BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net *Software*

Volume 11(4)

Omics Metadata Management Software (OMMS)

Martha O Perez-Arriaga[1], Susan Wilson[2], Kelly P Williams[3], Joseph Schoeniger[3], Russel L Waymire[5] & Amy Jo Powell[2,4,6]*

[1]Department of Computer Science, Mail stop: MSC01 1130, 1 University of New Mexico, Albuquerque, NM 87131-0001; [2]Center for Advanced Research Computing, University of New Mexico, Albuquerque, NM 87131; [3]Systems Biology, Sandia National Laboratories, Mail Stop 9291, Livermore, CA 94550; [4]Department of Biology, Mail stop: MSC03 2020, 1 University of New Mexico, Albuquerque, NM 87131-0001; [5]Security Systems Analysis, Sandia National Laboratories, Mail Stop 0757, Albuquerque, NM 87123; [6]Computational Simulation Infrastructure, Sandia National Laboratories, Mail Stop 0845, Albuquerque, NM 87123; Amy Jo Powell – Email: ajpowel@sandia.gov; Phone: +1 505 284 8050; *Corresponding author

Received October 11, 2014; Accepted March 16, 2015; Published April 30, 2015

Abstract:
Next-generation sequencing projects have underappreciated information management tasks requiring detailed attention to specimen curation, nucleic acid sample preparation and sequence production methods required for downstream data processing, comparison, interpretation, sharing and reuse. The few existing metadata management tools for genome-based studies provide weak curatorial frameworks for experimentalists to store and manage idiosyncratic, project-specific information, typically offering no automation supporting unified naming and numbering conventions for sequencing production environments that routinely deal with hundreds, if not thousands of samples at a time. Moreover, existing tools are not readily interfaced with bioinformatics executables, (e.g., BLAST, Bowtie2, custom pipelines). Our application, the Omics Metadata Management Software (OMMS), answers both needs, empowering experimentalists to generate intuitive, consistent metadata, and perform analyses and information management tasks via an intuitive web-based interface. Several use cases with short-read sequence datasets are provided to validate installation and integrated function, and suggest possible methodological road maps for prospective users. Provided examples highlight possible OMMS workflows for metadata curation, multistep analyses, and results management and downloading. The OMMS can be implemented as a stand alone-package for individual laboratories, or can be configured for web-based deployment supporting geographically-dispersed projects. The OMMS was developed using an open-source software base, is flexible, extensible and easily installed and executed. The OMMS can be obtained at http://omms.sandia.gov.

Availability: The OMMS can be obtained at http://omms.sandia.gov.

Keywords: Bioinformatics, relational database management system, omics, next-generation sequencing, biological curation, open-source software, integrated workflow

Background:
Next-generation sequencing has revolutionized research and medicine, and has been accompanied by increasingly challenging, if underappreciated, metadata management requirements **[1]**. Sequencing projects that have large, complex metadata demand lightweight, flexible, stable software to support dataset standardization, biological curation activities, and streamlining of pre- and post- processing steps with pipeline analyses **[2].** The few existing open-source tools for managing next-generation sequencing metadata have limited flexibility, thus hampering project-specific tailoring, and cannot be easily deployed and managed by a single administrator for multiple research sites **[3, 4].** Commercially-available laboratory information management alternatives are typically expensive, cumbersome and require proprietary administration and maintenance for the software lifecycle **[5],** and none of the available tools, open-source or otherwise, readily integrate curation, processing and advanced analyses.

We developed the Omics Metadata Management Software (OMMS), a flexible, extensible, open-source, web-based tool that provides semi-automated curation utilities, and integrated implementation with widely-used bioinformatics executables, such as BLAST **[6]** and Bowtie **[7],** for human-microbiome-oriented research in our laboratories **[8].** Example use cases with publicly available human microbiome and chimpanzee RNASeq datasets **[9, 10]** are detailed to demonstrate OMMS function and versatility, and operation as a pipeline frontend.

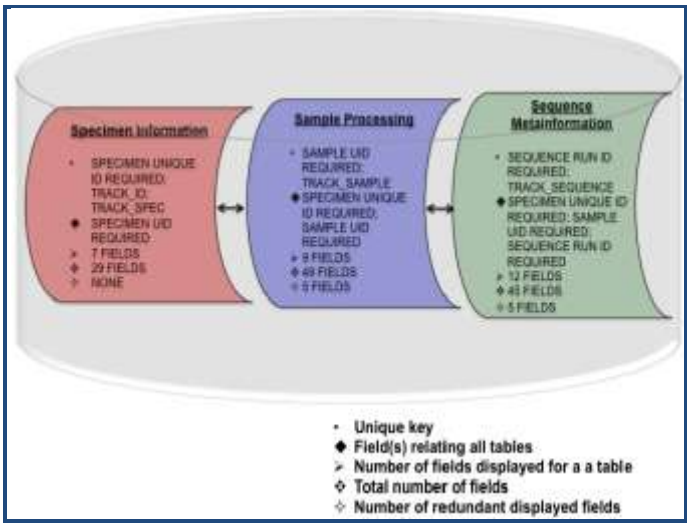


Figure 1: Omics Metadata Management Software. Core functionality resides in three tables, "Specimen Information," "Sample Processing," and "Sequence MetaInformation," which have fields with embedded automation supporting efficient data metadata entry, storage and intuitive entity relationships facilitating data sharing and analysis. These tables are accessed via the "MetaData" portal.

Omics Metadata Management Software

The OMMS was engineered to support a large, multidisciplinary, geographically-dispersed research team developing next-generation sequencing-based approaches for identifying potentially rare etiologic agents in human microbiomes across hundreds of distinct sample types **[8].** The OMMS graphical user-interface (GUI) enables semi-automated project-specific metadata entries in each table **(Figures 1 & 2**; **Table 1**, **3-5 (see supplementary material))**, and associated input sequence data are referenced to archiving locations in directories generated on Linux-based file systems. Intuitive point-and-click selection of input sequence files, analysis configuration and execution are carried out in the "Analysis" portal **(Figure 3**; **Table 2 (see supplementary material))**. In examples provided here, BLAST, Bowtie 2, TopHat and Cufflinks were integrated and implemented with the OMMS interface **[6, 7, 10],** and in principle, any open-source application can be integrated with the OMMS, including in-house pipelines, with custom scripts developed for that purpose.

Methodology:

Creating a record

The three main portals are displayed after login. Tables for detailed biological curation ("Specimen Info," "Sample Processing," "Sequence MetaInfo") reside in the "MetaData" portal. To create an entry, select "Create New," and then click on "New (Empty fields)" in the "Specimen Info" table, and provide required information (**Figure 1**; **Table 1(see supplementary material)**). The following (parentheses) were entered: Host Species (*Homo sapiens*); Tissue Sampled (Stool). Click the "Add Specimen" button to generate the "Specimen Unique ID" (HsStoo_01). For the second record, repeat the previous steps, but insert "*Pan troglodyte*" and "Brain" for the host and tissue, respectively, to generate "Specimen Unique ID" (PtBrai_02).

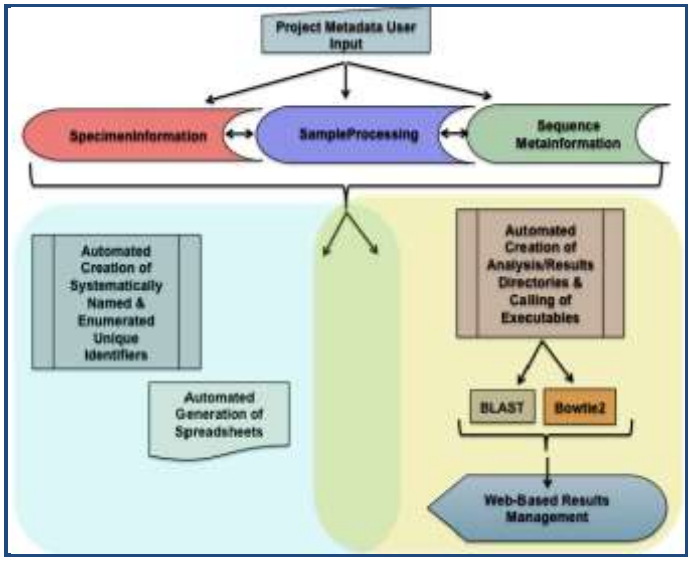
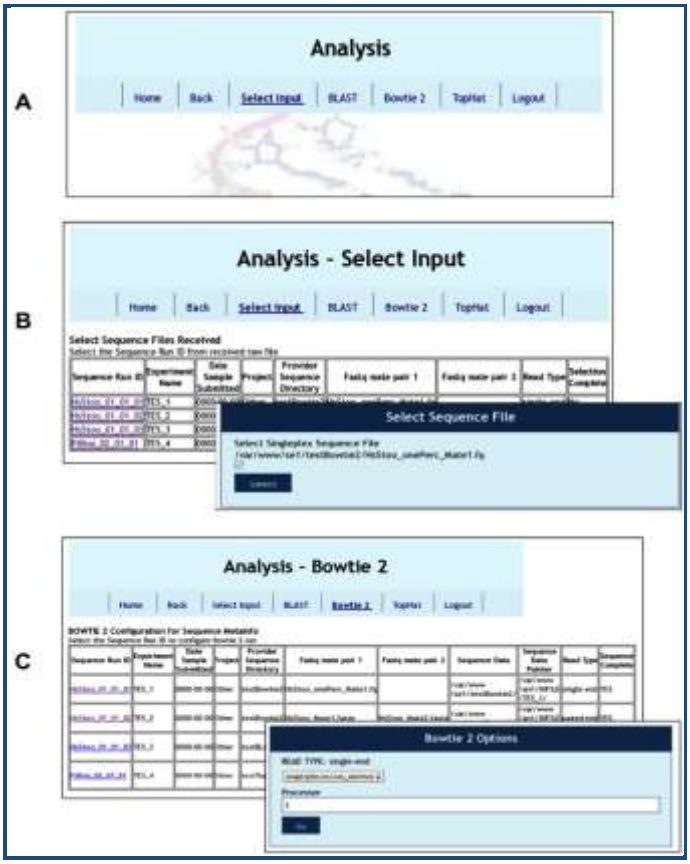


Figure 2: Unified framework for metadata management and state-of-the-art analyses. Curation (highlighted in aqua) and analyses (indicated in yellow) tasks are intrinsically related (overlap region) in next-generation sequencing studies, because sample handling and sequence production are multistep processes, and careful metadata tracking and management are required for downstream analyses and publication preparation. The OMMS supports user input of project metadata, automated creation of consistently named and enumerated unique identifiers for specimens, samples and sequence production information, and straightforward integration with bioinformatics utilities. Spreadsheets can be generated for structured data extraction and local download. Standard input and output of executables used here are stored in automatically-generated files and directories.

Corresponding records are generated in the "Sample Processing" table, with dropdown menus provided to streamline curation by explicitly linking the "Specimen Unique ID" (*e.g.*, HsStoo_01, PtBrai_02) and the user-defined "Sample Alias" with the new sample entries. "Sample Unique ID" entries are generated by clicking "Add Sample" (*e.g.*, HsStoo_01_01, PtBrai_02_01). Four corresponding records were created in the "Sequence MetaInfo" table to complete the curation exercise, and to illustrate functional integration with executables **Table 2 (see supplementary material)**; in the "Provider Sequence Directory Name" field, arbitrary directory names were given; for "Fastq File Mate Pair 1" and "Fastq File Mate Pair 2," test input file names were used, and appropriate options were chosen for "Read Type" field for testing in this order: Bowtie 2 (single–, then paired end with human microbiome stool fastq files), BLAST (with human microbiome stool fasta input), Tophat and Cufflinks (with the single-end

chimp RNASeq file). The "Sequence Run ID" and "Unique Experiment Name" were generated by clicking the "Add Sequence." In each of the tables, the "Update" function can be used to extend curation. Methods for generating and downloading custom metadata tables are further detailed in the "OMMS Integrated Workflow" link (under Quick Start).



Figures 3A-C: Omics Metadata Management Software (OMMS) curation and analysis interface. The OMMS was designed to integrate and implement with open-source bioinformatics tools, such as BLAST, Bowtie 2 and Tophat and/or custom pipelines. These tools are accessed via the "Analysis Portal" (panel A). End users select the identifier ("Sequence Run ID") of interest, which is referenced to particular sequence files (panel B and inset). Following input selection, the desired program is chosen and parameterized (panel C and inset) to launch a run. Output from a given analysis run can be downloaded via the OMMS "Results" portal (not shown).

Enabling integrated workflows

To call integrated executables via the "Analysis" portal, click on "Select Input" for the relevant Sequence Run ID (*e.g.*, HsStoo_01_01_01), and then choose the desired program (**Figure 2**). To launch a Bowtie 2 run on "Sequence Run ID" HsStoo_01_01_01, select the "*Staphylococcus_aureus*" index from the dropdown menu, and enter an integer value in "Processors" and click "Go." The results file name will appear, and standard output can be downloaded upon run completion. The same steps apply for paired-end analyses with Bowtie 2. For BLAST runs, select the pertinent Sequence Run ID and input (HsStoo_BLAST500.fa), and choose the desired program (blastn) and database (*Clostridium kluyveri*). Set the significance threshold expectation (E) value at 0.001 or higher, and indicate the desired output format in the dropdown menu, and click "Go." Similar steps are followed for splice-variant and/or differential expression analyses using TopHat and Cufflinks (**Figure 3; Table 2 (see supplementary material)**). The chimp "Sequence Run ID" was selected (PtBrai_02_01_01, referenced to RNASeq file SRR023838_RNASeq.fq) and aligned with the hg19 index **[7].** Standard output can be downloaded via the "Results" portal by choosing the "Sequence Run ID" of interest (*e.g.*, HsStoo_01_01_01). The website provides additional instructions for building integrated analyses (in the "OMMS Integrated Workflow" link under Quick Start).

Software:

Design and function

Interoperable, open-source software packages (*i.e.*, the LAMP bundle, **L**inux, **A**pache, **M**ySQL, **P**HP) wereused to develop the browser-based OMMS interface to support next generation sequencing-based research efforts in our laboratories **[8].** Real-world metadata associated with the test datasets were entered in the three tables, "Specimen Info," "Sample Processing," and "Sequence MetaInfo" (**Figure 1**; **Table 1, 3-5 (see supplementary material)**) to instantiate example database records. Most of the fields in the tables accommodate varied data types (*e.g.*, the "Sample_Alias" field in the "Sample_Processing" table), such as character strings, but in cases with fewer possibilities, dropdown menus are provided (*e.g.*, the "Nucleic Acid" field in the "Sample Processing" table).

Test datasets for validating installation and benchmarking integrated tools

Distinct hosts and tissues (*Homo sapiens*, *Pan troglodyte*; Stool, Brain) were used to demonstrate automated metadata tracking, storing and functional integration with utilities **[9, 10].** Test datasets were obtained from the GenBank Short Read Archive (accessions SRX025177: SRR063480 and SRX008322: SRR023838), and were pre-processed using the NCBI SRA and Fastx Toolkits, and in-house custom scripts. These pre-processing steps are explained in the README file included in the distribution and in the "OMMS Integrated Workflows" link (see the **Supplemental Materials** for fine details pertaining to curation and pre-processing steps).

Semi-automated curation and results downloading

After entering the minimum required information (indicated by asterisks) for a specimen, the OMMS generates a unique identifier under the "Specimen_UID" field (**Figure 1**; **Table 1** & **3 (see supplementary material)**) describing the subject/host and tissue/microhabitat from which nucleic acid preparations and sequence data will be derived **(Figure 1 & 2; Table 1 (see supplementary material)**). Unique identifiers are automatically propagated to corresponding fields in the other tables (**Figures 1 & 2**), intuitively linking specimen, sample and sequence data. Input sequence data files put can be uploaded, and results files (output) downloaded, as can metadata for specific entries, as well as table-overview custom spreadsheets (**Figures 2 & 3**).

Concluding Remarks:

The freeware reported here guarantees standardized, intelligible, automated curation and management of biological

metadata, and supports integrated analyses. Recent events, from the outbreak of Ebola Virus Disease in West Africa, to the emergence of antibiotic-resistant bacteria (*e.g.*, *Clostridium difficile*, Carbapenem-resistant Enterobacteriaceae), make it impossible to overstate the importance of rigorous metadata curation and management systems in high-intensity scenarios, clinical and otherwise. For our project, the OMMS frontend was foundational for handling metadata inherent to next-generation sequencing-based experiments involving large numbers of samples at a time. In the context of a host microbiome, potential etiologic agents are typically rare and difficult to detect using standard *in silico* and experimental approaches, and careful metadata curation is crucial for identifying signal (infectious disease) in the presence of overwhelming noise (background microbiota) and results interpretation. Looking ahead, the OMMS and OMMS-user tailored versions will represent easy-to-use promising tools for addressing microbiome-centric research, from clinical and public health challenges, to exploring new frontiers in agricultural research and development, where handing and tracking hundreds, if not thousands, of samples from diverse subjects at a particular location and time are necessary. Additionally, the OMMS enables development of integrated workflows with state-of-the-art utilities (Blast, Bowtie 2) and in-house pipelines (*e.g.*, local implementations of Galaxy), facilitating fine-grained comparative analyses, such as strain discrimination (*e.g.*, Zaire *vs.* Sudan ebolavirus) and microbiome composition and functional profiling.

Acknowledgement:
This R & D was supported by the Laboratory Directed Research and Development (LDRD) program at Sandia National Laboratories [under the auspices of the Rapid Threat Organism Recognition (RapTOR) Grand Challenge (LDRD # 142042)]. Sandia National Laboratories are multi-program laboratories managed and operated by Sandia Corporation, a wholly-owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration (contract DE-AC04-94AL85000). We offer our sincerest thanks to the RapTOR team, and Dr. Elisa LaBauve, in particular, for invaluable discussions around software design and development, and Professor Melanie Moses for guidance on manuscript preparation. As ever, I extend my deepest gratitude to Professors Donald O. Natvig and Gavin C. Conant for invaluable, ongoing scientific discussions, editorial assistance and software testing. Sincerest thanks to Steven Arroyo, Brian Nelson, Michael W. Folsom and Mark D. Murton for software testing and website development. We are grateful to the University of New Mexico Center for Advanced Research Computing and Dr. Susan Atlas, Director, for computational resources and technical assistance provided in support of this work.

References:

[1] Pareek C *et al. J Appl Genetics* 2011 **52**: 413 [PMID: 21698376]
[2] Wruck W *et al. Brief Bioinform* 2014 **15**: 65 [PMID: 23047157]
[3] Rocca-Serra P *et al. Bioinformatics* 2010 **26**: 2354 [PMID: 20679334]
[4] Wolstencroft K *et al. Bioinformatics* 2011 **27:** 2021 [PMID: 21622664]
[5] Gross J Labguru 2011, http://www.labguru.com/, Tel-Aviv, Israel: BioData Ltd
[6] Altschul SF *et al. Nucleic Acids Research* 1997 **25**: 3389 [PMID: 9254694]
[7] Langmead B & Salzberg SL, *Nat Meth* 2012 **9**: 357 [PMID: 22388286]
[8] Bent ZW *et al. Anal Biochem.* 2013 **438**: 90 [PMID: 23535274]
[9] Human Microbiome Project Consortium. *Nature* 2012 **486**: 207 [PMID: 22699609]
[10] Trapnell C *et al. Nat Protocols.* 2012 **7**: 562 [PMID 22383036]

Edited by P Kangueane
Citation: **Arriaga *et al.*** Bioinformation 11(4): 165-172 (2015)

BIOINFORMATION

open access

Supplementary material:

Table 1: Overview of tables with embedded automation supporting specimen, sample and sequence data curation and storage. The OMMS supports next-generation sequencing projects, which invariably have substantial metadata management overhead. Core software functionality resides in the three interoperable tables, which are accessed via the "MetaData" portal.

Table Name	Total number of fields	Number of required fields	Number of fields present in user interface "Consult" mode	Number of redundant fields	Number of automated fields	Fields used to relate other tables
Specimen Info	29	6	7	0	7	Specimen_UID_req
Sample Processing	48	5	9	5	7	Specimen_UID_req Sample_UID_req
Sequence MetaInfo	44	6	12	5	9	Specimen_UID_req Sample_UID_req SequenceRun_ID_req

Table 2: Integration of curation tasks and bioinformatics analyses. The "Omics Metadata Management System" (OMMS) is designed for integrated use with Bowtie 2, Tophat, Cufflinks and BLAST, and can be extended and configured to call and run other open-source executables, such as

Specimen Unique ID	Sample Unique ID	Sequence Run ID	Provider Sequence Directory Name[3]	Unique Experiment Name[3]	Sequence Input (Mate Pair 1)[3]	Sequence Input (Mate Pair 2)[3]	Executable	Input Name[3]	Output Name[3]
HsStoo_01	HsStoo_01_01	HsStoo_01_01_01	testBowtie2	TES_1	HsStoo_onePerc_Mate1.fq		bowtie2	INFILE/TES_1/TES_1	OUTFILE/TES_1/TES_1-1.sam
HsStoo_01	HsStoo_01_01	HsStoo_01_01_02	testBowtie2	TES_2	HsStoo_Mate1.fq	HsStoo_Mate2.fq	bowtie2	INFILE/TES_2/PAIR1/TES_2 INFILE/TES_2/PAIR2/TES_2	OUTFILE/TES_2/PAIR1/TES_2-1.sam
HsStoo_01	HsStoo_01_01	HsStoo_01_01_03	testBLAST	TES_3	HsStoo_BLAST500.fa		blastn	INFILE/TES_3/TES_3	OUTFILE/TES_3/TES_3-1
HsBrai_02	HsBrai_02_01	HsBrai_02_01_01	testTopCuff	TES_4	SRR023838_RNASeq.fq		tophat cufflinks	INFILE/TES_4/TES_4	OUTFILE/TES_4/TES_4-1/tophat_out/accepted_hits.bam

custom pipelines. Sequence files are referenced to automatically generated metadata entries (e.g., "Sample Unique ID", "Sequence Run ID"), and these files can be used as input for programs accessed via the "Analysis" portal. The OMMS also generates directories to store results (output) for each run of a program. Output can be accessed and downloaded via the "Results" portal.

Table 3: Specimen Information

Field Name	Data Type	Required Field	Automated Field	Drop-down Menu
Specimen_UID_req	varchar(100)		X	
Specimen_Alias_req	varchar(100)			
Date_Acquired_req	date			
Acquired_By_req	varchar(100)	X		
Storage_Location_req	varchar(100)			
Source_req	varchar(100)	X		
Specimen_Type_req	varchar(100)	X		X
Host_ID_req	varchar(100)	X		
Host_Organism_req	varchar(100)	X		
Host_Ploidy	varchar(100)			

Host_Gender	varchar(100)			X	
Host_Age	int(11)				
Host_Ethnicity	varchar(100)				
Tissue_Sampled_req	varchar(100)	X			
Diagnostic_Screen	varchar(5000)				
Pathogen_Type	varchar(100)				
Pathogen_Dose	varchar(100)				
Pathogen_Delivery	varchar(100)				
Specimen_Collection_Timing	varchar(100)				
Experimentalists_notes_upload	varchar(100)				
fileuploaded	mediumblob				
filetype	filetype				
Remarks	varchar(5000)				
createdby	varchar(20)		X		
timecreation	timestamp		X		
modifiedby	varchar(20)		X		
timemodified	timestamp		X		
track_id	int(11)		X		
track_spec	int(11)		X		

Table 4: **Sample Processing**

Field Name	Data Type	Required Field	Automated Field	Drop-down Menu
Specimen_UID_req	varchar(100)	X		X
Specimen_UID_sec	varchar(50)			
Sample_UID_req	varchar(100)		X	
Sample_Alias_req	varchar(100)			
Nucleic_Acid_req	varchar(100)			
Target_Nucleic_Acid_req	varchar(100)			
Purification_Nucleic_Acid_Method_req	varchar(100)	X		X
Purified_By_req	varchar(100)	X		
Date_Nucleic_Acid_Purified_req	date			
Purified_Nucleic_Acid_Yield_Nanograms	float(7,2)			
QC_Result	varchar(5000)			
Nucleic_Acid_Modification_Method	varchar(100)			X
Nucleic_Acid_Modified_By	varchar(100)			
Date_Nucleic_Acid_Modified	date			
Modified_Nucleic_Acid_Type	varchar(100)			
Modified_Nucleic_Acid_Yield_Nanograms	float(7,2)			
Modified_Nucleic_Acid_QC_Result	varchar(5000)			
Suppression_Product_Name	varchar(100)			
Suppression_Method	varchar(100)			X
Suppressed_By	varchar(100)			
Date_Nucleic_Acid_Suppressed	date			
Suppressed_Nucleic_Acid_Yield_Nanograms	float(7,2)			

Suppressed_Nucleic_Acid_QC_Result	varchar(5000)				
Barcode	varchar(200)	X			X
Primer_Set	varchar(30)	X			X
Primer_Name	varchar(100)				X
Barcode_Type	varchar(20)				X
Library_Preparation_ID_req	varchar(1000)				
Library_Preparation_Method_req	varchar(100)				X
Library_Prepared_By_req	varchar(100)				
Date_Library_Prepared_req	date				
Library_Nucleic_Acid_Input_Nanograms_req	float(7,2)				
Library_Prep_Final_Volume_Microliters_req	float(7,2)				
Final_Library_Concentration_Nanograms_Micro liter_req	float(7,2)				
Library_Nucleic_Acid_QC_Result	varchar(5000)				
Average_Insert_Size_basepairs	float(7,2)				
Sequencing_Provider_req	varchar(100)				
Date_at_Sequencing_Provider_req	date				
Sent_By_req	varchar(100)				
Experimentalist_Informatician_notes_upload	varchar(500)				
fileuploaded	mediumblob				
filetype	varchar(50)				
createdby	varchar(20)		X		
timecreation	timestamp		X		
modifiedby	varchar(20)		X		
timemodified	timestamp		X		
track_id	int(11)		X		
track_sample	int(11)		X		

Table 5: Sequence MetaInformation

Field Name	Data Type	Required Field	Automated Field	Drop-down Menu
Experiment_Name_req	varchar(50)		X	
Specimen_UID_req	varchar(100)	X		X
Specimen_UID_sec	varchar(50)			
Sample_UID_req	varchar(100)	X		X
Sample_Alias_req	varchar(100)	X		X
SequenceRun_ID_req	varchar(100)		X	
Sequence_Lane_req	int(11)			
Sequence_Data	varchar(200)		X	
Sequence_Data_Point	varchar(200)		X	
Date_Sample_Submitted	date			
Sequencing_Provider	varchar(100)	X		X
Sequencing_Platform	varchar(100)	X		X
ProviderSeqDirName	varchar(200)			
fastqMatePair1	varchar(50)			
fastqMatePair2	varchar(50)			

ProjectSeqDirName	varchar(200)				
Project	varchar(50)	X			X
Received	varchar(3)		X		
Run_Plexing	varchar(100)				
Primer_Set	varchar(30)				
Barcode_Type	varchar(20)				
Barcode	varchar(200)				
Library_Preparation_ID_req	varchar(100)				
Primer_Name	varchar(200)				
Primer_Seq	varchar(200)				
Date_Run_Performed	date				
Run_Number	int(11)				
Date_Data_Received	date				
Total_Reads_Millions	float(7,2)				
Read_Type	varchar(100)				X
Read_Length_Bases	int(11)				X
Total_Bases_Billions	int(11)				
Experimentalists_notes_upload	varchar(100)				
fileuploaded	mediumblob				
filetype	varchar(50)				
Remarks	varchar(100)				
Server_Results_Folder	varchar(100)				
createdby	varchar(20)		X		
timecreation	timestamp		X		
modifiedby	varchar(20)		X		
timemodified	timestamp		X		
track_id	int(11)		X		
track_sequence	int(11)		X		
track_exp	int(11)		X		