This is a preprint of an article accepted for publication in *IBM Journal of Research and Development*. 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),(DOI: 10.1147/JRD.2011.2172837).

Using IBM Content Manager for Genomic Data Annotation and Quality Assurance Tasks

Hong Huang[1], Jiang Lu [2*] , Wayne B. Hunter[3], Shuang Liang[4]

[1.] School of Information, University of South Florida, Tampa, FL, 33620, U.S.A.
[2.] Center for Viticulture and Small Fruit Research, Florida A&M University, Tallahassee, FL, 32308, U.S.A.
[3.] United States Department of Agriculture, Agriculture Research Service, United States Horticultural Research Laboratory, 2001 South Rock Road, Fort Pierce, FL, 34945, U.S.A.
[4.] Institute of Genetic Engineering, Southern Medical University, Guangzhou, China, 510515.
* Correspondent Author

Email: Hong Huang: honghuang@usf.edu
Jiang Lu: jiang.lu@famu.edu
Wayne Hunter: wayne.hunter@ars.usda.gov
Shuang Liang: itshuang@gmail.com

This is a preprint of an article accepted for publication in *IBM Journal of Research and Development*. 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),(DOI: 10.1147/JRD.2011.2172837).

Abstract

As the amount of massive and heterogeneous genomic data and the relative annotations continue to grow, A flexible and easy-to-access data management solution is required to integrate the heterogeneous genomics data that can accommodate the needs of diverse annotation tasks. This research expounds the benefits of using IBM DB2 Content Manager (CM) Software by conducting task-oriented grape genome annotations, and data quality assurance checks throughout the annotation process. To demonstrate the usability of this application, we describe the implementation of two real-life content-based, genome annotation case scenarios: 1) Expressed Sequence Tags annotation and 2) sequence annotation related to Simple Sequence Repeats (SSR) markers. The IBM DB2 CM allowed users to easily construct content-based genomic information applications as rapidly built and readily adapted customized content documents with attributes within an easy-to-use interface system. Users can simultaneously conduct the annotation quality checks while making annotations by utilizing a built-in, standardized data quality control assurance procedure [referred to as] annotation "routing." The system provides search features or cross-links with different annotation contents or data formats. The data quality workflow and procedure within the system also resulted in accuracy and consistency in the data annotation and curation lifecycle.

This is a preprint of an article accepted for publication in *IBM Journal of Research and Development*. 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),(DOI: 10.1147/JRD.2011.2172837).

Introduction

With the availability of advanced and affordable sequencing technology, tremendous amounts of genomic DNA (deoxyribonucleic acid), RNA (ribonucleic acid) sequences [1], Expressed Sequence Tags (ESTs) [2], Single Nucleotide Polymorphisms (SNPs) [3], microarray data profiles [4], and sequence tagged sites (STS) [5] that integrate genomic markers and genetic maps have been deposited in the public domain. For example, over one million nucleotide sequences data for over twenty grape species is freely available from the National Center of Biotechnology Institute (NCBI). Thus, the grape genomic research community has a strong need for a flexible data management solution that easily and coherently integrates, preserves, updates, and reuses genetic/genomic data and annotation products in various formats (e.g., documents, images, and other multimedia files). We defined "genome annotation" as the practice of attaching biological information to the genomic sequence by the users [6,7].

Previous studies indicate that Rational Database Management Systems (RDMS) such as IBM DB2 [8,9, 10], Oracle [11], and mySQL [12, 13] are applicable for biological data integration, but these solutions only focus on the relational organization of data, which does not correlate data based on content for efficient genome annotation document workflow management. Genomics information, which has traditionally been delivered in a text-based format dealing with genetic code, has become increasingly dependent on a large array of multimedia data. This is largely due to the advancements made in technologies related to gene chips, image processing, and genome-wide association analysis. Users interact with data at a file or document level in a variety of formats (relational tables, Extensible Markup Language: XML, documents, images, videos, and sounds). Since individual users conduct specific lab experiments that create unique genomic datasets for management, a data system must be highly flexible and adaptable for data sets produced by genome annotation work. Data administrators can then collaborate with different users to collectively organize these datasets to provide better service to the genomics research community.

Most genomic databases focus on specific biological functions or genomic data characteristics that applied to many species [14, 15, 16]. With the increased quantity and complexity of genomic data, annotation work became collaborative and required multiple users' input [17, 18, 19, 20]. Increased community-based genome annotation for multiple species was done by using flexible, interactive tools to manually curate, analyze and modify annotation products. Users in these communities had different annotation experiences and training backgrounds [21]. The genome annotation products, features, and attributes also required convenient, standardized pipelines and workflow for manual curation, as well as mechanisms for consistency and accuracy data quality checking for annotation records/documents [22].

Specialized genomic databases tailored to the data structures and attributes of a particular annotation product allow users to conduct specific genome annotation tasks and to link the heterogeneous genomic data sources together. IBM DB2 Content manager (CM) software is an Enterprise Content Management (ECM) repository [23]. ECM and its applications have been extensively reported in the literature [23, 24], and applied as

simplified functional tools for information exchanges in industry [25, 26] and academia[27]. EMC was found to be useful in certain research labs and institutes that had enormous amount of genomic data to deal with [28]. Particularly, the EMC enterprise architecture provided a scalable and flexible solution that met these needs [29]. In addition, EMC provided useful support throughout the life cycle of digital information [30], from creation of genomic annotation documents through review and approval, to a period of retention, and ultimately to re-use in new and unexpected contexts.

IBM DB2 CM can benefit genome research community in several ways. Firstly, it offers a flexible analysis and adaptive data management system when the changing experiments and ideas are continually developed within a research institution (or life science company). Secondly, the system can support community based annotations by defining different annotation role-based users groups and their task-related annotation documents when applying privilege settings, and access/control lists. Thirdly, the IBM DB2 CM build-in interfaces are easily to manipulate, so that either database administrators or desktop users can easily build up annotation task-oriented annotation modules. Moreover, the system provides annotation document routings for data quality assurance checking. This paper reports the implementations of real case scenarios of grape genomic resources data annotation by characterizing the system architecture and data modeling of IBM DB2 CM. It also demonstrates the enhanced features of easy-to-use accesses and controls toward effective community annotation and quality intervention procedures for reliable genome annotation and curation process managements.

Construction and Content

System Architecture

The IBM DB2 CM was installed on top of a secured Linux Suse 9.2 operation system, with a hardware setting of a x346 dual 3.2 GHz processor, 5GB RAM, 438 GB raw capacity with 330 GB usable storage array. The CM was based on the modularized triangular information architecture, which is robust to standardize dynamic data communication and integration, which contains a “Library Server”, a “Resource Manager,” and “Content Manager Clients” for three-way communication. The program uses a standard relational library server (LS), which is the central source, to store the meta-data for the content but applies specialized resource managers (RMs) to actually optimize management of the scalable content. The Resource Manager and eClient were configured to provide unified data storage/retrieval and importation/exportation that can eventually include features for automatic data processing without the many manual steps currently needed. During the three-way communications, the client sends an SQL request to the library server, then the library server processes the request by looking up the content index, then provides the client a security token and a locator pointing to the information in resource manager. The client then can use the token and locator to access the resource manager. When the resource manager accepts the token, it returns the information to client[31].This three-way communication architecture enables a dynamic workflow process and interoperable framework including the creation (RMs), storage

(LS), and deployment (eClient) of content knowledge collected from the heterogeneous genomic data.

Content life cycling data workflow

The IBM DB2 CM architecture provides scalable data curation support including data aggregations, both at the local (a specific set of experimental data) and global (links with various data sources) levels. The CM gives a set of processes that support the "evolutional life cycle of digital information" [29]. Users can dynamically conduct data recycling tasks including selection of data sources, creation of data models, importation/exportation of standardized data in various formats (e.g., text. or image files), and evaluation and allocation of annotation records, as well as transformation of the annotation product with updated genetic information for future reuse and preservation (Figure 1). The annotated genomic data is put into the document routing queue for data quality assurance. By doing so, users experience a standardized, consistent process of integrating various genomic resources via the life cycling management function.

The IBM DB2 CM can internally build up a data cycling process (Figure 1) with a series of steps defined by users for document routing management. In this research, the document contains the genomic data and the relative annotations. The document routing process can provide a very good information quality control method, specifically when there is a need for document routing management to allow user groups to evaluate the content of annotation and identify any errors. The documents/files are routed to a decision point, then the users decide to route it to one of several branches by rejecting or accepting the annotation documents.

CMS data model and data standard

The IBM DB2 CM data model uses "items" as basic components for building content repositories. An "item" normally encompasses a document and a set of user-defined "attributes," which can be placed within a hierarchical structure. These attributes were indexed in the library server so that they can be used to retrieve the document objects more easily. An item can have multiple document objects stored in the resource managers and, for this reason, can be viewed in a flexible way. Each item can be externally linked to each other as well as to child components or to other metadata [31]. "Items" and their associated "attributes" within the CM deal with the relationships to other items, define the access control roles (e.g., who can access the item and the actions that authorized users can perform), describe the hierarchical data storage structure, and characterize the life cycle, workflow, and retention of the data [30].

The IBM CM data model allows users to connect loosely coupled genomic data and related annotations that can satisfy annotation specific, content-based needs. Thus, we can build up or define "items" and their related "attributes" based on the genome annotation tasks we conducted. When these "items" and their associated "attributes" were created and ready, the related annotation data were collected or created by BLAST(Basic Local Alignment Search Tools) or other annotation tools [6, 32], and then were converted

into a standardized XML file correspondent to a data source. Subsequently, we imported these XML data into the corresponding items within the CMS on the fly using a XML loading program (a Java-based XML loader program using an in-house, API provided by CM that will import the XML file into the program). The customized "item" with related attribute lists within the CM can be matched with the attribute names from the XML that is imported. Once the program sets up the right "item" with its attribute lists, the XML file can be recognized and imported in the "item" folder, with each value residing in the attributes. In addition, the multimedia file, such as a chromatogram image or the text file with genomic sequences, can also be imported into the folder that links to the same attribute list for the specific genomic record. The XML loading program runs on the background mode and exams the XML file to check the consistency of the defined requirement for the "DTD" (Document Type Definition) and the XML attribute names defined in the CM program. The program will automatically store and display the loaded information as the web application in the eClient side if the XML file has been successful imported into the program. In Figure 2, it showed two searching interfaces for the "items," representing the two real case scenarios: ESTs and the Simple Sequence Repeats (SSR) marker developments. The IBM CM builds in and supports full text search via DB2 Netsearch Extender), wildcard search or combinational search. The search return results give the list of the annotated sequences with associated attributes and also provide additional features for manipulating (e.g., display, export, delete) the individual annotation record (Figure 2). Users can review the search results and click on the text or image file related to the specific sequence record with the file display (Figure 2).

Hierarchical user privilege and access control

Being community annotation oriented, the IBM DB2 CM allows users to administrate the configuration of different user privileges for who can access the item and for the actions that authorized users can perform. It provides very fine-grained access controls. The user with an administrator role could also add additional users to the CM, define the data model to represent the stored objects, and set up document routing (workflow). These controls are required to limit each class of user permissions to view and/or update only those portions of the metadata allowed and, likewise, to only see those content objects they are permitted. The program also provides access control settings such as permission to perform an action on an item or object (e.g., read, search, reindex, delete).

Utility and Discussion

Content sensitive annotation tasks

Here, we present examples of grape genomics data management using IBM DB2 CM software. We characterized the information integration challenge and describe the components of an architecture that provides illustrated examples for transparently managing the volume and diversity of data. These real-life genomics data examples specifically illustrate the genomic data management. We developed online information resources based on two genomics tasks: 1) Expressed Sequence Tags (ESTs) and

sequence annotations and 2) SSR Molecular markers development. These task-oriented genomic curation processing and life cycling examples demonstrate the usability of the IBM DB2 CM that provides a consistent data model and opportunities to integrate these resources via standardized workflow management functions.

The relational databases were required to define a static schema to hold all the annotation-related, predefined attributes within the database, and the genome annotation tasks had to match these attribute lists without much flexibility. The data requirements changed frequently within a research environment, thus, the static structure of relational database were incompatible in dealing with on-going unexpected life cycles or changing requirements of data [29]. However, the CM solution has a high flexibility for unstructured annotation content. As for the first scenario case, the constellation of the ESTs related annotation data including clone name, Genbank access number, BLAST homology search returns using the NCBI non-redundant database, as well as the additional ESTs annotation for KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway or gene ontology related information produced by Blast2go program [32] were obtained during our dynamic research process. In the future, annotation information for these grape EST data can easily be updated. The same situation goes for the second scenario: genomic data for SSR marker development and grape genome mapping task. The first scenario focused on functional annotation for the genomic data, while the second scenario concentrated on the discovery of the sequence repeat patterns, location, and mapping (Table 1). These two scenarios had different annotation needs, thus required separate settings of attributes for annotation contents. The IBM CMS data model can easily create different annotation content repositories as “items” to store these heterogeneous genomic annotation data types and file formats.

Flexible and value-added content management

The large-scale datasets present formidable challenges in data integration and sharing for genomic data. The IBM DB2 CM allows users easily modify the default data schema, create annotation clusters/ “items” dynamically, update the system, or add new concept [25]. The high variability of the genomic data could produce very unfavorable signal/noise ratios. From the genomic data obtained from the sequencing machines during the pre-annotation process, to the functional annotation stage, data could be messy without proper evaluation. The “workflow management” module in CM can help users evaluate the quality of annotation when “routing” the annotation documents, and provide value-added inputs during the life-cycling process.

Our implemented IBM DB2 CM is being constructed to conveniently link different items among the heterogeneous data sets including genetic sequences, chromatograms, blast results, and gene ontologies within the same biological domain in a parent-child hierarchy structure. In addition, the database provides a convenient index and full text searching. Such a program will better serve genomic data integration and information retrieval with a user-friendly interface. We keep on adding new data formats of genomics annotation such as microarray gene expression profiles, as well as the second generation sequencing data (e.g., data produced by Solexa sequencing technology) as new content repositories in our IBM DB2 CM database.

This is a preprint of an article accepted for publication in *IBM Journal of Research and Development*. 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),(DOI: 10.1147/JRD.2011.2172837).

Conclusion

The demand for improving genomic data management will continue to increase. The content-based data framework within the current IBM CM provided high flexibility to build up task-oriented, content-based genome annotation practices. Users within the community could easily access, and retrieve annotation records and contribute their annotations with a great flexibility to build up task oriented annotation module. The real case scenarios that can satisfy a diverse set of genomic annotation tasks illustrated the scope of the information integration challenge and sketches out the requirements for a solution. Coordinated content capture, production, (dynamic) organization, workflow, access, and search functions were integrated coherently in the program. It provides annotation products routing workflow, information life cycling management, data quality document routing, and sharing and reusing of the content. Such a software is particularly useful for a "small" research community, so the users can actively interact with contents in the database and obtain useful outputs. The grape genomics database is being constructed to conveniently link different content-based "items" among the heterogeneous data. In addition, the database provides a convenient index and full text searching.

We created an application of IBM DB2 CM that is easy to adapt, understand, and access for the users. By integrating heterogeneous data sources or formats, the program has the capacity to store, share, search, and aid in data annotation, in gene/protein discovery, and in characterization and data quality assurance in different contexts for all available grape genomics data. Such a system can also serve as an example to build similar genomic databases using IBM CM for easily managing genomic data in a research institute or life science company.

Acknowledgments

The authors wish to thank the funding support of the USDA Capacity Building Grant (#0205031) and FAMU-ARS Science Center for Excellence. The authors also want to thanks a very large number of people who provided insightful comments and helped formulate our thoughts as presented here.

References

[1] Morozova, O., M. Hirst and M. A. Marra, "Applications of new sequencing technologies for transcriptome analysis," *Annual Review.of Genomics and Human Genetics*, vol. 10, 135–151, 2009.

[2] J. Parkinson (ed), "Expressed Sequence Tags (ESTs) Generation and Analysis," Totowa, NJ., *Humana*, 2009.

[3] A. Rafalski, "Applications of single nucleotide polymorphisms in crop genetics."

Current Opinion in Plant Biology, vol. 5, no. 94-100, 2002.

[4] Y.F. Leung, and Cavalieri, D, "Fundamentals of cDNA microarray data analysis," *Trends in Genetics*, vol. 19, no. 649–659, 2003.

[5] K.D. Pruitt, Tatusova,T. and Maglott,D.R, "NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic Acids Research*. Vol. 33, no. D501–D504, 1-8, 2005.

[6] R. Madupu, Brinkac L.M., Harrow J., Wilming L.G., Böhme U., Lamesch P., and Hannick L.I., "Meeting report: a workshop on best practices in genome annotation," *Database : the journal of biological databases and curation*, Vol. 2010, no. baq001, 1-17, 2010.

[7] H. Huang, Stvilia, B., Jörgensen, C., and Bass, H, "Prioritization of data quality dimensions and skills requirements in genome annotation work," *Journal of the American Society for Information Science and Technology*. in press, 2011.

[8] E. Poustelnikova, A. Pisarev, M. Blagov, M. Samsonova, and J. Reinitz, "A database for management of gene expression data *in situ," Bioinformatics*, vol. 20, no. 14, 2212-2221, 2004.

[9] M. Attimonelli, M. Accetturo, M. Santamaria, D. Lascaro, G. Scioscia, G. Pappadà, L. Russo, L. Zanchetta, and M. Tommaseo-Ponzetta, "HmtDB, a human mitochondrial genomic resource based on variability studies supporting population genetics and biomedical research," *BMC Bioinformatics*, vol. 6, S4, 1-9, 2005.

[10] S. Schönherr, H. Weißensteiner, S. Coassin, G. Specht, and F. Kronenberg, A. Brandstätter, " eCOMPAGT – efficient Combination and Management of Phenotypes and Genotypes for Genetic Epidemiology," *BMC Bioinformatics*, vol. 10, no. 139, 1-11, 2009.

[11] X. Chen, Y. Lin, M. Liu, and M.K. Gilson, "The Binding Database: data management and interface design," *Bioinformatics*, vol. 18, no. 1, 130-139, 2002.

[12] Z. Du, X. Zhou, L. Li, and Z. Su, "plantsUPS: a database of plants' Ubiquitin Proteasome System," *BMC Genomics*, vol. 10, no. 227, 2009.

[13] X. Yang, Y. Ye, G. Wang, H. Huang, and S. Liang, "VeryGene: linking tissue-specific genes to diseases, drugs and beyond for knowledge discovery," *Physiological Genomics,* 2011, (doi:10.1152/physiolgenomics.00178.2010).

[14] J. Yu, S. Pacifico, G. Liu, and R.L. Finley, "DroID: the Drosophila Interactions Database, a comprehensive resource for annotated gene and protein interactions," *BMC Genomics*, vol. 9, no. 461, 1-9, 2008.

[15] D. Gonzalez-Ibeas, C. Blanca J, Roig, M. Gonzalez-To, B. Pico, V. Truniger, P. Gomez, W. Deleu, A. Cano-Delgado, and P. Arus, "MELOGEN: an EST database for melon functional genomics," *BMC Genomics*, vol. 8, no. 206, 1-17, 2007.

[16] A. Blenda, J. Scheffler, B. Scheffler, M. Palmer, J.M. Lacape, J.Z. Yu, C. Jesudurai, S. Jung, S. Muthukumar, P. Yellambalase, S. Ficklin, M. Staton, R. Eshelman, M. Ulloa, S. Saha, B. Burr, S. Liu, T. Zhang, D. Fang, A. Pepper, S. Kumpatla, J. Jacobs, J. Tomkins, R. Cantrell, and D. Main, "CMD: a Cotton Microsatellite Database resource for Gossypiumgenomics**,**" *BMC Genomics*, vol. 7, no. 132, 1-10, 2006.

[17] P. Gaudet, A. Bairoch, D. Field, S.A. Sansone, C. Taylor, T.K.,Attwood, A. Bateman, J.A. Blake, C.J., Bult, and J.M. Cherry, "Towards BioDBcore: a community-defined information specification for biological databases," *Nucleic Acids Research*,vol. 39, D7-D10, 1-4, 2011.

[18] D. Smedley, P. Schofield, C.K., Chen, V. Aidinis, C. Ainali, J. Bard, R. Balling, E. Birney, A. Blake, and E. Bongcam-Rudloff E, "Finding and sharing: new approaches to registries of databases and services for the biomedical sciences," *DATABASE*, doi:10.1093/baq014, 1-5, 2010.

[19] L.A. Flórez, S.F. Roppel, A.G. Schmeisky, C.R. Lammers, and J.A. Stülke, "Community-curated consensual annotation that is continuously updated: the Bacillus subtilis centred wiki SubtiWiki*," DATABASE*, vol. 2009, bap012, 2009.

[20] S.L. Salzberg, "Genome re-annotation: a wiki solution?" *Genome Biology*, vol. 8, no, 102, 1-5, 2007.

[21] T.W. Tan, J.C. Tong, M. De Silva, K.S. Lim, and S. Ranganathan, "Advancing standards for bioinformatics activities: persistence, reproducibility, disambiguation and Minimum Information about a Bioinformatics Investigation (MIABi)," *BMC Genomics*, vol. 11(Suppl. 4), no. S27, 1-7, 2010.

[22] R. Madupu, L.M. Brinkac, J. Harrow, L.G. Wilming, U. Bohme, P. Lamesch, and L. Hannick. "Meeting report: a workshop on Best Practices in Genome Annotation," *DATABASE*, vol.2010, baq001, 1-17, 2009..

[23] J. vom Brocke, A. Simons, and A. Cleven. "Towards a business process-oriented approach to enterprise content management: the ECM-blueprinting framework," *Information Systems and E-Business Management*, 2009, DOI: 10.1007/s10257-009-0124-6, 1-12.

[24] L. Wiltzius, A. Simons, and S. Seidel, "A Study on the Acceptance of ECM Systems," *Wirtschaftinformatik Proceedings 2011.* Paper 77. http://aisel.aisnet.org/wi2011/77

[25] A. Jhingran, N. Mattos, and H. Pirahesh, "Information Integration: A Research Agenda," *IBM Systems Journal*, vol. 41, no.4, 555-562, 2002.

[26] P. L. Bradshaw, K. W. Brannon, T. Clark, K. Dahman, S. Doraiswamy, and L. Duyanovich, "Archive storage systems design for long-term storage of massive

amounts of data," *IBM Journal of Research & Development*, vol. 52, no. 4/5, 1-10, 2008.

[27] S. Das, L. Girard, T. Green, and L. Weitzman, "Building biomedical web communities using a semantically aware content management system," *Briefings in Bioinformatics*, vol. 10, no. 2, 129-138, 2009.

[28] S.D. Mooney, and P.H. Baenziger, "Extensible open source content management systems and frameworks: a solution for many needs of a bioinformatics group," *Briefings in Bioinformatics*, vol. 9, no. 1, 69-74, 2008.

[29] J. Boyle, H. Rovira, C. Cavnor, D. Burdick, and S. Killcoyne, "Adaptable data management for systems biology investigations," *BMC Bioinformatics*, vol. 10, no. 79, 1-16, 2009.

[30] V.S. Smith, S.D. Rycroft, K.T. Harmen, B. Scott, and D. Roberts, "Scratchpads: A data-publishing framework to build, share and manage information on the diversity of life," *BMC Bioinformatics*, vol. 10, no. S14, S6, 2009.

[31] M. Herbach, "Manage your digital content effectively," IBM DB2 Whitepaper G325-2143-00, March 2003.

[32] S. Götz, J.M. García-Gómez, J. Terol, T. D. Williams, M.J. Nueda, M. Robles, M. Talón, J. Dopazo, and A. Cones, "A High-throughput functional annotation and data mining with the Blast2GO suite," *Nucleic Acids Research*, vol. 36, no. 10, 3420–3435, 2008.

This is a preprint of an article accepted for publication in *IBM Journal of Research and Development*. 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),(DOI: 10.1147/JRD.2011.2172837).

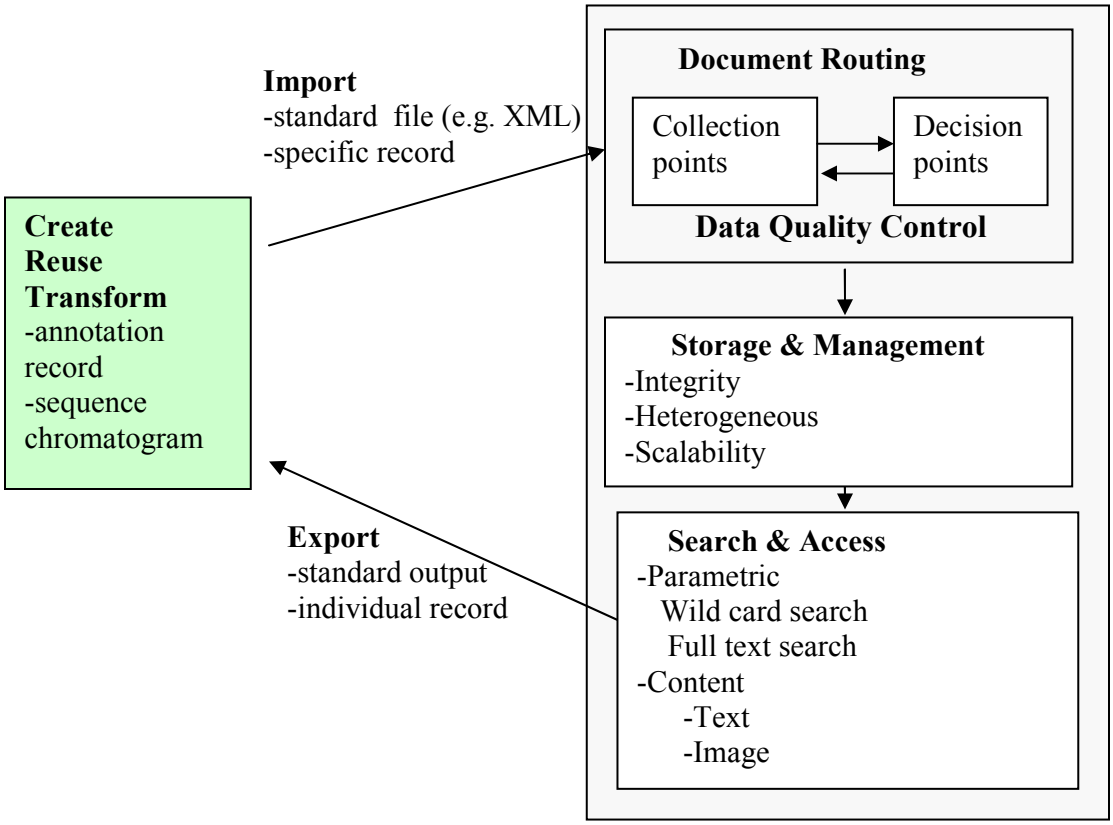
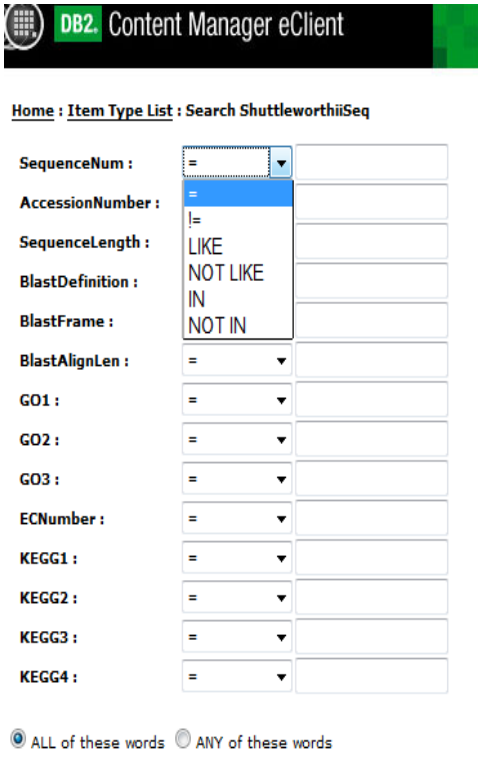


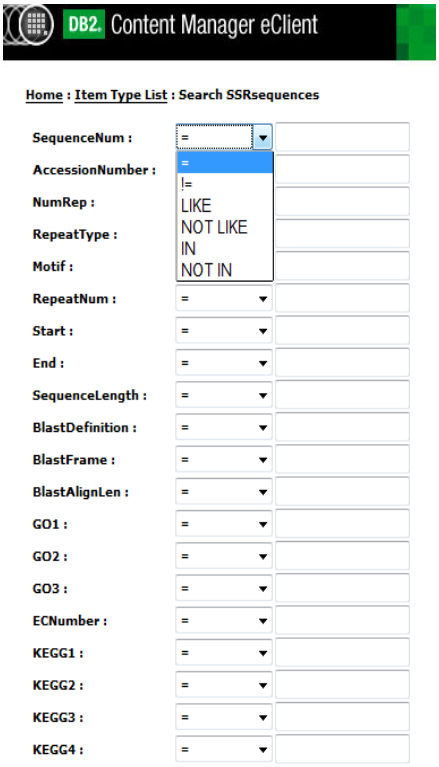
Figure 1. The data workflow in the genomics applications of the IBM CM. 1) create/reuse annotation: a user could conduct task-oriented genome annotation work. 2) Import: the genome annotation data were translated to a standardized XML file that can be imported into the IBM CM. 3) Data quality control: the imported data can be routed to the decision points for data quality checking. 4) Storage and Management: attributes or items can be cross-linked and integrated together. 5) Search and access: the system support parametric or content search. 6) Export: the individual or multiple annotation records as well as their sequence data can be exported or downloaded from the system. Users can continue the annotation work during the life cycling process.

This is a preprint of an article accepted for publication in *IBM Journal of Research and Development*. 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),(DOI: 10.1147/JRD.2011.2172837).

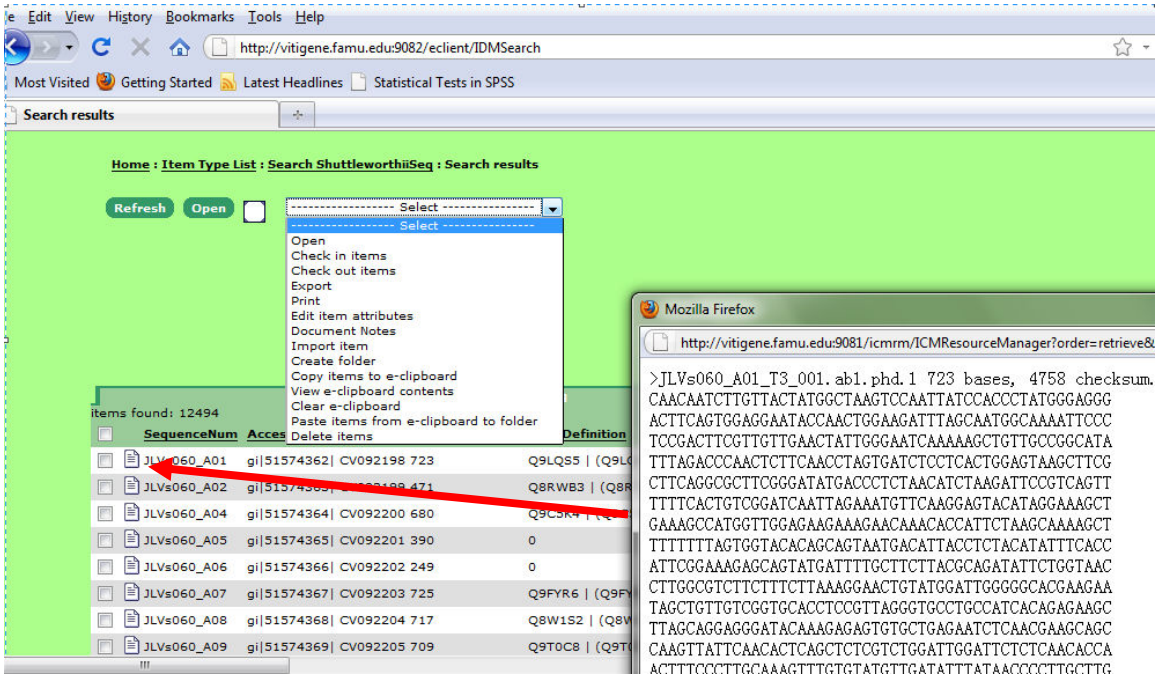
This is a preprint of an article accepted for publication in *IBM Journal of Research and Development*. Huang H., Lu J., Hunter W., Liang S., (2011). Using IBM Content Manager for genomic data annotation and quality assurance tasks. *IBM Journal of Research and Development* , 55(6),(DOI: 10.1147/JRD.2011.2172837).



A



B



This is a preprint of an article accepted for publication in *IBM Journal of Research and Development*. 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),(DOI: 10.1147/JRD.2011.2172837).

C

Figure 2. Attribute lists and search functions and result example for two "items". (A) ESTs item: it included a list of attributes related to blast searching, KEGG pathway and gene ontology annotation, as well as putative Enzyme Commission number. (B)SSR markers item: it contained attributes for SSR characteristics such as repeat type, motif, and repeat number information. It also provided genome annotation data for blast searching, molecular pathways, and gene ontology. Both items supported parametric, full-text and combined parametric and full-text search. (C) The online search on a specific item: ESTs sequences can return a list of sequence records. Users can further choose by checking in/out, exporting, importing, editing the sequence records. By further clicking on "document icon", users can view actual sequence data in a popup window. Definitions: GO1: Gene Ontology level 1; GO2: Gene Ontology level 2; ECNumber: Enzyme Commission number; KEGG1: KEGG level 1; KEGG2: KEGG level 2; BlastDefinition: homology search result for a genetic sequence; BlastFrame: BLAST search reading frame number; BlastAlignlen: the length of homology search result; SequenceNum: the EST clone sequencing number.

This is a preprint of an article accepted for publication in *IBM Journal of Research and Development*. 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),(DOI: 10.1147/JRD.2011.2172837).

Table 1. Genomic data annotation scenarios, defined items, attributes and the heterogeneous data types and formats.

Annotation Scenarios/items	ESTs	SSR marker sequences
Number of sequences	25,300	1,883
Grape species	V. shuttleworthii, muscadine	V. vinifera, shuttleworthii, muscadine
Annotation tasks/attributes	sequence ID BLAST results gene ontology sequence chromatograph files	sequence ID BLAST results gene ontology SSR patterns genome mapping files
file formats	text, image (chromatogram)	text, image (genetic map)

This is a preprint of an article accepted for publication in *IBM Journal of Research and Development*. 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),(DOI: 10.1147/JRD.2011.2172837).

Biographical sketches

Hong Huang, *School of Information, University of South Florida, Tampa, FL, 33620, U.S.A. (honghuang@usf.edu.)* Dr. Huang is an Assistant Professor in Health Informatics, at University of South Florida. He received two M.S. degrees in Genetics (2001) and Computer Science (2003), and PhD in Information Studies from the Florida State University in 2010. His industrial experiences included working as a medical representative in Novartis, and internship in Mainline Information System (an IBM business partner). He received over $1M grants supports for genomics and bioinformatics. He is an author or coauthor of over 50 journal publications or conference proceedings. Dr. Huang is a member of the American Medical Informatics Association, American Telemedicine Association, and The Healthcare Information and Management Systems Society.

Jiang Lu, *Center for Viticulture and Small Fruit Research, Florida A&M University, Tallahassee, FL, 32308, U.S.A. (jiang.lu@famu.edu.)* Dr. Lu is a Professor in genetic and functional genomics at Florida A&M University, and an adjunct professor at University of Florida. He developed and used bioinformatics data mining tools and databases for expediting grape breeding process. Dr. Lu received 2011 Florida A&M University Research Excellence Award. He has received over $3M in total grants, and published over 80 research papers and given more than 140 research presentations in local, regional, national and international conferences.

Wayne Hunter, *United States Department of Agriculture, Agriculture Research Service, United States Horticultural Research Laboratory, Fort Pierce, FL, U.S.A. (wayne.hunter@ars.usda.gov.)* Dr. Hunter is the Lead Scientist on the International Psyllid Genome Consortium and the Intl. Leafhopper Genome Consortium, conducting research on genomics and bioinformatics of hemipteran insects which spread plant diseases. His efforts were awarded 'Entomologist of the Year 2005', 'Team Research award 2005' both from the Florida Entomological Society; 'Distinguished Alumnus 2008', by Casper College, WY; and 'Recognition Award in Entomology 2011' from Southeastern Branch of Entomological Society of America, for Internationally recognized contributions to insect genomics, and the '2011 Achievement Award for Team Research' on psyllid genomics. He is a member of the International Aphid Genomics working group. Research includes researchers from the USA, France, Spain, Japan, Australia and Italy.

Shuang Liang, *College of Medicine, Southern Medical University, Guangzhou, 510515, China. (itshuang@gmail.com.)* Dr. Liang is a principal investigator and professor affiliated with the College of Medicine where he is responsible for developing research programs by means of medical bioinformatics. He received his doctoral degree in Molecular Genetics from University of P. & M. Curie (Paris VI), his master degree in Computer Science from

Northeastern University (USA). He spent five years as postdoc fellow at Case Western Reserve University and worked his way up to various senior positions for international firms including Bell Lab of Lucent Technologies and Wyeth Pharmaceuticals in the areas of IT and drug development. He has involved in the development of telecommunication billing systems as well as e-procurement platform and large scale data-mining and building a biological knowledge discovery platform. He has authored over twenty articles in some of the well known peer-reviewed journals.