8
Believability	6.0	58.6	Consistency	7.3	63.4	Consistency	7.4	62.2
Consistency	6.0	64.7	Appro amount of info	6.3	69.7	Appro amount of info	5.3	67.6
Interpretability	6.0	70.7	Ease of manipulate	5.6	75.3	Unbiased	5.3	72.9
Unbiased	6.0	76.7	Traceability	5.6	80.9	Ease of manipulate	4.8	77.7
Up-to-date	6.0	82.8	Unbiased	5.1	86	Up-to-date	4.8	82.4
Concise repres	4.3	87.1	Interpretability	3.4	89.3	Traceability	4.3	86.7

Ease of manipulate	4.3	91.4	Understandability	3.1	92.5	Understandability	3.7	90.4
Relevance*	***2.6***	***94***	***Relevance***	***1.7***	***94.2***	***Value-added***	***2.7***	***93.1***
Traceability	***2.6***	***96.6***	***Security***	***1.7***	***95.9***	***Relevance***	***2.1***	***95.2***
Reputation	***1.7***	***98.3***	***Value-added***	***1.7***	***97.6***	***Reputation***	***2.1***	***97.3***
Value-added	***1.7***	***100***	***Reputation***	***1.5***	***99***	***Concise repres***	***1.6***	***98.9***
Security	***0.0***	***100***	***Concise repres***	***1.0***	***100***	***Security***	***1.1***	***100***

*DQ dimensions are Bold/Italics indicating their accumulated rankings over 90%.

Appendix 3. Cumulated percentiles of the ranked DQ skills in domain knowledge

	Bioinfo			Biology			Both	
	%	Cumulated %		%	Cumulated %		%	Cumulated %
Data-error detection	13.6	13.6	Data-error detection	15	15	Data-error detection	17.5	17.5
DQ measurement	9.1	22.7	DQ measurement	8.8	23.8	Data-mining skills	13.5	31
Statistical techniques	9.1	31.8	Data-mining skills	8.8	32.6	Statistical techniques	9.9	40.9
Data-mining skills	9.1	40.9	Statistical techniques	7.7	40.3	DQ dimensions	8.8	49.7
DQ implication	8.2	49.1	DQ implication	7.8	48.1	DQ measurement	8.2	57.9
DQ audit	8.2	57.3	DQ dimensions	7.5	55.6	Software tools	7.0	64.9
Software tools	8.2	65.5	Data-entry improvement	6.9	62.5	Organization policies	5.9	70.8
DQ dimensions	6.4	71.8	DQ audit	6.2	68.7	Data-warehouse set-up	5.2	76
Data-entry improvement	6.4	78.2	Software tools	6.0	74.7	DQ implication	4.7	80.7
User requirement	6.4	84.6	User requirement	5.9	80.6	Data-entry improvement	4.1	84.8
Organization policies	4.5	89.1	Analytic models	4.9	85.5	DQ audit	4.1	88.9
Information overload	4.5	93.6	Organization policies	4.2	89.7	Analytic models	3.5	92.4
Analytic models*	***2.8***	***96.4***	***Data-warehouse set-up***	***4.1***	***93.8***	***User requirement***	***2.9***	***95.3***
Change process	***1.8***	***98.2***	***Change process***	***2.8***	***96.6***	***DQ cost/benefit***	***1.8***	***97.1***
SQL	***1.8***	***100***	***Information overload***	***1.9***	***98.5***	***Change process***	***1.1***	***98.2***
DQ cost/benefit	***0.0***	***100***	***DQ cost/benefit***	***1.5***	***100***	***Information overload***	***1.2***	***99.4***
Data-warehouse set-up	***0.0***	***100***	***SQL***	***0.0***	***100***	***SQL***	***0.6***	***100***

*DQ skills are Bold/Italics indicating their accumulated rankings over 90%.