This is a preprint of an article accepted for publication in Journal of the American Society for Information Science and Technology. Huang, H., Stvilia, B., Jörgensen, C., & Bass, H. (in press, 2011). Prioritization of data quality dimensions and skills requirements in genome annotation work. *Journal of the American Society for Information Science and Technology*,

Prioritization of Data Quality Dimensions and Skills Requirements in Genome Annotation Work

Hong Huang

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

Telephone: (813) 974-6361; Fax: (813) 974-6840; E-mail: honghuang@usf.edu

Besiki Stvilia and Corinne Jörgensen

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

Hank W. Bass

Department of Biological Science, Florida State University, Tallahassee, Florida, 32306-4295.

Telephone: (850) 644-9711; Fax: (850) 645-8447; E-mail: bass@bio.fsu.edu

This is a preprint of an article accepted for publication in Journal of the American Society for Information Science and Technology. Huang, H., Stvilia, B., Jörgensen, C., & Bass, H. (in press, 2011). Prioritization of data quality dimensions and skills requirements in genome annotation work. *Journal of the American Society for Information Science and Technology*,

Abstract

The rapid accumulation of genome annotations, as well as their widespread reuse in clinical and scientific practice, poses new challenges to management of the quality of scientific data. This study contributes towards better understanding of scientist perception and priorities for data quality and data quality assurance skills needed in genome annotation. Our study was guided by a previously developed general framework for assessment of data quality and by a taxonomy of data quality skills, and intended to define context-sensitive models of criteria for data quality and skills for genome annotation. Analysis of the results revealed that genomics scientists recognize specific sets of criteria for quality in the genome-annotation context. Seventeen data quality dimensions were reduced to five factor constructs, and 17 relevant skills were grouped into four factor constructs. The constructs defined by this study advances the understanding of data quality relationships and is an important contribution to data and information quality research. In addition, the resulting models can serve as valuable resources to genome data curators and administrators for developing data-curation policies and designing DQ-assurance strategies, processes, procedures, and infrastructure. The study's findings may also inform educators in developing data quality assurance curricula and training courses.

Introduction

The objectives of genome annotation are to mark the key features of the genome and to link them to the related literature (Stein, 2004, p. 501). Such annotation is a collaborative activity, involving participation by many actors from different domains (e.g., researchers, clinical doctors) who might have different needs for and uses of identical information. For example, genome annotation links knowledge with specific gene products useful to develop personalized genomic medicine. Genome-annotation tasks include collecting raw genomic data and applying various tools for analysis of the primary data, i.e., utilizing available genomic information, and secondary data for production of functional genomics interpretation and new knowledge for promotion of human health.

Because of genome annotation's complexity, annotation errors can occur during the process (Brenner, 1999; Frohlich, Speer, Poustka, & Beissbarth, 2007; Pruitt, Tatusova, & Maglott, 2007; Samuel, Gussman, & Klumke, 2008; Schlueter, Wilkerson, Huala, Rhee, & Brendel, 2005). Ignoring these errors may cause serious problems for database users, curators, research scientists, or clinical doctors. Indeed, the affordability of genome sequencing has elevated the technology from pure science to employment in clinical practice like direct-to-consumer personal genome testing (McGuire, Diaz, Wang, & Hilsenbeck, 2009) with corresponding potentially strong social impacts on human life. Genomic medicine requires proper integration of genomic and clinical data from the molecular level (e.g., "*in vivo,*" "*in vitro,*" or "*in silico*") to the population level (e.g., public health genomics), demanding procedures that deliver high-quality products.

Genome annotation work may include different roles: annotation users, providers, and curators. Genomics scientists may play some or all of these roles in different task contexts. They can be users of and providers of genome annotations, as well as curators of their own, or community genome data. Data curation is a process a of managing data, including ensuring its quality –the availability and 'fitness' for use and re-use (Curry, Freitas, & O'Riáin, 2010; Lord & Macdonald, 2003). Similar to other domains, genome annotation work and annotation curation now moves to community level collaborations and collaborative content creation and sharing systems (Huss et al., 2008; Mons et al., 2008; Salzberg, 2007). Genomics scientists can use annotation records from a community database to produce their own annotated work and then deposit it to the same database. Some of them may also serve as curators or data quality stewards, formally or informally, and ensure the accuracy, completeness, as well as the consistency of annotations across the database and with the community's standards and literature (Bragge, Merisalo-Rantanen, & Hallikainen, 2005; Hermann, 2007; Marco, 2006; Stein, 2001). Indeed, genome annotations are "work in progress" metadata and genome annotation is an ongoing process. As new research findings become available, current annotations have to be updated and expanded (McNeal et al, 2007)

The concept of data quality (DQ) – 'fitness for use' - is a contextual, multidimensional concept (Strong, Lee, Wang, 1997; Stvilia et al, 2007). Several quality models and frameworks have been proposed in the literature (see Ge & Helfert, 2007, for a recent review) and provide a knowledge base valuable to researchers and practitioners alike, but they are not directly applicable to the context of genome annotation. The need remains for a context-sensitive model for genome annotation that would account for process- and community-specific sources of DQ variation, requirements, and priorities and for trade-offs among DQ criteria.

Indeed, data of high quality are those that meet the user's requirements (Evans & Lindsay, 2005). Understanding a user's perception of DQ and the requirements for it in a specific context is the first step to developing a user-specific contextual DQ model (McGilvray, 2008). Surprisingly little investigation has addressed the genomics scientist's perspective on DQ needs and skills in genome annotation. Our study addressed that perspective gap. Guided by earlier frameworks of information and DQ assessments, this study has defined empirically grounded models of DQ and DQ skills for genome annotation.

Literature Review

Data-quality issues in genome annotation work are actively discussed in the genomic literature. Errors in annotation can arise from the mismatches for sequence similarity searches, then propagate and amplify into discrepancies in specific descriptions of gene-function annotations (Devos & Valencia, 2001). Also reported in the literature are issues of genomic data and types of annotation activities, such as genomic context-based prediction (Kolesov, Mewes, & Frishman, 2001); of structure alignment and structure patterns (Shindyalov & Bourne, 1998); and of a single concept or particular step within the genome-annotation process, for example, homology-based transfer (Hsiang & Goodwin, 2003), genome properties/patterns (Emmersen, Rudd, Mewes, & Tetko, 2007), phylogenic considerations (Mikkelsen, Galagan, & Mesirov, 2004), expression microarray–based predictions (Kim & Falkow, 2003), and semantic variations in gene ontology annotations (Jones, Brown, & Baumann, 2007; MacMullen, 2006). Other studies related to DQ are found in the information-system literature, addressing genomic databases (Müller & Freytag, 2003), detection of errors in data, and the biocurator's role in DQ (Burkhardt, Schneider, & Ory, 2006). In addition, there have been community efforts to establish evaluation frameworks, standards and annotated test datasets for evaluating data extraction and mining software, including systems used for gene and protein name extraction from the literature and mining for association to existing database entries (e.g., BioCreAtIvE; Colosimo, Morgan, Yeh, & Colombe 2005). Finally, genomics communities regularly perform large scale data consistency checks to identify potentially erroneous annotations (McNeal et al, 2007).

Understanding the general concepts and relationships in DQ can help us define DQ in the genome-annotation context. There is a consensus that the quality of data or information is contextual and multidimensional, and must be evaluated relative to the context of its use (Strong, Lee, & Wang, 1997; Stvilia, Gasser, Twidale, & Smith, 2007). Indeed, Wang and Strong (1996) define quality as “fitness for use,” pointing to the importance of the context of use in determining an item’s quality. Likewise, Evans and Lindsay (2005) characterize quality as “user satisfaction” or “meeting or exceeding user expectation,” suggesting that the user’s perception and value structure for data characteristics play a critical role in the evaluation of DQ (Stvilia et al., 2009a; Wang, Pierce, Madnick, & Zwass, 2005). In the context of genome annotation, the community of genomics scientists –users, providers, and curators of genome annotations – determines what aspect makes an annotation higher or lower quality.

Data quality is a multidimensional concept. A DQ dimension, as defined by Wang and Strong (1996), is “a set of data quality attributes that represent a single aspect or construct of data quality.” That is, a DQ dimension is a conceptualization of measurable variations for a single aspect of DQ (Stvilia et al., 2007). Considerable research has addressed the definition of general taxonomies of DQ dimensions (Ge & Helfert, 2007; Stvilia, 2006; Wand & Wang, 1996). In addition, researchers have sought to define and operationalize data-quality dimensions specific to a particular domain, community, or document genre. Lankes (2008) reported the creditability shift from authority to reliability in an online community. Frické and Fallis (2004) defined a set of indicators for evaluating accuracy of consumer health information web pages. Stvilia, Mon, and Yi (2009) formulated a model for evaluating the quality of online consumer health information consisting of five constructs: Accuracy, Completeness, Authority, Usefulness, and Accessibility. Rieh (2002) defined a set of dimensions for evaluating the quality of scholarly information: Usefulness, Goodness, Currency, Accuracy, and Trustworthiness for the research scholar. MacMullen (2006) used five quality-assessment facets/dimensions (Consistency, Specificity, Completeness, Validity, and Reliability) to evaluate the curator’s gene-ontology annotation performance. Although the names and meanings of the quality of dimensions may persist across different domains (e.g., Accuracy, Completeness), their operationalizations and metrics may change with changes in context (Stvilia & Gasser, 2008; Stvilia et al., 2009). For example, the concept of completeness means the same in different contexts, but its operationalization will differ from one context to another. An intrinsic—i.e., context neutral—measurement of completeness can be defined in terms of all missing values, but as a contextual dimension, completeness can be measured in terms of missing values used or needed by a specific data user or needed for a particular activity (Lee, Pipino, Funk, & Wang, 2006; Stvilia, 2006).

Developing and applying a DQ-assessment model and analyzing results require a specific set of skills. Furthermore, assessment is only one task of DQ control or assurance. Other tasks may include DQ intervention and preventive DQ maintenance. Like DQ dimensions, DQ-assurance tasks and skills are context specific and are ultimately shaped by the processes of information production and use, the size of the organization or community, and the scale and complexity of the information product (Stvilia et al., 2008). In addition, like a DQ assessment model, DQ-assurance skills can be identified through conceptual analysis (e.g., task analysis) or by collection and analysis of empirical data – data collected through observations, interviews, and/or surveys. Chung, Fisher, and Wang (2002) surveyed DQ professionals to identify their understanding and priority ranking of DQ-assurance skills. They found that DQ professionals regarded interpretive capabilities as critical for understanding organizational implications of DQ. Adaptive capabilities that can identify user requirements and measure user DQ needs were also important (Chung et al., 2002). Lee and Strong (2003) discussed the impact of domain knowledge on DQ. Data collectors felt more strongly than data custodians that "knowing why" knowledge generated higher-quality data.

Research Questions

The primary focus of our exploratory study was to determine the perceptions of genome annotation DQ and skills needed to perform DQ assurance work by genomics scientists. In particular, the study addressed three research questions:

RQ1: What are the quality criteria considered to be important in genome annotation? This question is investigated by determining priority rankings and factor constructs of DQ dimensions.

RQ2: What are some of the DQ-assurance skills needed in genome annotation? This question is investigated by determining priority rankings and factor constructs in DQ skills.

Method

Genome annotation is a complex process that may involve multiple agents, playing different roles, and mediation of complex tools. To conceptualize the activity system of the genome annotation process, our study used a methodology consisting of activity theory and scenario based design previously proposed by Carroll (1997) and applied by Stvilia et al. (2007) in developing an information-quality assessment framework. In particular this study used activity theory (Leontiev, 1978; Nardi, 1996; Vygotsky, 1981) as the conceptual framework and a set of principles to reason about annotation and related quality-assurance activities, roles, expectations, and tool mediation occurring in the genome-annotation process (Figure 1). Activity theory can guide researchers in identifying relationships among different components of an annotation activity, activity-specific requirements for quality, and moderating effects of the activity's organizational, social, and cultural contexts.

[Insert Figure 1 here]

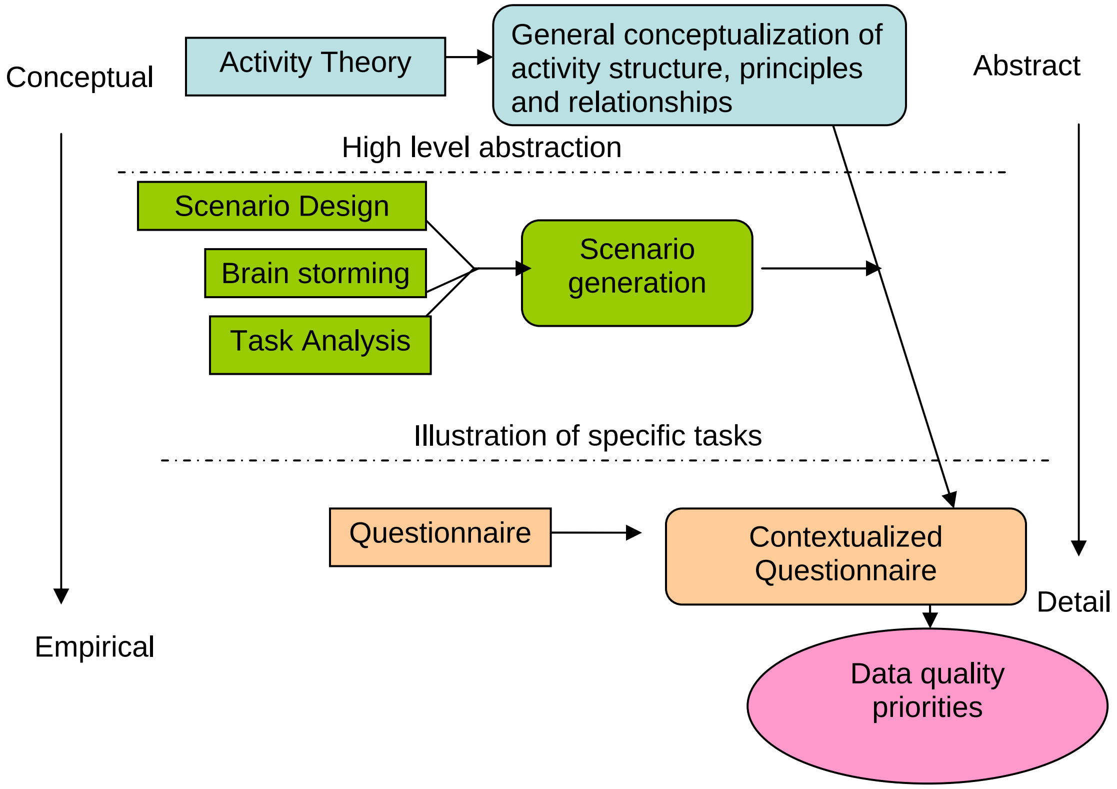

Figure 1. Study's methodology

In addition to the high-level conceptualization of activity structures and relationships, more detailed, activity 'instance' specific methods are needed for identifying and codifying requirements for genome annotation quality, and the expectations of quality assurance skills in genome annotation quality assurance work. The method of scenario based task analysis (Carroll, 1997; Go & Carroll, 2004) can be used to conceptualize specific requirements for quality and quality assurance tasks. This combines scenario-based system design and task analysis (Diaper, 2004) and encourages user involvement in process analysis to build shared understanding and knowledge of activities. Scenario-based task analysis serves as a useful tool for developing hypothetical stories for conceptualizing genome-annotation related activities (Figure 1). It allows to decompose a complex task into detailed lists of steps, procedures, and procedural descriptions conceptualized through rich, concrete, detailed scenarios describing creation and/or use of annotation artifacts and annotation outcomes.

The study used scenario based task analysis to develop two representative scenarios conceptualizing genome annotation related activities. The scenarios were used to develop a survey instrument of genome annotation quality perception and quality assurance skills by framing survey tasks and questions into the contexts that were meaningful and familiar to survey participants. The survey questions were adapted from DQ dimensions and skills requirement questions from previous data quality survey instruments found in the literature (Wang & Strong, 1996; Chung, et al., 2002; Lee et al., 2006) (see Figure 2).

The researchers obtained an institutional review board (IRB) approval for this study from the Florida State University's Human Subjects Committee in March, 2009. To develop the scenarios, the lead researcher interviewed two professors in plant genomics (one at Florida State University: FSU, and another at University of Florida: UF), and three postdoctoral researchers in bioinformatics and genomics (FSU). The five subjects were asked to provide detailed descriptions of the genome-annotation process and activities they performed as a part of their work and research. All subjects believed that genome annotation was a sequential (Kunin, Copeland, Lapidus, Mavromatis, & Hugenholtz, 2008) and multidimensional process (Reed, Famili, Thiele, & Palsson, 2006) based on the molecular-biology central dogma, which indicates the sequential order of genetic data flow: Deoxyribonucleic acid (DNA) → Ribonucleic acid (RNA) → protein (Crick, 1958). In addition to conducting interviews, we searched the literature for examples of genetic and bioinformatics tasks (Bartlett & Toms, 2005; Stevens, Gobe, Baker, & Brass, 2001) and standard genome-annotation-process pipelines (Samuel et al., 2008).

The first scenario conceptualized a functional-genomics sequence-annotation analysis, with one goal, and five sequential annotation actions. Each action consisted of preprocessing, structure-annotation, and functional-annotation operations. The task execution also involved peripheral use of other annotations or DQ-assurance tools (e.g., vector-trimming tools for elimination of ambiguous sequencing reads; see Table 1).

The second scenario conceptualized three goals for participants, with five specific curation tasks. The tasks were related to data-record creation, and data quality control. Task procedures required a complete range of genome-annotation approaches and use of DQ-assurance tools (see Table 1). Both scenarios included annotation and data quality assurance activities requiring intermediate knowledge of functional genomics and bioinformatics skills (see Table 1).

[Insert Table 1 here]

Table 1: Genome annotation scenarios.

Scenarios	**Data scales**
Scenario 1: **Production, annotation, and submission of Expressed Sequence Tags (ESTs) data** In this scenario, you will be a genetic database user, generating primary sequence data. For this purpose, you will process, 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.	Medium
Scenario 2: **Whole genome data curation in a model organism** In this scenario, you will be a genome data curator, generating 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. 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.	Large

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. *The phrase "sequence records" refers to both the primary DNA sequences themselves and all the associated annotations.	

The survey instrument was pilot-tested with eleven researchers (five professors at Arizona State University, FSU, and UF; two scientists at the U.S. Department of Agriculture (USDA); three postdoctoral researchers at the Noble Foundation (Ardmore, OK) and FSU, and one research staff member at Pfizer). The researchers asked the pilot test participants to read the survey, and comment on the validity and understandability of survey questions. The comments then were used to revise the questions, and optimize the survey as a whole. In addition, the pilot test suggested that adding pop-up windows with term glossaries to the instrument might be helpful for participants. Participants could use the glossaries to get the definitions of genetic and bioinformatics terms. In the final version of the survey, subjects could rate the quality criteria and skills on a seven point Likert scale, or select "unable to decide" and "not applicable" if they could not provide a judgment. In addition, , the survey included an open-ended question: "Do you have any comments or concerns (accuracy of terms, comprehensiveness, clarity of questions etc.) for this scenario and its question sets?" to allow participants to comment about the survey questions and scenarios (see Appendix 1).

The population for this study consisted of people who do genome-annotation work and conduct genomic research. To determine the survey population, the lead researcher searched the PubMed database (http://www.ncbi.nlm.nih.gov/pubmed/) with the following phrase " genome annotation" and the publication period limited to the 09/01/2006 to 09/01/2009. The search returned 1,504 articles. In a next step the researcher extracted 2,782 email addresses of the authors of those articles, and then randomly sampled 240 email addresses. Emails were to the sampled scientists in September 2009 to recruit them for survey participation. 158 scientists responded and completed the survey. Although compensation was offered, only 30% of survey participants accepted it, suggesting good buy-in to the goals of the research. The study used the Qualtrics software (http://www.qualtrics.com) to distribute the survey online, and collect data. The survey data were analyzed with SPSS software to produce descriptive statistics, factor analysis, and correlation reports and graphs.

Findings

Survey Participants' Characteristics

Survey participants self-identified their annotation roles as follows: users (n=93, 59%), curators (n=47, 30%), and dual roles (n=18, 11%). Over half of the participants (n=92, 58%) had a biology background, working in higher education in the U.S. or Canada, and holding a doctorate (see Table 2). Almost half the participants had more than five years of genome annotation experience.

[Insert Table 2 here]

Table 2. Demographics of survey participants (*n* = 158).

Demographic category	***n***
Annotation role	
User	93 (59%)
Curator	47 (30%)
Both	18 (11%)
Disciplines	
Biology	92 (58%)
Both	38 (24%)
Bioinformatics	28 (18%)
Residency	
U.S. and Canada	101 (64%)
Europe	35 (22%)
Asia	14 (9%)
South America	5 (3%)
Oceania	3 (2%)
Education level	
Ph.D.	128 (81%)
M.S.	30 (19%)
Years of annotation experience	
>5 years	70 (44%)
3–5 years	44 (28%)
1–2 years	28 (18%)
<1 year	16 (10%)
Organization	
University and higher education	114 (73%)
Government agency	16 (10%)
Nonprofit organization	13 (8%)
Industrial or private sectors	5 (3%)

Clinical practice	5 (3%)
Other	5 (3%)
Age (in years)	
<30	47 (30%)
30–39	56 (35%)
40–49	35 (22%)
50–59	19 (12%)
>60	1 (0.1%)
Gender	
Male	114 (72%)
Female	44 (28%)

RQ1: The Ranking of DQ Dimensions

To identify the domain specific perception of annotation quality, participants were asked to assess the importance of data quality dimensions (see Table 3), relative to the contexts of the first scenario (see Table 1). The descriptive statistics of the quality criteria rankings are given in Table 3. Mean, median, and standard deviation were calculated for each data quality dimension. On average, the participants ranked Accuracy as of the highest importance and Security the lowest, indicating that genome-annotations were expected to be highly accurate in an open-access environment (Ouyang et al., 2007).

[Insert Table 3 here]

Table 3. Survey participants' ranking of data-quality dimensions by mean importance.

Data-quality dimensions	***Num. of responses ****	***Mean***	***Median***	***Mode***	***Standard deviation***
Accuracy: Sequence records are correct and free of error	157	6.27	7	7	1.21
Believability: Sequence records are regarded as credible and believable	154	6.19	7	7	1.19
Accessibility: Sequence records are easily and quickly retrievable for access	157	6.02	6	7	1.22
Consistent representation: Sequence records are presented in a consistent format	156	5.77	6	7	1.26
Interpretability: Sequence records are in appropriate languages, symbols, and units, and the definitions are clear for interpretation	158	5.71	6	6	1.27
Completeness: Annotated sequence records are not missing and are fully annotated according to the steps described in this scenario.	156	5.67	6	7	1.31
Unbiased: Sequence records are unbiased and objective	154	5.56	6	6	1.35
Understandability: Sequence records are easily understandable	155	5.56	6	6	1.21
Ease of manipulation: Sequence records are easy to manipulate and make it easy to carry out various tasks described in this scenario	157	5.44	6	6	1.36
Traceability: The derivation history of the sequence records is documented and traceable	157	5.34	6	6	1.4
Up-to-date: Sequence records are sufficiently up-to-date for this scenario	154	5.34	6	6	1.39
Appropriate amount of information: The volume of the sequence records is appropriate for this scenario	151	5.19	5	6	1.36
Relevancy: Sequence records contain information relevant to the scenario	153	5.14	5	6	1.28
Value added: Sequence records contain additional annotations from the tasks in this scenario and these annotations are beneficial and add value	155	5.1	5	6	1.35
Reputation: Sequence records are highly regarded and reputable in terms of their source or content	155	4.98	5	5	1.42
Concise representation: Sequence records are concisely represented	155	4.82	5	5	1.51
Security: Access to sequence records is restricted appropriately to maintain their security	158	3.78	4	5	1.76

*Responses of "not applicable" were not included in the statistical analysis.

Factor Constructs for DQ Dimensions

To identify the value structure for quality reflected by the 158 survey respondents' rankings, the study conducted exploratory factor analysis using principal-components analysis as the extraction method and varimax with Kaiser normalization as the rotation method (see Table 4). Since the sample size was n=158, the cutoff size for criterion loadings was set to 0.45 (Hair, 2005). Both the Bartlett ($\chi^2 = 702.21$, $p < 0.001$) and measure of sampling adequacy (MSA = 0.789) tests for the sample pointed to a significant level of correlation among the dimensions. A scree-plot analysis suggested selecting the first five components for DQ dimension constructs.

[Insert Table 4 here]

Table 4. Factor loadings for the data-quality dimensions. Principal-components analysis served as the extraction method and varimax with Kaiser normalization as the rotation method. Values above the cutoff size for criterion loadings (0.45) are marked with asterisks.

	Components				
Data-quality dimensions	***1***	***2***	***3***	***4***	***5***
Accessibility	0.21	0.03	0.04	**0.82***	0.15
Accuracy	0.28	0.09	**0.75***	0.02	–0.03
Appropriate amount of information	0.24	**0.49***	–0.10	**0.52***	–0.04
Believability	0.02	0.04	**0.54***	**0.67***	–0.02
Completeness	0.14	0.42	0.20	0.36	–0.44
Concise representation	0.21	**0.69***	0.02	0.11	–0.20
Consistent representation	**0.62***	0.14	0.22	0.25	–0.08
Ease of manipulation	**0.73***	0.01	0.08	0.18	–0.18
Interpretability	**0.76***	0.15	0.21	0.14	–0.01
Relevance	0.11	**0.76***	0.24	0.05	0.10
Reputation	0.20	**0.49***	0.28	0.23	0.20
Security	–0.13	0.41	–0.09	0.21	**0.65***
Traceability	**0.49***	–0.02	0.35	0.08	**0.56***
Unbiased	0.08	0.35	**0.71***	0.12	0.02
Understandability	**0.75***	0.22	0.06	0.01	0.20
Up to date	0.09	**0.67***	0.17	–0.03	0.18
Value added	**0.62***	**0.49***	–0.07	–0.09	0.10

The dimensions loaded on the first factor construct could be categorized as those of Usefulness (see Table 5). The second construct was related mainly to Relevance. The third included both Accuracy and trust-related dimensions; Accuracy had the highest loading. The fourth construct was mostly heavily loaded with Accessibility. The Believability dimension loaded on both the third and fourth constructs, although semantically it is closer to accuracy than to accessibility. The criteria loaded on the fifth construct could be categorized as related to Security. The constructs were then ranked by the arithmetic averages of the mean ratings of the individual dimensions loaded on the constructs (see Table 5). The Accuracy construct was ranked the highest, followed by the Accessibility, Usefulness, Relevance, and Security constructs.

[Insert Table 5 here]

Table 5. The five data-quality constructs generated from the data-quality dimensions, and ranked by the arithmetic averages of the mean ratings of the individual dimensions loaded on the constructs.

Data-quality constructs	***Ranking***	***Dimensions***
Accuracy	6.01	Accuracy, Unbiased, Believability, Traceability
Accessibility	5.80	Accessibility, Believability, Appropriate amount of information
Usefulness	5.52	Interpretability, Understandability, Ease of Manipulation, Consistent representation, Value added
Relevance	5.08	Relevant, Concise representation, Up-to-date, Reputation, Value added
Security	4.56	Security, Traceability

RQ2: The Ranking of DQ Skills

In the second part of the survey the participants were asked to rank a set of data-quality skills on the basis of their importance in genome-annotation work in the context of the second scenario described in Table 1. The descriptive statistical summary of the results (see Table 6) revealed that the Data-error-detection skill was ranked the highest (6.04) and the Data-quality-cost/benefit skill the lowest (4.84). This outcome implied that the genome-annotation communities are concerned mainly with sufficient skills to detect annotation errors.

[Insert Table 6 here]

Table 6. Survey participants' ranking of data-quality skills by mean importance in the context of Scenario 2.

Data-quality skills	Num. of responses*	Mean	Median	Mode	Standard deviation
Data error detection: Ability to detect and correct errors in databases	149	6.04	6	7	1.26
Data mining skills: Data mining and knowledge discovery skills for analyzing data in a data warehouse	148	5.95	6	6	1.06
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.	140	5.81	6	7	1.16
Data quality implication: Understanding pervasiveness of data quality problems and their potential impacts	143	5.73	6	6	1.04
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)	141	5.6	6	6	1.21
Data quality audit: Ability to conduct data quality auditing (formal review, examination, and verification of data quality)	145	5.57	6	6	1.35
Statistical techniques: Ability to apply statistical techniques to manage and control data quality	148	5.54	6	6	1.32
Data entry improvement: Skills and ability to analyze and improve the data entry process in order to maintain data quality	144	5.51	6	6	1.22
Software tools: Experience and ability to use diverse commercially available data quality software packages	145	5.5	6	6	1.32
Organization policies: Ability to establish and maintain organizational policies and rules for data quality management	143	5.38	6	5	1.42
Data warehouse setup: Ability to integrate multiple databases into an integrated data warehouse	147	5.34	6	5	1.39
User requirement: Ability to translate subjective user requirements for data quality into objective technical specification (such as use of Quality Function Deployment)	139	5.25	5	6	1.33
Analytic models: Ability to apply diverse analytic models (such as regression model and multidimensional model) for data analysis	144	5.21	5	5	1.44
Change process: Ability to manage the change process/transitions resulting from the data quality management project	141	5.2	5	6	1.35
Structural Query Language (SQL): Skills and ability to apply SQL to estimate the accuracy of data	136	5.12	5	6	1.42
Information overload: Understanding the information overload that managers often face and ability to reduce information overload	144	4.94	5	5	1.3

Data quality cost/benefit: Skills and ability to conduct cost/benefit analysis of data quality management	144	4.84	5	6	1.29

*Responses of "not applicable" were not included in the statistical analysis.

Factor Constructs for Data-Quality Skills

Data collected from the survey were explored by factor analysis for underlying constructs for quality skills. Both the Bartlett ($\chi^2 = 790.8$, $p < 0.001$) and MSA (0.842) tests for the sample pointed to a significant level of correlation among the criteria. The scree plot suggested selecting the first four components for DQ-skills constructs. The first construct represents technical skills required for high-quality genome annotation, including the Statistical techniques skill (see Table 7 for details). Construct two represents Adaptive skills and includes Data-quality cost/benefit analysis. The third-construct loadings could be categorized as Interpretive skills, including Data-error detection. Factor four represents DQ literacy skills, including DQ dimensions.

[Insert Table 7 here]

Table 7. Factor loadings for the data quality skills. Principal-components analysis served as the extraction method and varimax with Kaiser normalization as the rotation method. Values above the cutoff size for criterion loadings (0.45) are marked with asterisks.

	Components			
Data-quality skills	***1***	***2***	***3***	***4***
Data-quality dimensions	0.20	0.37	**0.62***	0.15
Data-quality measurement	–0.04	0.10	**0.73***	0.18
Data-quality implication	0.12	0.15	**0.85***	–0.01
Data-entry improvement	0.37	**0.60***	0.19	0.17
Organization policies	0.23	0.71	0.26	–0.05
Data-error detection	0.28	0.05	0.36	**0.60***
Change process	–0.03	**0.58***	0.33	**0.55***
Data-quality cost/benefit	–0.11	**0.82***	0.10	0.12
User requirement	0.41	**0.52***	0.12	0.17
Information overload	0.38	**0.69***	0.07	0.02
Data-quality audit	**0.70***	0.12	0.25	0.10
Statistical techniques	**0.76***	0.14	0.25	–0.01
Data-mining skills	**0.66***	–0.03	0.08	0.16
Data warehouse set-up	**0.66***	0.37	–0.24	0.14
Analytic models	**0.59***	0.30	–0.08	0.33
Structured query language	**0.66***	0.29	–0.05	0.39
Software tools	0.27	0.08	0.04	**0.83***

The four constructs were then ranked by the arithmetic averages of the mean ratings of the individual skills loaded on the constructs (Table 8). Data-quality literacy skills ranked the highest, and Adaptive skills lowest. The results indicated that DQ literacy skills are of the highest priority to genome annotation users; they value high basic training on the definition of data quality (Data-quality dimension), how to assess data quality (Data-quality measurement), and DQ impact or outcomes (Data-quality implication). The Change process skill can be regarded as an Interpretive or Adaptive skill in a highly dynamic environment, especially when the genome-annotation process comprises numerous complex and contextual annotation tasks.

[Insert Table 8 here]

Table 8. The four skills constructs generated from the data-quality skills listed in Table 7, ranked by the arithmetic averages of the mean ratings of the individual skills loaded on the constructs.

Data-quality skills construct	*Ranking*	*Data-quality skills*
Data-quality literacy skills	5.71	Data-quality dimension, Data-quality measurement, Data-quality implication
Interpretive skills	5.58	Data-error detection, Software tools, Change process
Technical skills	5.46	Data-quality audit, Statistical techniques, Data-mining skills, Data warehouse set-up, Analytical models, Structured query language
Adaptive skills	5.26	Data-entry improvement, Organization policy, Change process, Data-quality cost/benefit, User requirement, Information overload

Discussion

The first research question focused on genomics scientists perception of annotation data quality. The survey participants ranked both the Accuracy dimension and the accuracy construct the highest; this matches findings from earlier studies of different user communities, such as the study of data-quality perception of online-health-information consumers by Stvilia, Mon, and Yi (2009.), However, is study's participants ranking of Accessibility second is in disagreement with the quality model of that study. This incongruity can be attributed to differences between the types of information objects used by these two communities. Online consumer-health information usually consists of web pages created for the purpose of information sharing, often with accessibility in mind (e.g., language tailored to the general audience). Genome annotations, conversely, are scientific metadata, which may not be always created for global sharing. Like any other metadata, genome annotations can be incomplete, may not include locally shared knowledge, and may not follow the community's standards and conventions.

The genome annotation DQ model identified shared similarities with that of Wang and Strong (see Figure 2), even though the factor labels differ. Wang and Strong's (1996) framework was developed from the rankings of quality dimensions by survey participants selected mainly from industry and a business school. Interestingly, in the genome-annotation-quality model, the Concise presentation dimension fell into the Relevance construct rather than the Useful Information/Representational DQ construct as it did in Wang and Strong's model (Figure 3). This suggests that the genome-annotation community may recognize the contextuality of the concept of "concise representation." A concise presentation for one task can be an incomplete or a lengthy presentation for another task.

[Insert Figure 2 here]

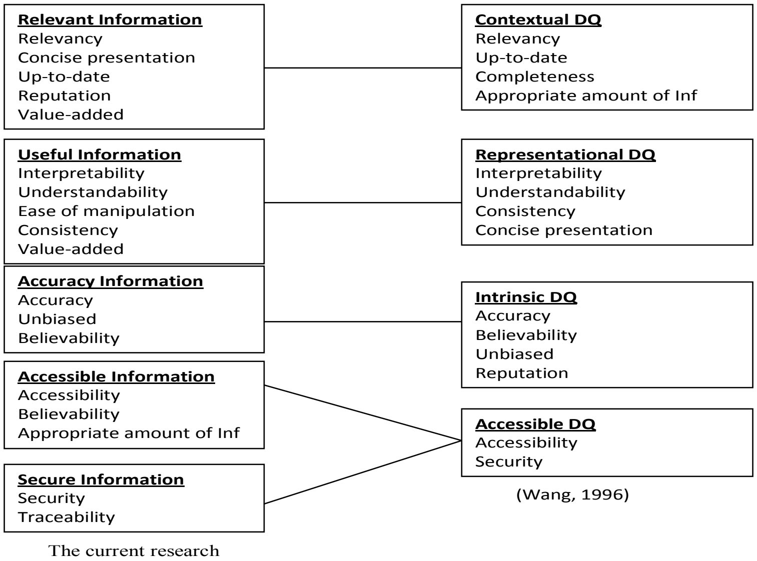

Figure 2. Comparison of our data-quality (DQ) constructs to those of Wang and Strong (1996).

The Relevance factor of the genome-annotation quality model also included the Reputation dimension, whereas Wang and Strong's model placed Reputation in the Intrinsic factor (Figure 2). This difference suggests that the Reputation concept might be more contextually sensitive for the genome-annotation community than for other groups and need reevaluation as task context changes.

In our study, the Value-added dimension occurred in both the Relevance and Usefulness constructs. In the Relevance construct, the Value-added dimension is primarily concerned with whether data fits the task's needs. It may also evaluate the value or utility of the annotation to the task. Hence, its grouping with these two factors is not surprising.

The Accuracy construct combined the Accuracy dimension itself and the trust-related dimensions of Unbiased and Believability and matches the structure of the Intrinsic factor of Wang and Strong's model. The genome annotation users' high rankings of the Accuracy and trust-related quality dimensions is also in agreement with the participants' perception of DQ from Wang and Strong's study.

Security is a separate construct in our model and is not a part of the Accessible construct as it was in Wang and Strong's model. This could be an indicator of the genome community's openness to data sharing, and the community not associating security with data accessibility. Traceability improves accuracy of data by providing verification methods. From the security and data assurance perspective, Traceability increases information integrity by maintaining provenance information at every step in the genome-annotation process.

Overall, our findings indicate that high-quality data should be accurate, contextually appropriate to the annotation task, and useful and accessible to genome-annotation users.

The second research question examined genomics scientists' perception of the skills needed to ensure the quality of genome annotations. A comparison of the four DQ-skills constructs identified by this study with the model developed by Chung et al. (2002), shown in Figure 3, reveals that the Data-quality literacy skills construct does not have a matching construct in the Chung et al. model. This construct includes the skills needed to understand the basic concepts of DQ, such as DQ dimensions, measurements, and their implications (see Figure 3). The findings suggest that the genome-annotation community assigns a high value to DQ literacy and the skills needed to improve annotation quality. One reason that the Chung et al. model does not include the literacy construct could be that DQ professionals, who were the subjects of their study, were expected to know basic DQ concepts, and, therefore might not find DQ literacy skills as important as genomics scientists who might not be trained in DQ assurance.

[Insert Figure 3 here]

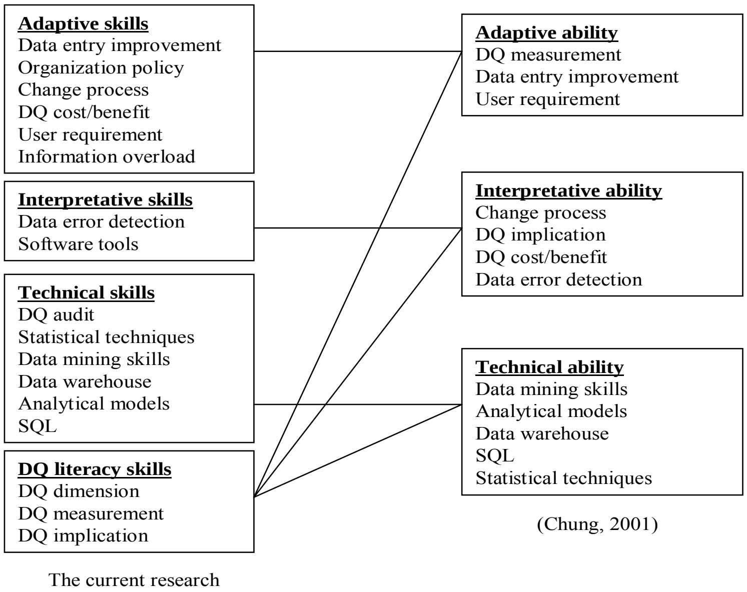

Figure 3. Comparison of our DQ-skills constructs to those of Chung et al. (2001).

Genome-annotation is an ongoing process, and genome annotations are dynamic objects, which may require regular updates, expansion, and consistency checks as new knowledge, annotation data, and new uses for these data emerge. Genomics scientists may need adaptive skills to deal with context-related aspects of DQ to determine whether the annotation data are valuable and relevant to a particular context. For example, the skills related to Information overloading help scientists to distill important information from large-scale sets of data. Technical skills, including Data mining skills, are helpful in addressing intrinsic DQ problems such as the accuracy and consistency of annotation data.

Unlike those of Chung et al., this study placed DQ cost/benefit skills in the Adaptive skills rather than in the Interpretative skills construct. Data-quality-assurance activity is highly domain specific and context sensitive. At each annotation step, users have to determine the usefulness of the annotation: whether it has an expected and perceived value for contemporary or future use in that particular context. Scientists apply cost/benefit analysis to strike a balance between the cost and benefits of annotations. Data-quality cost/benefit analysis may serve as a tool for evaluating the cost effectiveness of an annotation strategy or the process itself and then recalibrating or adapting the process on the basis of that analysis.

The DQ skills model identified by this study placed the Software Tools in Interpretive skills construct. The Software tools is absent in the Chung et al model. This could be explained by the highly technical nature of genome-annotation work. The Software Tools was placed with the Interpretative skills. The Software Tools covers not only data mining tools but also tools used for interpreting annotations. The Data quality audit was grouped within the technical skills construct. Despite several attempts to establish task and software type specific evaluation frameworks (e.g., BioCreAtIvE), genome annotation work has not matured yet and may lack concrete standards and procedures for formal review and verification of genome annotation DQ. To generate and interpret the annotation data, annotation users apply multiple software tools to align, compare, and interpret genomic data. Thousands of bioinformatics-related software tools are available in the public domain. For example, one of the popular bioinformatics resource sites, Bioinformatics.org, lists more than 300 such tools. Since so much genomic data are available (e.g., over 3 billion DNA base pairs for the human genome), knowledge of appropriate annotation software can help accelerate the annotation process and create useful annotation products.

Conclusion

The study identified the genome-annotation community's perception of DQ and DQ skills by extracting the factor constructs of genome-annotation quality dimensions and skills from survey data. To the best of our knowledge, this is the first study that examines both DQ skills to DQ dimensions and problems in genome annotation. The quality constructs defined by this study advances the understanding of DQ relationships and is an important contribution to data and information-quality research. In addition, the resulting models can be a valuable resource to genome data curators and administrators for developing data-curation policies and designing DQ-assurance strategies, processes, procedures, and infrastructure. The study's findings may also inform educators in developing data quality assurance curricula and training courses.

The study is not without limitations. The rankings of DQ dimensions and skills used to develop the genome-annotation-quality models represent the participants' perceptions of the concepts' importance. These are only approximations of the actual value models of quality and the skills used in practice. 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.

To identify the rankings of the DQ dimensions and skills constructs, the study used the arithmetic average of the mean ratings of the individual dimensions or skills loaded on the factor. Future research may explore the use of different metrics to calculate the factor rankings. Also, data-quality dimensions and skills are only one segment of the overall quality-assurance process structure. The other components may include metrics, roles, tools, and procedures. Future research related to this study may include developing the registries of these components along with data-type-specific templates for quality requirements and relationships. Future research may also include exploring data quality needs in other activities of genomic research, other than genome annotation work.

Finally, Genomics scientists can play more than one role in annotation work. They can be both users of and providers of genome annotations in the context of a community or institutional database, as well as curators of their own, or community genome data. One may theorize that genomics scientists who play a curator role on a regular basis for the community's genome information database can be more acutely aware and knowledgeable of the importance of data quality control, and as well the skills needed to annotate genome data, or ensure the quality of annotations submitted by different providers. The objective of this study was to identify general, domain level, models of quality and quality assurance skills needed in genome annotation work. The scope of the current study did not include identifying role specific models of annotation quality and quality assurance skills. It would be interesting, however, as a part of future research related to the current study, to explore how curators' perception of genome annotation quality and quality assurance skills needed may differ from of those genomics scientists who do not play that role on a regular basis.

Acknowledgements

The authors would like to express their gratitude to Paul Marty, Michelle Kazmer, and Larry Dennis, the College of Communication and Information, at the Florida State University, for insightful discussions and comments for this research. The authors also want to thanks for three anonymous reviewers' insightful comments for this study.

References

Bartlett, J. C., & Toms, E. G. (2005). Developing a protocol for bioinformatics analysis: An integrated information behavior and task analysis approach: Research Articles. *Journal of the American Society of Information Science and Technology, 56*(5), 469–482.

BioCreAtIvE. http://www.biocreative.org/about/background/description/

Bragge, J., Merisalo-Rantanen, H., & Hallikainen, P. (2005). Gathering innovative end-user feedback for continuous development of information systems: A repeatable and transferable e-collaboration process. *IEEE Transactions on Professional Communication, 48*(1), 55–67.

Brenner, S. (1999). Errors in genome annotation. *Trends in Genetics, 15*, 132–133.

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

Carroll, J.M. (1997). Scenario-Based Design. In M. Helander & T.K. Landauer, (Eds.) *Handbook of Human-Computer Interaction, second edition*, (pp. 383-406), Amsterdam: North Holland.

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*, Boston, MA, USA.

Colosimo, M.E., Morgan, A.A., Yeh, A.S., & Colombe J.B. (2005). Data preparation and interannotator agreement: BioCreAtIvE. *BMC Bioinformatics, 6*, S12.

Crick, H.F. (1958). One protein synthesis. *Symposia of the Society for Experimental Biology. 12*, 138–163.

Curry, E., Freitas, A., & O'Riáin, S. (2010). The role of community-driven data curation for enterprises. In D. Wood (Ed.), *Linking enterprise data* (pp. 25-47). New York, NY: Springer.

Devos, D., & Valencia, A. (2001). Intrinsic errors in genome annotation. *Trends in Genetics, 17*, 429–431.

Diaper, D. (2004). Understanding task analysis in human computer interaction. In D. Diaper & N. Stanton (Eds.), *The handbook of task analysis for human-computer interaction* (pp. 117-133). Mahwah, NJ: Lawrence Erlbaum Associates.

Emmersen, J., Rudd, S., Mewes, H., & Tetko, I. (2007). Separation of sequences from host-pathogen interface using triplet nucleotide frequencies. *Fungal Genetics and Biology, 44*(4), 231–241.

Evans, J. R., & Lindsay, W.M. (2005). *The management and control of quality (6th ed.).* Cincinnati, OH: South-Western, Thomson Learning.

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.

Frohlich, H., Speer, N., Poustka, A., & Beissbarth, T. (2007). GOSim—an R-package for computation of information theoretic GO similarities between terms and gene products. *BMC Bioinformatics, 8*, 166.

Ge, M., & Helfert, M. (2007). A review of information quality research—develop a research agenda. Paper presented at *the International Conference on Information Quality 2007*, 76–91.

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: Lawrence Erlbaum Associates.

Hair, J., Black, B., Babin, B., Anderson, R., & Tatham, R. (2005). *Multivariate data analysis*. Upper Saddle River, NJ: Prentice-Hall.

Hermann, D. (2007). Are you a good data steward? *Healthcare Purchasing News, 31*(1), 58.

Hsiang, T., & Goodwin, P. (2003). Distinguishing plant and fungal sequences in ESTs from infected plant tissues. *Journal of Microbiological Methods, 54*, 339–351.

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.

Jones, C., Brown, A., & Baumann, U. (2007). Estimating the annotation error rate of curated GO database sequence annotations. *BMC Bioinformatics, 8*, 170–174.

Kim, C., & Falkow, S. (2003). Significance analysis of lexical bias in microarray data. *BMC Bioinformatics, 4*,12–15.

Kolesov, G., Mewes, H., & Frishman, D. (2001). SNAPping up functionally related genes based on context information: a colinearity-free approach. *Journal of Molecular Biology, 311*, 639–656.

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., & Hugenholtz, P. (2008). A bioinformatician's guide to metagenomics. *Microbiology and Molecular Biology Reviews, 72*(4), 557–578.

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

Lord, P., & Macdonald, A. (2003). *E-Science curation report: Data curation for e-Science in the UK: An audit to establish requirements for future curation and provision*. Bristol, UK: The JISC Committee for the Support of Research. Retrieved from http://www.jisc.ac.uk/media/documents/programmes/preservation/e-sciencereportfinal.pdf

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

Lee, Y., Pipino, L., Funk, J., & Wang, R. (2006). *Journey to data quality*. Cambridge, MA: MIT Press.

Leontiev, A. (1978). *Activity, consciousness, and personality*. Englewood Cliffs, NJ: Prentice Hall.

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

Marco, D. (2006). Understanding data governance and stewardship, Part 2. *DM Review, 16*(10), 17.

McGilvray, D. (2008). *Executing data quality projects: ten steps to quality data and trusted information.* San Francisco, CA: Morgan Kaufmann.

McGuire, A., Diaz, C., Wang, T., & Hilsenbeck, S. (2009). Social networkers' attitudes towards direct-to-consumer personal genome testing. *American Journal of Bioethics, 9*(6–7), 3–10.

McNeil, L. et al (2007). The national microbial pathogen database resource (NMPDR): A genomics platform based on subsystem annotation. *Nucl. Acids Res. 35*(suppl 1): D347-D353. Retrieved 19 June, 2011, from doi:10.1093/nar/gkl947

Mikkelsen, T., Galagan, J., & Mesirov, J. (2005). Improving genome annotations using phylogenetic profile anomaly detection. *Bioinformatics, 21*(4), 464–470.

Mons, B., Ashburner, M., Chichester, C., van Mulligen, E., Weeber, M., den Dunnen, J., van Ommen, G.-J., Musen, M., Cockerill, M., Hermjakob, H., Mons, A., Packer, A., Pacheco, R., Lewis, S., Berkeley, A., Melton, W., Barris, N., Wales, J., Meijssen, G., Moeller, E., Roes, P., Borner, K., & Bairoch. A. (2008). Calling on a million minds for community annotation in WikiProteins. *Genome Biology,* 9, R89.

Müller, F., & Freytag, J. (2003). Data quality in genome databases. In *Proceedings of the Eight International Conference on Information Quality (ICIQ03).* Cambridge, MA: MIT.

Nardi, B. (1996). Activity theory and human-computer interaction. In B. A. Nardi (Ed.), *Context and consciousness: Activity theory and human-computer interaction* (pp. 69-103). Cambridge and London: MIT Press.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F,, Malek, R., Lee, Y., Zheng, L., Orvis, J., Haas, B., Wortman, J., Buell, C. (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Research*, *35*, D883–D887.

Pruitt, K., Tatusova, T., & Maglott, D. (2007). NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research, 35m* D61–65.

Reed, J., Famili, I., Thiele, I., & Palsson, B. (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.

Salzberg S. (2007). Genome re-annotation: A wiki solution? *Genome Biology 8*, 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.

Schlueter, S. D., Wilkerson, M. D., Huala, E., Rhee, S. Y. & Brendel, V. (2005). Community-based gene structure annotation. *Trends in Plant Science, 10*(1), 9–14.

Shindyalov, I., & Bourne P. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering, 11*(9), 739–747.

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

Stein (2004). [Please complete this entry]Stevens, R., Gobe, C., Baker, P., & Brass, A. (2001). A classification of tasks in bioinformatics. *Bioinformatics, 17*, 180–188.

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

Stvilia, B. (2006). *Measuring information quality*. Doctoral thesis, University of Illinois at Urbana-Champaign, Urbana.

Stvilia, B. (2007). A model for ontology quality evaluation. *First Monday, 12*(12).

Stvilia, B. (2008). A workbench for information quality evaluation. In R. Larsen, A. Paepcke, J. Borbinha, & M. Naaman (Eds.), *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries. JCDL '08* (p. 469). New York, NY: ACM.

Stvilia, B., & Gasser, L. (2008). *An activity theoretic model for information quality change. First Monday, 13*(4).

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

Stvilia, B., Twidale, M., Smith, L. C., & Gasser, L. (2008). Information quality work organization in Wikipedia. *Journal of American Society of Information Science and Technology, 59*(6), 983–1001.

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

Vygotsky, L. (1981). The development of higher forms of attention in childhood. In J. V. Wertsch (Ed.), *The concept of activity in Soviet psychology*. Armonk, NY: Sharpe.

Wand, Y., & Wang, R. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM, 39*(11), 86–92.

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

Wang, R., Pierce, E., Madnick, S., & Zwass, V. (2005). *Information quality (advances in management information systems).* Armonk, NY: Sharpe.