This is a preprint of an article accepted for publication in *Library & Information Science Research*. Huang, H., Jörgensen, C., Stvilia, B. (in press, 2014). Genomics data roles, skills, and perception of data quality. *Library & Information Science Research*.

Genomics Data Curation Roles, Skills, and Perception of Data Quality

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

Corinne Jörgensen and Besiki Stvilia

School of Library and Information Studies, Florida State University, Tallahassee, Florida, 32306-2100.

Telephone: (850) 645-7366, (850) 644-5775; Fax: (850) 644-6253; E-mail: {bstvilia, cjorgensen}@fsu.edu

*Corresponding Author

Abstract

Compared to a decade ago, genomics scientists, driven by technical changes and availability of massive genomic data, are performing a wider plurality of curation roles

including those of end-users, curators, or dual-role users. Scientists with different curation roles (including that of end user) may focus on different data quality aspects and skills requirements in a community curation environment. This study examines how genomics scientists' perceived priorities for data quality and data quality skills differ when assuming different roles played in genomics data curation work. The analysis of survey data collected from 147 genomics scientists found that curators of genomic data valued quality criteria that can be assessed through direct examination of the data more highly, while end-users placed a high value on the quality criteria that can be assessed indirectly such as believability. With regard to data quality skills, curators appeared to care more about understanding user's requirements and specific data management skills than end-users, while end-users valued the skills needed to deal with information overload more highly – those needed to identify useful, relevant information from large amounts of data. The study found that scientists with different curation roles, given common curation tasks with the same skill requirements, prioritized different data quality criteria. The data quality, skill priorities, and tradeoffs identified by this study can inform the development of effective data curation mandates and policies, data quality assurance planning and training, and the design of curation role specific tool dashboards and visualization interfaces for genomics data.

Introduction

The widespread use of information and communication technologies has globalized genomic research, and vastly increased both the number of collaborators on projects and the size of the genomics data involved in a project (Özdemir et al., 2013; Ward, Schmieder, Highnam, & Mittelman, 2013). Genome curation involves multiple steps that integrate genomic data with disparate pieces of validated experimental evidence and literature across databases (Pruitt, Tatusova, Brown, & Maglott, 2012). Manual and automatic genomic data curation are processed using standardized terminologies and metadata schemas (Kuhn, Haussler, & Kent, 2013; MacMullen & Denn, 2005; Reed et al., 2006; Shimoyama et al., 2009). Genome curation products provide value-added information for interpreting genome structure and function (Pruitt, Tatusova, Brown, & Maglott, 2012). This type of work has resulted in a new kind of professional: The biocurator, who conducts data quality (DQ) assurance work to ensure the accuracy and precision of biological curation activities prior to its release to the public (Sanderson, 2011).

Traditionally, biological knowledge was collected and created manually by domain experts. Expert curators now have to remain updated to process the rapid growth in diverse biological data and literature (Baumgartner et al., 2007). Accordingly, community curation has become a solution that can collect community intelligence for knowledge creation when dealing with the flood of biological knowledge (Good, Clarke, de Alfaro, & Su, 2012; Mons et al., 2008; Salzberg, 2007). The online community curation environment is highly pluralistic, participatory, and social, presenting challenges to the conventional professional role and working practices of genomics scientists (Hoffmann 2008). In this environment, genomics scientists have moved from their traditional role solely dealing with in-house data to multiple roles that involve diverse and interactive ways to share their curated work to others. Scientists today need to obtain sufficient data skills (i.e., data "wrangling" skills) and other necessary curation skills – in addition to their domain knowledge – in order to make a meaningful use of genomic data. Given the increasing need for genomics scientists to incorporate multiple curation roles into their working practices, understanding the relationships that exist among these roles and scientists' perspectives regarding data quality and skills requirements in genome curation work becomes

increasingly important. Therefore, the primary focus of this exploratory study is to examine the relationships among the perceptions of data quality dimensions, data quality skills, and varying curation roles of 147 genomics scientists in two hypothetical genome curation scenarios.

Problem Statement

With both single and multiple varieties of curation roles in existence among genomic scientists, the respective complexities of data roles for genome curation might consequently affect scientists' decision making for data quality assurances. Identifying the role-based data quality and skills requirements could help scientists to have better preparation for their specific curation roles and could potentially stimulate development of data management architectures to support role-based online genome curation systems. While substantial research (Wang et al., 2012; MacMullen 2006) has explored the genome curation strategies and specific data quality issues, there is little empirical evidence regarding how data quality issues are perceived by scientists with different data roles. Huang et al. (2012) found that genomics scientists shared some of the expectations for the data quality requirements and proposed a data quality model dealing with overall genome curation activities. The proposed model however did not address issues related to curation roles. In addition, the relationship between data roles for genome curation and data quality assurance activities remains unknown. Scientists with single or multiple curation roles may make conflicting data quality decisions when assigned similar genome curation tasks, resulting in the current gap in understanding curation problems associated with data quality assurance.

The current study addresses this gap by specifically examining the respective performances of three different user groups who play different curation roles as end-users, dual-role users, and curators in order to identify their perceptions of data quality and skill requirements. The study explores user perceptions of skills or knowledge required for genomic data curation that may be lacking within the professions engaged in the work, recognizing the unique and significant curation roles in the genomics community. Librarians and practitioners, especially those involved with institutional repositories, actively seek collaborations with allied professionals (e.g., scientists) for data practice and curation activities (Friedlander & Alder, 2006). Data curators and end

users, along with practitioners, need to have appropriate expertise, skills and consolidated curation policies to carry out required curation work. The findings can suggest development of role based data quality and skill curation strategies that yield both improved resource integration and more cost-effective collaborative solutions in the context of e-science. The results provide examples that could enrich data curation curriculum in data quality and skill requirements for different stakeholders, including scientists and practitioners. The results also serve to support development of data quality and curation skill training modules that will enable scientists to smoothly transition the job skills necessary in data quality as they switch roles from end-users to curators.

The primary focus of this exploratory study was to examine the relationships among the perceptions of DQ dimensions, DQ skills, and the different curation roles of genomics scientists. In particular, the study investigated the following research questions:

RQ1: *How do genomics scientists with different genome curation roles prioritize DQ dimensions?*

This question was investigated by comparing survey rankings of DQ dimensions among curators, end-users and dual-role users working with genome curation.

RQ2: *How do genomics scientists with different genome curation roles prioritize DQ skills?*

This question was investigated by comparing the survey rankings of DQ skills among curators, end-users and dual-role users working with genome curation.

Literature Review

Data quality is a contextual and multidimensional concept that must be defined, operationalized, and evaluated within the contexts of its use (Strong, Lee, & Wang, 1997; Stvilia, Gasser, Twidale, & Smith, 2007). A Data Quality (DQ) dimension is a single aspect or component of a data quality concept (Stvilia, Gasser, Twidale, & Smith, 2007; Wang and Strong, 1996, p. 6), some of which are identified by researchers as important, or top priorities by users in a particular domain (Bade, 2007; Kahn, Strong, & Wang, 2002; Homburg, Droll, & Totzek, 2008). For example, specific sets of DQ dimensions were determined to be important for evaluating consumer health information (Frické & Fallis, 2004; Lankes, 2008; Stvilia, Mon, & Yi, 2009), online

scholarly information (Rieh, 2002), and gene-ontology curation behaviors (MacMullen, 2006). Data quality models, including a taxonomy of DQ dimensions, were developed to capture and describe the context and value structure for DQ for a Wikipedia community (Stvilia, Gasser, Twidale, & Smith, 2007) and the genomics research community (Huang, Stvilia, Jörgensen, & Bass 2012). Wang and Strong (1996) defined quality as "fitness for use," indicating the importance of determining data quality in its context of use (Strong, Lee, & Wang, 1997; Stvilia, Gasser, Twidale, & Smith, 2007). Thus, understanding "user satisfaction" or "meeting or exceeding user expectation" (Evans and Lindsay, 2005) could prove helpful in characterizing data quality in a specific context.

Lynch (2009) stated that data and its associated software/infrastructure are integral parts of the scientific record. As such, the quality of the data, software and systems, as well as required skills should therefore be considered simultaneously (Kahn et al., 2002). Information content and technical challenges for data processing were considered as interrelated issues for data quality. Similarly, a user's perception of data quality involves more than data itself (Klein, 2002). When discussing data quality for data processing, three data roles were identified – data collectors or producers, data custodians, and data consumers (Lee & Strong, 2003). Lee & Strong (2003) also prioritized role-based DQ dimensions differently during the data process. For instance, data collectors are knowledgeable about data collection, and data custodians are familiar with data storage and maintenance; thus data collectors and custodians might know more about collecting and maintaining accurate, complete, and accessible data. However, since data consumers may know more about the utility of the data, they might be more knowledgeable about data relevancy (Lee & Strong, 2003). Knowing the data roles and understanding data processing patterns also benefits the DQ in genome curation. In addition, technical problems and DQ skills should be considered to ensure DQ in genomics research community. Guided by General Systems Theory, Chung, Fisher, and Wang (2002) developed an educational framework based on a survey of data quality professionals with different job responsibilities while categorizing respective DQ skills into three categories: Technical capabilities, Adaptive capabilities, and Interpretive capabilities. Reported findings indicated that executives and managers regarded interpretive capabilities as critical for understanding

organizational implications of DQ, while consultants, project managers, and analysts ranked highly the adaptive capabilities that can identify user requirements and measure user DQ needs (Chung, Fisher, & Wang, 2002).

Genomics scientists may play multiple roles in genome curation work. They are called end-users when utilizing others' curation data or genomic information from public databases. End-users might collaborate with others on the curation of community shared genomics data (Good, Clarke, de Alfaro, & Su, 2012; Mons et al., 2008). Scientists might also adopt or be assigned a data curator role and focus on genome data curation produced by a project team, lab or community (Shimoyama et al., 2009). In many of these situations the curated data is reused within their lab/group in continuing experimental work. A number of genome scientists are required to play dual-roles (as both end-users and curators) due to the budget limitations or research purposes. They help develop genome curation resources for model organisms and specialized genomic databases. They curate their own genomics data as curators, while utilizing these data or similar information from public domains as end-users. Some of their curation work can be shared with others outside their lab. For example, scientists generated primary sequence data for their community curation efforts (Good, Clarke, de Alfaro, & Su, 2012), or curated data for a model organism (Shimoyama et al., 2009), or worked at National Center for Biotechnology Information (NCBI) curating large scale of genomics data (Pruitt, tatusova, Brown, & Maglott, 2012).

A special group of professional scientists were hired as curators to collect, annotate, and validate genome curation information by utilizing, cataloging, and archiving curated genomics data that are shared by genomic databases. As data collectors or producers do, curators play an active role in many genome curation data quality assurance activities. They also comprise points of views collected from the genome community to draft the genome curation policy for guiding best practices for data practice and curation (Pruitt, Tatusova, Brown, & Maglott, 2012). In addition to creating collection-building tools for genomic information repositories (Vasilevsky et al, 2012; Yang et al., 2011), curators preserve curation records, check the discrepancies and redundancy of the data, and then produce high quality, manually curated reference sequence sets (Pruitt, Tatusova, Brown, & Maglott, 2012). Furthermore, curators help develop data standards that can fully capture necessary curation data elements (Liolios et al., 2010; Samuel, Gussman, & Klumke, 2008).

Curators are also required to take responsibility for data quality control in addition to the other duties of various genome curation tasks. For example, major public genome database curators play an active role in supporting both the data preparation and quality of data submission for the data submitter. To ensure the quality of an automated data acquisition process, they also trace the provenance of curation records to verify the sources of extracted information (Shimoyama et al., 2009). Particularly, when referring to the 'Evidence-based' genome curation practice (Zhou et al., 2008), the data quality assurance process was also integrated as part of the curation process in a community-based system (Huang, Lu, Hunter, & Liang 2011). For example, curators implemented a two-tier review system for curation quality assurance consisting of an initial phase of peer review followed by detailed review by senior curators (McCarthy et al., 2007). The review system allowed internal interactions to assure the quality of curation before release of records to the public (Shimoyama et al., 2009). Curators also communicated externally with researchers to ensure the accuracy of curated information and to foster data exchanges with research laboratories (Shimoyama et al., 2009).

MacMullen (2006) reported gene ontology curation variations and data quality perceptions from the curator's perspective. The content analysis of curation reports indicated that perfect consistency for gene ontology curation is not necessary as long as the data were acceptable to both end-users and curators (MacMullen, 2006). The study, however, lacked empirical evidence concerning data quality perception from the end-user side (MacMullen, 2006). The curation role-based data quality assurance activities for end-users or curators can be quite different. For example, end-users actively provide feedback, comments, or suggestions for curation process improvements and system evaluations (Good, Clarke, de Alfaro, & Su, 2012). Data curators play important roles in data quality control activities including data validation, and data provenance etc. (Good, Clarke, de Alfaro, & Su, 2012; Shimoyama et al., 2009; Vasilevsky et al., 2012). Dual-role users help coordinate necessary communications in order to solicit data quality requirements from user groups for better data quality assurance (Shimoyama et al., 2009; Good, Clarke, de Alfaro, & Su, 2012).

As for the genome curation skills, curators, end-users, and dual-roles users need an intermediate level of genomics knowledge to interpret the genomic data (Burge et al., 2012). In addition to knowing routine data mining and other bioinformatics skills, curators are expected to obtain critical thinking and technical writing

skills for content selection using a “gold standard” (Howe et al., 2008). Communication skills, along with other interpersonal skills were regarded together as adaptive skills to demonstrate the ability to effectively interact with data users, managers, and other stakeholders (Pierce, 2003). Curators need communication skills to effectively exchange ideas with end-users, IT supporters, and database developers. When developing curation guidelines and databases, curators often work with a diverse, interdisciplinary curation team and must ensure consistency in data representation and interpretation (Shimoyama et al., 2009). Curators are also responsible for developing the curation manual, which provides concrete instructions regarding the approaches and steps for capturing and annotating complex and detailed data from the literature (Salimi & Vita, 2006).

Genomics researchers and curators usually make consensus judgments or best “guess” judgments based on the literature they found. Working together, they have to determine if they identify contradictions between the literature and what is used for annotation (Stein, 2001). In the earlier study, the authors explored the genome curation community’s perceptions of data quality requirements and developed a proposed data quality (DQ) model (Huang, Stvilia, Jörgensen, & Bass 2012). This work extends that study by examining role-specific data quality needs and priorities. The study can contribute towards identifying, explaining and bridging differences in data quality perception among different curation roles for the same set of data curation tasks. By identifying curation role-based needs and priorities for data quality, scientists and other stakeholders can generate tools and applications that can help the genomics research community design data policies, training modules, and infrastructure configurations which are attentive to those role based differences.

Procedures

Research Design

The study used a survey method to collect data, employing a survey instrument that was adapted from DQ dimensions and skills requirement questions from previous data quality surveys found in the literature (Chung, Fisher, & Wang 2002; Huang, Stvilia, Jörgensen, & Bass 2012; Wang & Strong, 1996). The scenario based task analysis method (Carroll, 1997; Go & Carroll, 2004) was used to develop two hypothetical scenarios representing common curation tasks that can build shared understanding and knowledge of activities by

codifying requirements for genome curation quality dimensions and skills. Subjects were given these two scenarios (See Appendix 1). In Scenario One, subjects were asked to rank *the top five* data quality dimensions (from a total 17 DQ dimensions) by their importance (see Figure 1); in Scenario Two subjects were asked to rank *the top five* data quality assurance skills (from a total of 17 DQ skills) also by their importance (see Figure 1). Scientists with different curation roles answered questions from both hypothetical scenarios, so that their perception differences due to their curation roles could be identified, given the same set of survey questions. At the end of the rankings, scientists were asked to provide open-ended comments on the survey questions' comprehensibility and clarity. Scientists were also given the opportunity to share as well, any concerns they had regarding data quality or skills related to genome annotation and curation.

Methodology

In order to identify potential survey participants with domain knowledge in genome research and annotation, literature searches were conducted within the PubMed database, (http://www.ncbi.nlm.nih.gov/pubmed/), using the search query "genome annotation". Results yielded 1,504 articles published between 09/01/2006 and 09/01/2009 that provided 2,782 email addresses for article authors. The email addresses were randomly sampled resulting in sending 240 survey invitations via emails. Of the 147 survey responses, demographics revealed 85% of the survey participants held PhD's, 75% worked in academia, and 69% resided in the United States. Educational demographics revealed that 60% had backgrounds primarily in biology and 30% in bioinformatics. Other demographic information such as age, gender, and work experience were also collected.

The subjects were grouped by their curation roles (curator, end-user, and dual-roles). In particular, the survey asked for the participants' curation roles using the following question: "How do you identify your role(s) in genome curation work (select all that apply)?" The subjects could choose the following options: "A regular user accessing the public database for sequences retrieval and analysis;" "a regular user to make sequence submission;" "work with the Genbank or other public database on data curation;" and "other." The total survey

participants (n = 147) identified themselves with the three groups: end-users (87), curators (42) and those who played both roles (18) in relation to genomics data. The study used the Qualtrics software (http://www.qualtrics.com) to distribute the survey and collect data online. The survey data was analyzed with STATA 11 software (College Station, Texas, USA) to conduct descriptive statistics, and chi-square analysis. Contingency 2x2 tables were created and tested using chi-square tests based on the number of top-five ranking occurrences (Tables 1 and 2).

To further understanding the preference ratings of the DQ dimensions and skills in selected DQ dimensions and skills categories from three user groups, the lead author firstly calculated the percentages of the frequencies for each of seventeen DQ dimensions and DQ skills being ranked by the users as *the top five* DQ dimensions or skills. Secondly, the calculated percentages of each DQ dimension and skill were sorted from the largest to the smallest, and then the cumulative percentage for each DQ dimension and skill were also calculated (see Appendices 2 and 3). The cumulative percentage for each DQ dimensions or skills were calculated as follows:

$$Yi = \sum_{i=1}^{k} Xi$$

Yi was defined as the cumulative percentage for the *ith* ranking of DQ dimensions or skills accumulated from the sum of the percentiles from the first DQ dimension or skill ranking to the *k*. *i* has values from 1 to *k*. *k* *k* takes the *1st* to *17th* rank (the total number of DQ dimensions or skills is 17). *X* is 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. To this end, the value of *Y* for the last accumulated ranking (*17th*) for DQ dimensions or skills is 100% (see Appendices 2 and 3). Only those with cumulative rankings of less than 90% were kept since the remaining 10% only counted for a trivial portion. DQ dimensions and skills ranked within the top 90% accumulated ranking lists were further grouped into the categories using previous data quality dimensions and skills models (see Figure 1). Finally, the percentages of DQ dimensions and skills in each category were aggregated with the sum of the percentile of each dimension or skill in a category. These aggregated percentages were ranked in the decreasing order (see Appendices 2 and 3).

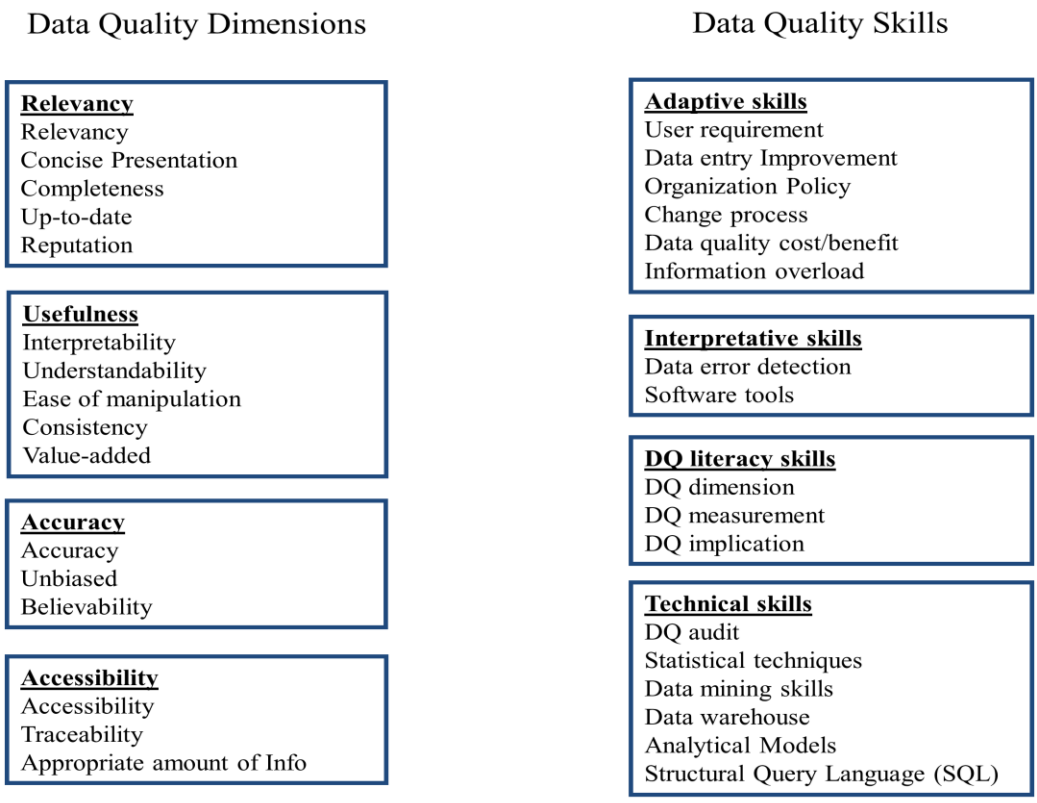

Figure 1. DQ dimensions (Wang and Strong, 1996; Huang, Stvilia, Jörgensen, & Bass 2012) and skills (Chung, Fisher, &Wang, 2002; Huang, Stvilia, Jörgensen, & Bass 2012) used for accumulated percentage analysis.

Findings

Chi-square analysis revealed several statistically significant differences ($p < 0.05$) in the top five DQ dimensions and DQ skills by survey participants who perform different curation roles (Table 1). Analysis further revealed particularly for end-users and curators that trade-offs exists for certain DQ dimensions as well as DQ skills. Data curation models or policies can be defined more specifically to meet role based users' needs and new curation procedures and data standards need to be developed in order to accommodate different needs among users.

The descriptive statistical analysis of the survey data (Table 1) revealed the following top-five most important DQ dimensions ranked by the end-users: Accuracy, Accessibility, Believability, Completeness, and Up-to-date. In addition, Accuracy, Accessibility, Traceability, Completeness, and Up-to-date were identified as the five most important DQ dimensions by curators. As for the dual-role scientists, the five most important DQ dimensions were the following: Completeness, Accuracy, Consistent Representation and Interpretability (tie), and Accessibility.

Table 1. Rankings of DQ dimensions based on the curation roles.

Attribute	End user(N=87)		Both(N=18)		Curator(N=42)		χ^2	χ^2	χ^2
	Mean rank	Ranked by	Mean rank	Ranked by	Mean rank	Ranked by	(End user vs Both)	(End user vs Curator)	(Both vs Curator)
Accessibility	2.1	59 (67.8%)	1.5	8 (44.4%)	1.8	30 (78.9%)	3.528	0.173	***3.951***
Accuracy	1.9	66 (75.9%)	2.1	10 (55.6%)	1.6	30 (78.9%)	3.076	0.292	1.429
Appropriate amount of information	2.8	23 (26.4%)	2.6	8 (44.4%)	3.3	12 (31.6%)	2.324	0.065	1.429
Believability	2.7	41 (47.1%)	3.6	5 (27.8%)	1.7	9 (23.7%)	2.268	***7.881***	0.284
Completeness	2.7	39 (44.8%)	3.5	13 (72.2%)	4.0	18 (47.4%)	***4.477***	0.045	***4.351***
Concise representation	4.4	6 (6.9%)	5.0	1 (5.6%)	3.0	6 (15.8%)	0.231	0.701	0.932
Consistent representation	3.5	27 (31.0%)	2.4	9 (50.0%)	3.0	6 (15.8%)	2.381	***4.174***	***8.571***
Ease of manipulation	3.9	23 (26.4%)	2.7	3 (16.7%)	3.5	12 (31.6%)	0.764	0.065	0.952
Interpretability	3.8	22 (25.3%)	3.9	9 (50%)	3.5	6 (15.8%)	***4.377***	2.017	***8.571***
Relevance	3.7	8 (9.2%)	3.0	4 (22.2%)	2.0	3 (7.9%)	2.500	0.153	2.780
Reputation	3.3	7 (8.0%)	3.0	1 (5.6%)	4.0	3 (7.9%)	0.131	0.032	0.051
Security	3.8	6 (6.9%)	4.0	1 (5.6%)	5.0	3 (7.9%)	0.043	0.003	0.051
Traceability	3.8	13 (14.9%)	3.25	4 (22.2%)	3.9	21 (55.3%)	0.582	***17.93***	***4.000***
Unbiased	3.5	22 (25.3%)	1.6	5 (27.8%)	5.0	6 (15.8%)	0.048	2.017	1.532
Understandability	4.0	19 (21.8%)	4.5	4 (22.2%)	4.0	6 (15.8%)	0.001	***3.739***	2.414
Up-to-date	3.9	32 (36.8%)	4.7	3 (16.7%)	3.6	15 (39.5%)	0.998	0.014	2.177
Value added	4.4	9 (10.3%)	5.0	1 (5.6%)	3.0	3 (7.9%)	0.397	0.344	0.051

Note. Bold/Italics: Pearson's chi-square test values were statistically significant (p <0.05). Top five DQ dimensions for each group have the cell highlighted.

As for the perceptual differences identified among participants who play different curation roles, compared to end-users, curators ranked Traceability higher and they ranked Believability, Consistent Representation, and Understandability lower. These differences in the data quality priorities could be attributed to the ways these groups evaluate data. Curators cared more about provenance (Traceability), than about indirect quality dimensions such as Believability. Curators may have direct access to the curated data with

enough domain knowledge for quality assessment, while end-users may often have to rely on indirect evaluations.

As for the DQ skills, all user groups perceived Data error detection and Data mining skills in the top five skills (see Table 2). Chi-square analysis of DQ skills also revealed significant differences ($p < 0.05$) in the quality skills priorities among three groups of participants (Table 2). Curators, with their data roles and responsibilities, rate Data-quality Audit skills as one of the top-five skills. Curators were also found to care more about understanding user's requirements than the other two groups. In addition, curators had higher priorities for the SQL than end-users. Furthermore, end-users and scientists playing dual roles had high priorities for data quality literacy skills (DQ dimensions, DQ measurement), and Statistical skills (Table 2). End-users valued more highly the skills needed to deal with information overload and DQ Literacy skills such as DQ Dimensions. Users with dual-roles had higher priorities for statistical skills than did the other two groups.

Table 2. Rankings of DQ skills based on the curation roles.

Attribute	End user (N=87)		Both(n=18)		Curator(n=42)		χ^2	χ^2	χ^2
	Mean rank	Ranked by	Mean rank	Ranked by	Mean rank	Ranked by	(End user vs Both)	(End user vs Curator)	(Both vs Curator)
Analytic models	3.7	17 (19.5%)	3.7	3 (16.7%)	4.5	6 (14.3%)	0.080	0.534	0.056
Change process	4.0	8 (9.2%)	5.0	1 (5.6%)	3.5	6 (14.3%)	0.252	0.759	0.932
Data mining skills	3.2	41 (47.1%)	2.0	7 (38.9%)	2.6	15 (35.7%)	0.408	1.502	0.055
Data-entry improvement	3.0	22 (25.3%)	3.4	5 (27.8%)	4.5	6 (14.3%)	0.048	2.017	1.532
Data-error detection	2.8	60 (69.0%)	2.4	11 (61.1)	1.9	27 (64.3%)	0.420	0.283	0.055
Data-quality audit	3.6	20 (23.0%)	2.7	3 (16.7%)	3.4	15 (35.7)	0.348	2.32	2.177
Data-quality cost/benefit	4.3	4 (4.6%)	3.0	2 (11.1%)	5.0	3 (7.1%)	1.174	0.358	0.260
Data-quality dimensions	2.0	33 (37.9%)	1.6	8 (44.4%)	1.0	3 (7.1%)	0.266	***13.34***	***11.71***
Data-quality implication	2.4	25 (28.7%)	2.8	6 (33.3%)	3.0	6 (14.3%)	0.152	3.240	2.857
Data-quality measurement	2.3	36 (41.4%)	3.8	4 (22.2%)	1.3	18 (42.9%)	2.321	0.025	2.310
Data-warehouse set-up	3.4	18 (20.7%)	4.0	1 (5.6%)	3.5	6 (14.3%)	2.305	0.767	0.932
Information overload	3.2	10 (11.5%)	0.0	0 (0.0%)	0.0	0 (0.0%)	2.287	***5.233***	N/R
Organization policies	3.2	20 (23.0%)	2.3	3 (16.7%)	2.7	9 (21.4%)	0.348	0.040	0.179
Software tools	3.5	25 (28.7%)	4.8	4 (22.2%)	3.5	12 (28.6%)	0.317	0.0004	0.260
Statistical techniques	3.6	31 (35.6%)	3.5	10 (55.6%)	2.7	9 (21.4%)	2.487	2.671	***6.782***
Structural query language	0.0	0 (0.0%)	4.0	1 (5.6%)	1.0	3 (7.1%)	4.880	***6.362***	0.051
User requirement	3.1	17 (19.5%)	3.5	2 (11.1%)	4.3	18 (42.9%)	0.715	***7.789***	***5.714***

Note. Bold/Italics: Pearson's chi-square test values were statistically significant ($p < 0.05$). Top five DQ dimensions for each group have the cell highlighted.

As for the role-based difference in the rankings of four DQ dimension categories (Figure 2) from the previous reported DQ dimension grouping model (Wang and Strong, 1996; Huang, Stvilia, Jörgensen, & Bass 2012), curators ranked the Accessibility category highest, which in addition to the Accessibility dimension included Traceability and the Appropriateness of Information dimensions. End-users ranked the Accuracy category highest, which also included the indirect perceptual dimension of Believability. Dual-role-users, however, favored the Usefulness category which included Ease of Manipulation, Interpretability and Understandability dimensions (Figure 2).

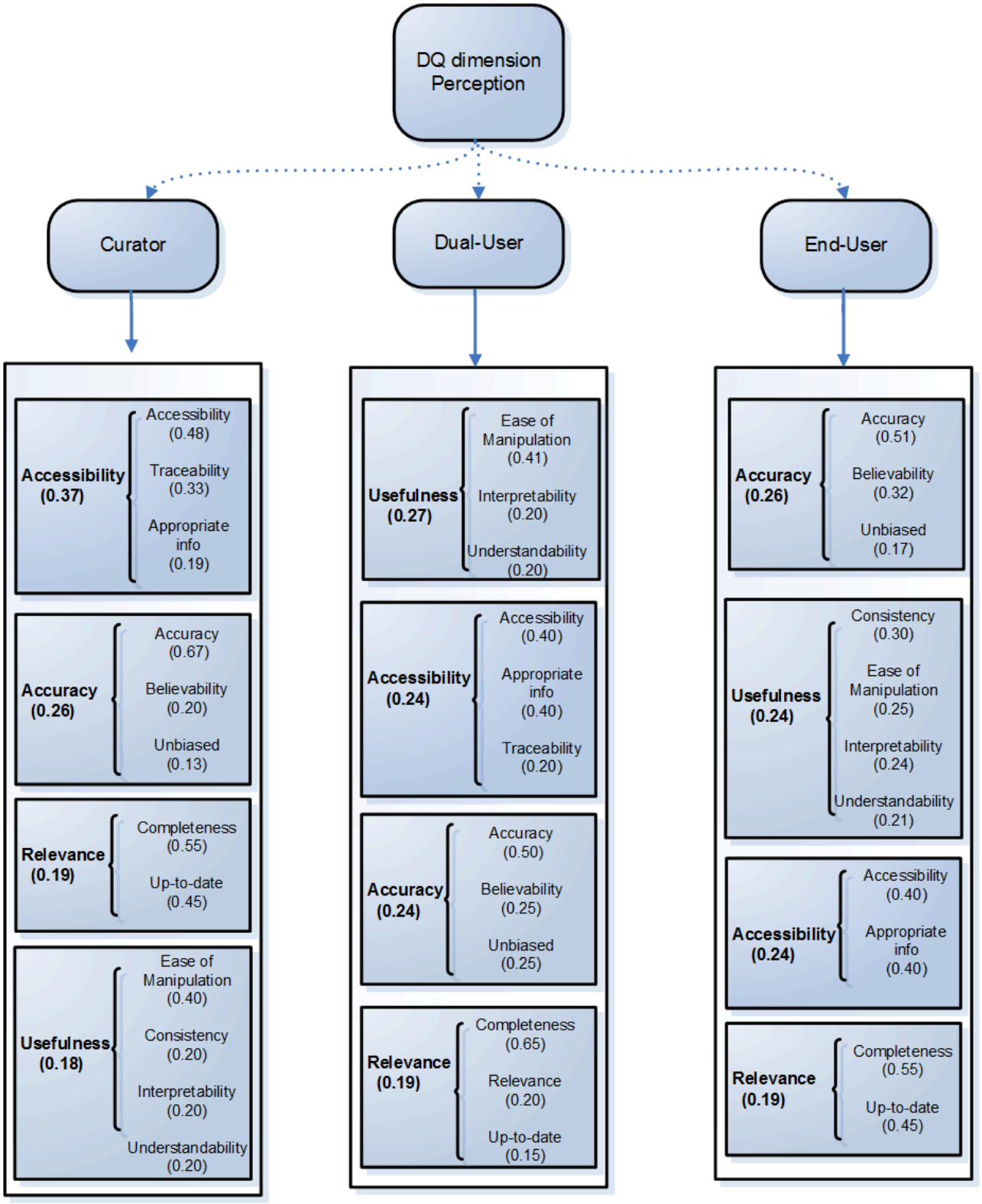

Figure 2. Role based DQ dimension grouping priorities. Only those with cumulative rankings of less than 90% were kept (see Appendix 2).

Similarly, in Figure 3, the study compared the rankings of the four DQ skill categories by participants playing different roles based on previous data skills models (Chung, Fisher, & Wang, 2002; Huang, Stvilia, Jörgensen, & Bass 2012). The technical skills category had the highest rankings among three groups, comprised of Statistical techniques, Data mining, DQ audit, and Analytical modeling skills (Figure 3). The Adaptive skills category was ranked higher by curators than others'. This category included the User requirement, Organization policy, and Data entry improvement skills.

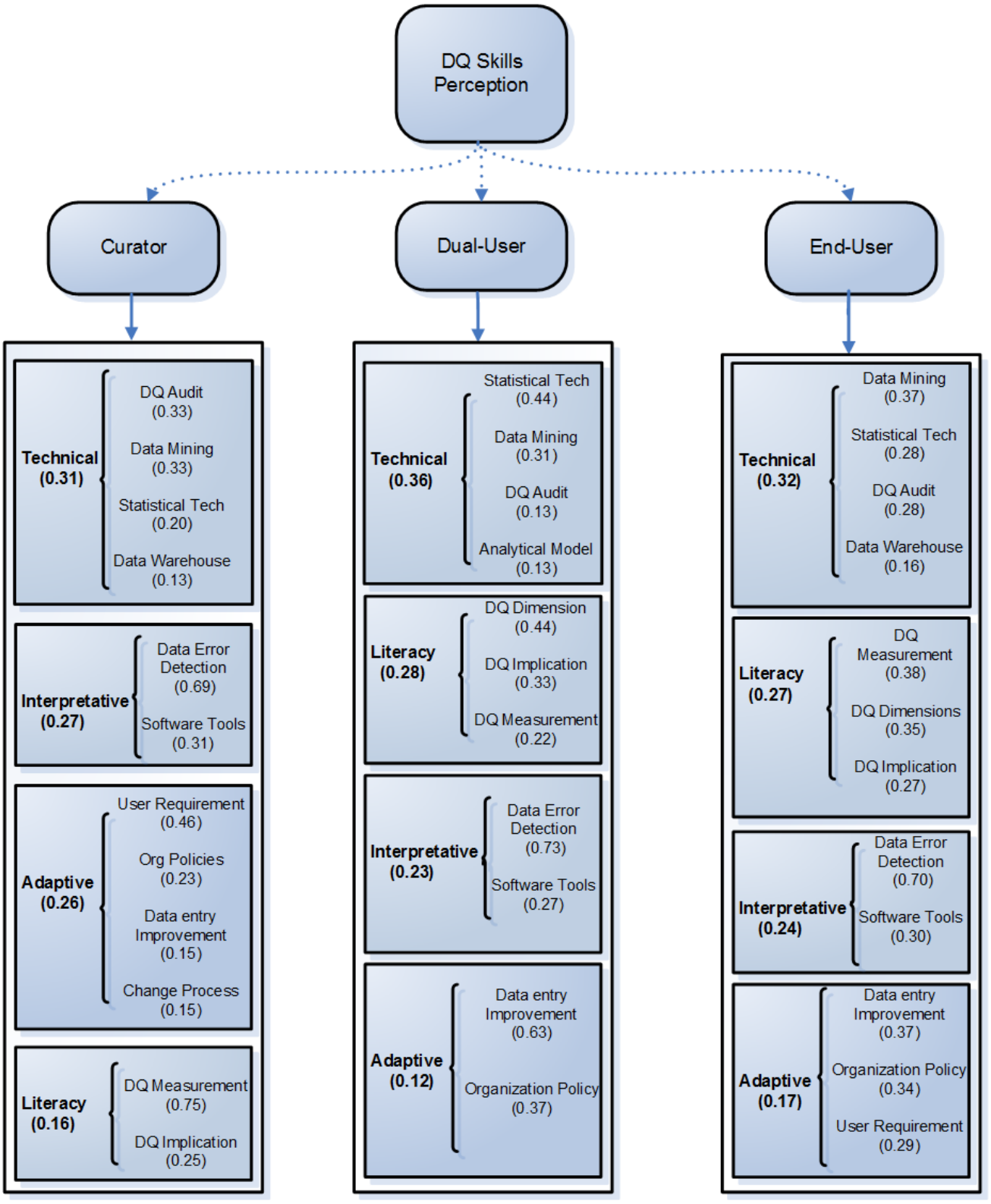

Figure 3. Role based DQ skills grouping priorities. Only those with cumulative rankings of less than 90% were kept (see Appendix 3).

Discussion

The findings of this study demonstrate the perceptual differences among curators, end-users and dual-role users in the prioritization of both DQ dimensions and DQ skills. The first research question focused on the DQ dimension perception gaps among users with different curation roles. Significant perception differences were found in Accessible Information (Accessibility, Traceability), Accuracy Information (Believability), and Useful Information (Consistency, Interpretability, Understandability) among curators, dual-role users, and end-users (Figure 2). Users with different curation roles also assigned different priorities in DQ skills requirements. These differences are seen in DQ Literacy skills (DQ dimensions), Adaptive skills (Information overload, User requirement), and Technical skills (Statistical techniques, SQL) among three user groups (Figure 3). As the concept of data quality is regarded as domain-specific and context-sensitive (Stvilia and Gasser, 2008), prioritization of DQ dimensions and skills requirements is found to be dependent on user roles. It is necessary to evaluate the role based data quality activities and related skills in genome curation for further development and support for these roles.

Impact of Curation Roles on DQ Dimensions Prioritization

The genomics profession has moved from a data production role to an activist role for promoting the wider use and data sharing of genomic data (Burge et al., 2012). As data intermediaries, curators directly assess data quality and "virtue" in genome curation. In addition, curators organize, interpret, edit genomics data, and make them accessible to other users. Like data consumers, end-users browse, retrieve, and download curation records and indirectly assess the data quality of curation records by relying on source reputation.

Similar to data collectors, or custodians, curators are expected to manage the community's data while maintaining the data integrity and accessibility to end-users. As evidenced in this study, curators demonstrated preference for the DQ criteria of Accessibility and Traceability. Traceability, which demonstrates provenance, was regarded as an important DQ dimension within the Accessibility construct by curators in genome curation. An ability to establish the provenance of data, including the types of changes made by whom and when, is essential for effective data management. Tracing the history of records allows curators to develop provenance-based data quality metrics that capture some of the dynamics of the quality of data and to enable better tailored,

quality assurance actions (Stvilia & Gasser, 2008). Recording data provenance and data curation actions also enables better reporting, communication, and coordination as well as more effective development of data curation strategies and policies to include not just members within the community but also outside stakeholders. Additionally, the logs of data curation activities and the conversations around those activities can be also used as a knowledge source to acculturate and train new curators and users of the database (Stvilia, Twidale, Smith, & Gasser, 2008).

Dual-role users, across the data roles of curators and end-users, deal with both in-house and external data. They would pay more attention to issues of data consistency and interpretation. The survey data also indicate that there is a trend or merging of roles for end-users and curators. In other words, the users actually access and utilize the genomic data and at the same time provide value-added data curation for long term preservation and data reuse. As end-users, genomics researchers utilize the genomic data to benefit their own research, but at the same time they contribute to the community with value-added annotation for valuable data reuse. Huss et al. (2008) predicted that in a community curation system the roles of curators and end-users will eventually merge and end-users will be empowered to participate in the curation process so that the community can cope with the ever-increasing amount of genomic data. This trend was shown in the current survey comments as well:

"I created a public database and submitted sequence and other data there."

"I worked as both end-user and curator previously."

"Created and managed genome curation resources."

Genome curation work is conducted in a highly collaborative environment. The end-users often act as "liaisons" in the genome curation process by supplying local contextual information through annotation and quality checking:

"Work on internal curation."

"When I found a discrepancy, I tell the curators."

"Work with bioinformatics specialists to supply biological context."

Scientists collaborate together and usually make consensus agreements or judgments with a "best guess" based on the literature and identify any contradictions between the literature and curation use. They also make recommendations for the improvement of the process infrastructure and associated data. The recommendations also aid in the evaluation of systems that can lead to understanding reasons for inconsistent or erroneous outcomes. Thus, the curation process could be transparent, and scientists could make improvements to the state of the art across the community, as the example below from the open-ended survey comments:

> "Often genome curation is like a game of telephone: a gene function is assigned based on the assigned function of a homologue, which was assigned based on the function of its homologue, etc.... Genome curation should take into account reliability of functional assignments based on a ***'consensus of curations'*** rather than actual experimental data which directly suggests or supports functional assignment for the organism and gene being annotated."
>
> "Would like to have seen more detail on tentative assignment by ***similarity***"

This trend of end-users actively participating in and collaborating with curators on quality assurance and knowledge creation activities (e.g., data curation) ultimately evolves into higher expectations for bioinformatics literacy, including data quality literacy on the part of users. A recent survey (Burge et al., 2012) showed that the biocuration community believed that a curator having research experience and a strong biological or bioinformatics background will tremendously benefit their curation work.

Impact of Curation Roles on DQ Skills Prioritization

End-users assigned higher priorities to the data quality assessment skills and the skills needed to identify useful information (i.e., "finding a needle in a haystack"), while curators valued more highly on the skills needed to make data useable. Previous studies indicated that job responsibilities influence what skills users

perceive as important (Chung, Fisher, & Wang, 2002). Users such as consultants, project managers, and analysts are responsible for converting user requirements to technical specifications; this was perceived as the most important set of skills (Chung, Fisher, & Wang, 2002). Similarly, curators believe that information concerning user requirements enables the curator to stay sensitive to the needs of users by having knowledge of the methods and techniques for use requirements.

In fact, the biocurator community has started addressing the critical issues in order to obtain a clear picture for user needs. This will help to develop curation and annotation pipelines and tools for the genomic research community (Tan et al, 2010). This is particularly important when curating high quality genome sequences, as well as satisfying community needs in the long run. Curators can effectively translate the users' requirements into technical specifications to develop extensive and deeply curated resources for researchers (Shimoyama et al., 2009). Curators also regard using database tools (e.g., SQL) for curation as important, since their jobs involve dealing with data storage and retrieval in large scale databases. Given the fact that genomic data is less structured and more complex than in business domains or other fields, curators or dual-role users would like to have additional database tools, statistical tools for data mining and innovative data representation tools that help the genomics community design operational quality metrics for assessing curation data (Shimoyama et al., 2009).

End-users indicated stronger priorities for the skills that would help them deal with information overload than did other groups. Information overload occurs when a person has difficulty understanding an issue and making decisions due to an overabundance of information that prohibits effective utilization (Waddington, 1998). Genomics research is data driven. There are significant technical and organizational challenges in ensuring provenance and reproducibility of genomic resources. Often, the information overload problem exists simply because the end-users are not aware of or do not know how to use available data search, retrieval and analysis tools. It is also because end-users are not familiar with the policies or workflows of the genome curation organization. In genome curation, information sources are highly technical and involve acquiring, processing, analyzing and distributing data to users using different kinds of tools and systems (Burge et al.,

2012; Davis et al., 2011). End-users have to deal with the flood of genomic data, thus they have a strong need for data skills which facilitates coping with the information overload problem.

The DQ Technical skills were ranked higher for the three user groups, while the DQ Adaptive skills were ranked higher for curators (Figure 2). Adaptive skills are the set of skills that allow effectively interacting with different users (Pierce, 2003), which can help curators define the best strategies to digitize and integrate disparate pieces of scientific data to facilitate users' needs. Well curated genomic resources that enable easy discovery and are reusable can help reduce the burden of information overload. For instance, accessibility barriers might be perceived differently among users. Curators might think genomic data is simply physically unavailable rather than inaccessible. However, end-users might also consider the barriers as technical, for various reasons: 1) data may be represented in different formats which cannot be recognized; 2) the coded data may scarcely be interpreted; and 3) there is a large volume of data which is difficult to locate (Strong et al., 1997).

In this study, users with a dual-role are familiar with both the needs of research work for data and analysis tools needed to perform data curation work. Hence, it was not surprising that dual-role users ranked statistical analysis skills highly as they can benefit both for their curation and research work. They ranked the ability to define data quality dimensions highly as well. This study suggests that curators can utilize genomic data to benefit their own research, and the same time their curation roles can promote value-added curation for long term preservation and reuse.

In the research reported here the survey respondents were genomic research domain experts and highly educated scholars, most of them possessing PhD degrees. End-users and dual-role users demonstrate a strong requirement for understanding data quality concepts and skills needed to perform data quality assurance tasks. Understanding DQ dimensions is ranked the highest among the data quality skills for these groups. This indicates that scholarly users would like to learn more about data quality issues and the related data management skills needed both for genomic data preservation and facilitating user understanding of and engagement with genomic data resources (Howe, et al., 2008; Goth, 2012). Grasping the necessary knowledge

of data quality concepts and their related assessment methods, as well as ways in which to identify potential data quality problems, is a prerequisite for every user to accomplish high quality curation activity. Without the fluency of data quality literacy, it is hard to facilitate data usage with other data quality skills to provide efficient data quality assurance for genome curation work.

Role based Data Quality Trade-offs

The differences in users' roles appear to eventually lead them to prioritize data quality criteria differently in genome curation work. It is necessary to collect, benchmark and communicate the differences among users' perceptions through empirically-based feedback in order to design strategies to improve curation quality for the community. The results of this study can be used to extend the evaluation of curation role-specific data quality concerns. The study found that curators ranked the Accessibility dimension higher than the Accuracy dimension, while end-users did the opposite (Figure 2). This may also point to the presence of trade-offs among different data quality dimensions depending on the users' role. When a user makes a decision, it is affected by various factors including information overload, task complexity, time, and other costs. Examples include the trade-offs between representation quality and accuracy, completeness and accessibility, and accessibility and timeliness. Users may have to accept and tolerate data with errors or incomplete data in exchange for having access to important information fast. These data quality trade-offs can be justified when complete data is difficult to obtain and access in a given time period. Accessibility is expensive when dealing with a high volume of the data. As the literature suggests, data quality assurance is not free and it needs to be optimized based on the user's priorities which include attending to possible tradeoffs among different quality dimensions (Ballou and Pazer, 1995; Stvilia et al, 2007). Some of the trade-offs identified by survey comments are as follows:

"Some uncalled nucleotides ok, as long as they are designated as Ns."

"Requiring completeness may cause lags in data entry.…"

"I think the most important thing in this is the accessibility, which determines the time spent on it."

These DQ trade-offs are acceptable as long as the overall quality of data is "good enough" for its utilization in the research task at hand. The current research collected empirical data using a survey of the members of a specific scientific community regarding their perceived priorities for data quality criteria and related data quality skills in the context of genomic data curation work (see Figures 2 to 3). Scientists with different roles, given common genome curation tasks with the same set data quality requirements, reach different decisions to prioritize the importance of DQ dimensions and skills. It would be beneficial to inform the genomics research community of such role-based data quality perception differences, in particular data quality dimensions and skills requirements. The findings of this study can also be used in developing genomic data curation policies, procedures and training modules that could be used not only by the current curation team, but also by future institutional participants or end-users who may not have extensive training in data management and data curation.

Limitations

There are limitations to the current study. The rankings of DQ dimensions and skills used to fine-tune the genome curation quality models are based on recollected behavior and the importance of these concepts perceived by these specific survey participants at that particular time. These are only approximations of the participants' actual value models for quality and data quality skills used in practice and should be considered pointers for future system and best practices development considerations rather than definitive results. Future research collecting additional empirical data on the community's data-curation and quality-assurance practices through observations and interviews can provide further insight into the genome-annotation quality relationships. In addition, the DQ skills used as items in the survey instrument were based on the findings from previous studies (Chung, Fisher, & Wang, 2002). As new data management technologies evolve (e.g., 'NoSQL' databases, 'cloud' technologies), these items and related constructs will need to be revisited to update the community's priorities for data quality assurances skills.

Conclusion

Genome curation activity is a collaborative process constructed through a dynamic complex interaction among its users who contribute to different tasks and play different roles. It requires curators to read tremendous amounts of literature and have solid domain and background related knowledge. It also requires curators to be flexible and adaptive in order to deal with different scales of genomic related data, to make confident judgments to annotate related information in the genome context, and to make sure to capture all the related information within the data model. Stakeholders in the genome curation community need to clarify the transparent and explicit role assignments and responsibilities for their curation tasks, and identify the expected skills and expertise for required curation tasks and activities.

The study recognized the unique and significant genome curation roles that play in the assessment of trustworthiness, and evidential value of the curation resources. The study also identified the important curation roles played by traceability to document the origin of curated information. As for the curation skills, the study recognized the important skills of inter-process communication that can possible reduce the ambiguity and disagreement of collaborative curation work. The role based DQ rankings developed in this study can help develop collaborative curation system architecture to support role based curation quality assurance activities, and strategies tailored to the genomics community. Practitioners in institutional repositories also need to understand what scientists are doing and their data quality and skills requirements for data curation, so that they can provide the necessary support and services to facilitate collaborative data curation and research data management. The results also supported the development of curation polices, institutional mandates, and educational curriculum in data practice and curation for both scientists and practitioners. Furthermore, this study serves to support development of training modules that will enable scientists to smoothly transition the job skills necessary in data quality as they switch roles from end-users to curators.

Genomics research is very data intensive. This study found that genomics researchers are aware of data quality trade-offs when dealing with enormous amounts of genomic data and facing information overload

challenges. Scientists with different curation roles, given common curation tasks, lack a consensus for selecting data quality criteria for genome data curation. Scientists' data quality expectations change as their work roles are pluralistic and evolving, and the curators must strive to keep up with newer or emerging skills. Identification of these differences can help develop data management architectures to support role-based community curation. Genome curation is no longer conducted solely by professional curators. The genome curation work now resides in a community system in which the records are cross-referenced and curated by scientists with consensual understandings of expected data quality requirements (Stein, 2004). In order to achieve a consensus in the decision making process, it would be beneficial for the genomics community to conduct seminars and data quality education prior to deciding upon data quality assurance activities. Future research studies should involve collection of additional data and the development of operational models of existing and potential tradeoffs so these can be used in practice to optimize quality assurance activities.

References

Bade, D. (2007). Rapid cataloging: Three models for addressing timeliness as an issue of quality in library catalogs. *Cataloging & Classification Quarterly*, *45*(1), 87-123.

Ballou, D. P., & Pazer, H. L. (1995). Designing information systems to optimize the accuracy-timeliness tradeoff. *Information Systems Research*, *6*(1), 51-72.

Baumgartner, W. A., Cohen, K. B., Fox, L. M., Acquaah-Mensah, G., & Hunter, L. (2007). Manual curation is not sufficient for annotation of genomic databases.*Bioinformatics*, *23*(13), i41-i48.

Bartlett, J. C., & 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*, *56*(5), 469-482.

Burge, S., Attwood, T.K., Bateman, A., Bateman, A., Berardini, T.Z., Cherry, M., et al. (2012). Biocurators and biocuration: surveying the 21st century challenges. *Database: The Journal of Biological Databases and Curation*, bar059.

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

Castro, A. G., Thoraval, S., Garcia, L. J., & Ragan, M. A. (2005). Workflows in bioinformatics: meta-analysis and prototype implementation of a workflow generator. *BMC bioinformatics*, *6*(1), 87.

Chung, W., Fisher, C., & Wang R. (2002). What skills matter in data quality? Paper presented at the 7th International Conference on Information Quality (ICIQ-02), Cambridge, MA.

Davis, A.P., Wiegers, T.C., Murphy, C.G., & Mattingly, C.J. (2011). The curation paradigm and application tool used for manual curation of the scientific literature at the Comparative Toxicogenomics Database. *Database: the journal of biological databases and curation*, bar034.

Dubay, C., Gorman, P., & Hersh, W. (2004). Applying task analysis to describe and facilitate bioinformatics tasks. In *Medinfo 2004: Proceedings of the 11th World Conference on Medical Informatics,* p. 818, *San Francisco,* IOS Press.

Evans, J. R., & Lindsay, W.M. (2005). *The management and control of quality*. Cincinnati, OH: Thomson Learning.

Friedlander, A. & Alder, P. (2006). To stand the test of time: Long-term stewardship of digital data sets in science and engineering. *A report to the National Science Foundation from the ARL workshop on New collaborative relationships: The role of academic libraries in the digital data universe.* Retrieved from http://www.eric.ed.gov/PDFS/ED528649.pdf.

Frické, M., & 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*, *55*(3), 238–245.

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

Goth, G. (2012). Preserving digital data. *Communications of the ACM*, *55*(4), 11-13.

Good, B.M., Clarke, E.L., de Alfaro, L., & Su, A.I. (2012). The Gene Wiki in 2011: community intelligence applied to human gene annotation. *Nucleic acids research*, *40*(D1), D1255-D1261.

Heer, J., & Kandel, S. (2012). Interactive analysis of big data. *XRDS: Crossroads, The ACM Magazine for Students*, *19*(1), 50-54.

Hoffmann, R. (2008). A wiki for the life sciences where authorship matters.*Nature genetics*, *40*(9), 1047-1051.

Homburg, C., Droll, M., & Totzek, D. (2008). Customer prioritization: does it pay off, and how should it be implemented? *Journal of Marketing*, *72*(5), 110-130.

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., et al. (2008). Big data: the future of biocuration. *Nature*, *455*(7209), 47–50.

Huang, H., Lu, J., Hunter, W., & Liang, S. (2011). Using IBM Content Manager for genomic data annotation and quality assurance tasks. *IBM Journal of Research and Development*, *55*(6), 13.

Huang, H., Stvilia, B., Jörgensen, C., & Bass, H. W. (2012). Prioritization of data quality dimensions and skills requirements in Genome annotation work. *Journal of the American Society for Information Science and Technology*, *63*(1), 195-207.

Huss, J., Orozco, C., Goodale, J., Wu, C., Batalov, S., Vickers, T., Valafar, F., & Su, A. (2008). A gene wiki for community annotation of gene function. *PLoS Biology 6*(7), e175.

Kahn, B. K., Strong, D. M., & Wang, R. Y. (2002). Information quality benchmarks: product and service performance. *Communications of the ACM, 45*(4), 184-192.

Klein B (2002). When do users detect information quality problems on the World Wide Web? In *Proceedings of the Americas Conference on Information Systems AMCIS 2002*, Paper 152. Available at: http://aisel.aisnet.org/

Kuhn, R. M., Haussler, D., & Kent, W. J. (2013). The UCSC genome browser and associated tools. *Briefings in bioinformatics, 14*(2), 144-161.

Lankes, D.R. (2008). Credibility on the internet: shifting from authority to reliability. *Journal of Documentation*, *64*(5), 667–686.

Lee, Y., Strong, D. (2003). Knowing – why about data processes and data quality. *Journal of Management Information Systems*, *20*(3), 13–39.

Liolios, K., Chen, I.M., Mavromatis, K., Tavernarakis, N., Hugenholtz, P., et al. (2010). The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research, 38*, D346-D354.

Lynch, C. (2009). Jim Gray's fourth paradigm and the construction of the scientific record. In T. Hey, S. Tansley, & K. Tolle (Eds.), The fourth paradigm: Data intensive scientific discovery. Redmond, WA: Microsoft Research, pp. 177-183.

MacMullen, W. (2006). Contextual analysis of variation and quality in human-curated gene ontology annotations. (Unpublished doctoral dissertation). University of North Carolina: Chapel Hill.

MacMullen, W. J., & Denn, S. O. (2005). Information problems in molecular biology and bioinformatics. *Journal of the American Society for Information Science and Technology*, *56*(5), 447-456.

McCarthy, F.M., Bridges, S.M., Wang, N., Magee, G.B., Williams, W.P., Luthe, D.S., et al. (2007). AgBase: a unified resource for functional analysis in agriculture. *Nucleic acids research*, *35*(suppl 1), D599-D603.

Mons, B., Ashburner, M., Chichester, C., van Mulligen, E., Weeber, M., et al. (2008). Calling on a million minds for community annotation in WikiProteins. *Genome Biology*, *9*(5), R89.

Özdemir, V., Badr, K. F., Dove, E. S., Endrenyi, L., Geraci, C. J., Hotez, P. J., et al. (2013). Crowd-Funded Micro-Grants for Genomics and "Big Data": An Actionable Idea Connecting Small (Artisan) Science, Infrastructure Science, and Citizen Philanthropy. *Omics: a journal of integrative biology*,*17*(4), 161-172.

Pruitt, K.D., Tatusova, T., Brown G.R., & Maglott, D.R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Research*, *40*(D1), D130-D135.

Reed, J. L., Famili, I., Thiele, I. & Palsson, B. O. (2006). Towards multidimensional genome annotation. *Nature Reviews Genetics*, *7*(2), 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*, *53*(2), 145–161.

Salimi, N. & Vita, R. (2006). The biocurator: connecting and enhancing scientific data. *PLoS Computational Biology*, *2*(10), e125.

Salzberg, S. (2007). Genome re-annotation: a wiki solution? *Genome Biology*, *8*(1), 102–102.

Samuel, V., Gussman, A., & Klumke, W. (2008). Toward an online repository of Standard Operating Procedures (SOPs) for (meta)genomic annotation. *OMICS: A Journal of Integrative Biology*, *12*(2), 137–141.

Sanderson, K. (2011). Bioinformatics: curation generation. *Nature*, *470*(7333), 295–296.

Shimoyama, M., Hayman, G. T., Laulederkind, S. J. F., Nigam, R., Lowry, T. F., et al. (2009). The rat genome database curators: who, what, where, why. *PLoS Computational Biology*, *5*(11), e1000582.

Stein, L. (2001). Genome annotation: From sequence to biology. *Nature Reviews Genetics, 2*, 493–503.

Strong, D., Lee Y., & Wang R. (1997). Data quality in context. *Communication of the ACM*, *40*(5), 103–110.

Stvilia, B., Gasser, L., Twidale, M., & Smith L. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, *58*(12), 1720–1733.

Stvilia, B., Mon, L., Yi, Y. (2009). A model for online consumer health information quality. *Journal of the American Society for Information Science and Technology*, *60*(9), 1781-1791.

Vasilevsky, N., Johnson, T., Corday, K., Torniai, C., Brush, M., et al. (2012). Research resources: curating the new eagle-i discovery system. *Database: the journal of biological databases and curation*, bar067.

Waddington, P. (1998). Dying for information? *A report on the effects of information overload in the UK and Worldwide. CNI: Coalition for Networked Information*. Retrieved from http://old.cni.org/regconfs/1997/ukoln-content/repor~13.html

Wang, R., & Strong, D. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–35.

Yang, X., Ye, Y., Wang, G., Huang, H., Yu, D., & Liang, S. (2011). VeryGene: linking tissue-specific genes to diseases, drugs, and beyond for knowledge discovery. *Physiological genomics*, *43*(8), 457-460.

Zhou, Y., Ramachandran, V., Kumar, K.A., Westenberger, S., Refour, P., et al. (2008). Evidence-based annotation of the Malaria Parasite's genome using comparative expression profiling. *PLoS ONE*, *3*(2): e1570.

Figure Legends

Figure 1. DQ dimensions (Wang and Strong, 1996; Huang, Stvilia, Jörgensen, & Bass 2012) and skills (Chung, Fisher, &Wang, 2002; Huang, Stvilia, Jörgensen, & Bass 2012) used for accumulated percentage analysis.

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

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