Form Approved Through 06/30/2012 OMB No. 0925-0001

Department of Health and Human Services
Public Health Service

Grant Application

Do not exceed character length restrictions indicated.

LEAVE BLANK-FOR PHS USE ONLY.		
Type	Activity	Number
Review Group		Formerly
Council/Board (Month, Year)		Date Received

1. TITLE OF PROJECT *(Do not exceed 81 characters, including spaces and punctuation.)*
A Data Coordinating Center for ENCODE

2. RESPONSE TO SPECIFIC REQUEST FOR APPLICATIONS OR PROGRAM ANNOUNCEMENT OR SOLICITATION □ NO ☒ YES
(If "Yes," state number and title)
Number: RFA-HG-11-026 Title: Data Analysis and Coordination Center for the Encyclopedia of DNA Elements (ENCODE)

3. PROGRAM DIRECTOR/PRINCIPAL INVESTIGATOR

3a. NAME *(Last, first, middle)*
Cherry, J. Michael

3b. DEGREE(S)
Ph.D

3h. eRA Commons User Name
CHERRY.MIKE

3c. POSITION TITLE
Associate Professor

3d. MAILING ADDRESS *(Street, city, state, zip code)*
Stanford University
School of Medicine
Department of Genetics
1501 S. California Ave, Rm 2419
Palo Alto, CA 94304-5577

3e. DEPARTMENT, SERVICE, LABORATORY, OR EQUIVALENT
Genetics

3f. MAJOR SUBDIVISION
School of Medicine

3g. TELEPHONE AND FAX *(Area code, number and extension)*
TEL: (650) 723-7541 FAX:

E-MAIL ADDRESS:
cherry@stanford.edu

4. HUMAN SUBJECTS RESEARCH
☒ No □ Yes

4a. Research Exempt
☒ No □ Yes

If "Yes", Exemption No.

4b. Federal-Wide Assurance No.
00000935

4c. Clinical Trial
☒ No □ Yes

4d. NIH-defined Phase III Clinical Trial
☒ No □ Yes

5. VERTEBRATE ANIMALS ☒ No □ Yes

5a. Animal Welfare Assurance No. A3213-01

6. DATES OF PROPOSED PERIOD OF SUPPORT *(month, day, year-MM/DD/YY)*
From 07/01/12 Through 06/30/16

7. COSTS REQUESTED FOR INITIAL BUDGET PERIOD
7a. Direct Costs ($) 4,185,339
7b. Total Costs ($) 4,999,903

8. COSTS REQUESTED FOR PROPOSED PERIOD OF SUPPORT
8a. Direct Costs ($) 17,709,052
8b. Total Costs ($) 20,980,779

9. APPLICANT ORGANIZATION
Name Board of Trustees of the
Address Leland Stanford Junior University
Research Management Group
301 Ravenswood Ave., 2nd Floor

Menlo Park, CA 94025-3434

10. TYPE OF ORGANIZATION
Public: → □ Federal □ State □ Local
Private: → ☒ Private Nonprofit
For-profit: → □ General □ Small Business
□ Woman-owned □ Socially and Economically Disadvantaged

11. ENTITY IDENTIFICATION NUMBER
1941156365A1

DUNS NO. 009214214 Cong. District CA-014

12. ADMINISTRATIVE OFFICIAL TO BE NOTIFIED IF AWARD IS MADE
Name Uuve Sauga
Title Research Process Manager
Address Stanford University
Research Management Group
301 Ravenswood Ave., 2nd Floor
Menlo Park, CA 94025-3434
Tel (650) 736-0592 FAX (650) 498-5876
E-mail usauga@stanford.edu

13. OFFICIAL SIGNING FOR APPLICANT ORGANIZATION
Name Mary Palmer
Title Research Process Manager
Address Stanford University
Research Management Group
301 Ravenswood Ave., 2nd Floor
Menlo Park, CA 94025-3434
Tel (650) 725-3991 FAX (650) 498-5876
E-Mail mary.palmer@stanford.edu

14. APPLICANT ORGANIZATION CERTIFICATION AND ACCEPTANCE: I certify that the statements herein are true, complete and accurate to the best of my knowledge, and accept the obligation to comply with Public Health Services terms and conditions if a grant is awarded as a result of this application. I am aware that any false, fictitious, or fraudulent statements or claims may subject me to criminal, civil, or administrative penalties.

SIGNATURE OF OFFICIAL NAMED IN 13.
(In ink."Per" signature not acceptable.)
Mary Palmer

DATE
12/13/11

Program Director/Principal Investigator (Last, First, Middle): Cherry, J. Michael

PROJECT SUMMARY (See instructions):

The goals of the ENCODE Data Coordinating (DCC) component to the ENCODE Database Coordination and Analysis Center are to support the ENCODE Consortium by defining and establishing pipelines that connect all participants to the data and by creating avenues of access that distribute these data to the greater biological research community. The ENCODE Consortium brings together laboratories that generate complex data types via experimental assays with laboratories that integrate these unique data using computational analyses to discover how chromosomal elements function together to define the human cell. The DCC's participation enhances the data created by these laboratories through the creation of structured pipelines for the verification and validation of all submitted data and providing processes for the documentation of metadata that describe each biological sample and assay method. To facilitate access to all the data created by the previous ENCODE projects as well as data from the modENCODE project and any other large data collections that are determined to be appropriate for incorporation, the DCC will construct a state of the art data storage repository called the Big Data Hub. The DCC will design and development new software to enhance the data submission and processing pipeline, the organization and access to metadata and the Big Data Hub. In addition, we will create the ENCODE Portal that will be the primary entry point to the wealth of experimentally determined information as well as results of computational analyses. The Portal will integrate these data resources and make them available via enhanced search and browsing capabilities. Tools will be implemented to aid discovery by both experienced bioinformaticians and naïve laboratory staff. The DCC will evolve into a substantial service organization allowing biomedical research to take full advance of the ENCODE results. To this end the DCC will provide documentation via many media including written documentation, video tutorials, webinars, and meeting presentations. The DCC, DAC, and AWG will be tightly woven together to create the EDCAC.

RELEVANCE (See instructions):

The relevance of this work for public health is that the comprehensive determination of functional elements encoded by the human genome is essential for understanding the nature of human health and the treatment of disease.

PROJECT/PERFORMANCE SITE(S) (if additional space is needed, use Project/Performance Site Format Page)

Project/Performance Site Primary Location

Organizational Name: Stanford University

DUNS: 009214214

Street 1: 1501 S. California Avenue
Street 2:

City: Palo Alto
County: Santa Clara
State: CA

Province:
Country: USA
Zip/Postal Code: 93404-5120

Project/Performance Site Congressional Districts: CA-014

Additional Project/Performance Site Location

Organizational Name: * see additional performance sites on page 5

DUNS:

Street 1:
Street 2:

City:
County:
State:

Province:
Country:
Zip/Postal Code:

Project/Performance Site Congressional Districts:

SENIOR/KEY PERSONNEL. See instructions. *Use continuation pages as needed* to provide the required information in the format shown below. Start with Program Director(s)/Principal Investigator(s). List all other senior/key personnel in alphabetical order, last name first.

Name	eRA Commons User Name	Organization	Role on Project
J. Michael Cherry	cherry.mike	Stanford University	PI
W. James Kent	jkent123	UC, Santa Cruz	co-Investigator
James Taylor	jpt.nyu	Emory University	co-Investigator
Ewan Birney	ebirney	EBI	co-Investigator
Kate Rosenbloom		UC, Santa Cruz	Sr. Project Eng.
Eurie L. Hong		Stanford University	Sr. Data Wrangler
Benjamin C. Hitz		Stanford University	Sr. Scientific Prog.

OTHER SIGNIFICANT CONTRIBUTORS

Name	Organization	Role on Project

Human Embryonic Stem Cells ☒ **No** □ **Yes**

If the proposed project involves human embryonic stem cells, list below the registration number of the specific cell line(s) from the following list: http://stemcells.nih.gov/research/registry/eligibilityCriteria.asp. *Use continuation pages as needed.*

If a specific line cannot be referenced at this time, include a statement that one from the Registry will be used.

Cell Line

Number the *following* pages consecutively throughout the application. Do not use suffixes such as 4a, 4b.

Use only if additional space is needed to list additional project/performance sites.

Additional Project/Performance Site Location

Organizational Name: University of California, Santa Cruz

DUNS: 125084723

Street 1: 1156 High Street | Street 2:

City: Santa Cruz | County: Santa Cruz | State: CA

Province: | Country: USA | Zip/Postal Code: 95064-1077

Project/Performance Site Congressional Districts: CA-017

Additional Project/Performance Site Location

Organizational Name: Emory University

DUNS: 066469933

Street 1: 1510 Clifton Road NE | Street 2:

City: Atlanta | County: Dekalb | State: GA

Province: | Country: USA | Zip/Postal Code: 30322-4218

Project/Performance Site Congressional Districts: GA-005

Additional Project/Performance Site Location

Organizational Name: European Molecular Biological Laboratory

DUNS: 321691735

Street 1: European Bioinformatics Institute | Street 2: Wellcome Trust Genome Campus

City: Hinxton, Cambridge | County: Cambridgeshire | State:

Province: | Country: United Kingdom | Zip/Postal Code: CB10 1SD

Project/Performance Site Congressional Districts: 00-000

Additional Project/Performance Site Location

Organizational Name:

DUNS:

Street 1: | Street 2:

City: | County: | State:

Province: | Country: | Zip/Postal Code:

Project/Performance Site Congressional Districts:

Additional Project/Performance Site Location

Organizational Name:

DUNS:

Street 1: | Street 2:

City: | County: | State:

Province: | Country: | Zip/Postal Code:

Project/Performance Site Congressional Districts:

RESEARCH GRANT

TABLE OF CONTENTS

Page Numbers

* Follow the page limits for these sections indicated in the application instructions, unless the Funding Opportunity Announcement specifies otherwise.

DETAILED BUDGET FOR INITIAL BUDGET PERIOD DIRECT COSTS ONLY

	FROM	THROUGH
	07/01/12	06/30/13

List PERSONNEL *(Applicant organization only)*
Use Cal, Acad, or Summer to Enter Months Devoted to Project
Enter Dollar Amounts Requested *(omit cents)* for Salary Requested and Fringe Benefits

NAME	ROLE ON PROJECT	Cal. Mnths	Acad. Mnths	Summer Mnths	INST.BASE SALARY	SALARY REQUESTED	FRINGE BENEFITS	TOTAL
J. Michael Cherry	PI	4.80				[redacted]	[redacted]	[redacted]
Eurie Hong	Lead Data Wrangler	12.00				[redacted]	[redacted]	[redacted]
Esther Chan	Data Wrangler	12.00				[redacted]	[redacted]	[redacted]
To Be Named	Data Wrangler	12.00				[redacted]	[redacted]	[redacted]
To Be Named	Data Wrangler	12.00				[redacted]	[redacted]	[redacted]
To Be Named	Data Wrangler	12.00				[redacted]	[redacted]	[redacted]
Gail Binkley	Database Administrator	3.00				[redacted]	[redacted]	[redacted]
Personnel continued on next page	Personnel TOTALS ►					945,887	294,172	1,240,059

CONSULTANT COSTS		-
EQUIPMENT *(Itemize)*		
Small Linux cluster of 12 nodes	$15,000	
Backup server and disk array	$35,000	
Servers for cutomized local analysis (2 x $5,000)	$10,000	60,000
SUPPLIES *(Itemize by category)*		
Desktop computers (7 x $3,000)	$21,000	
Hardware & software maintenance	$22,500	
Cloud backup via Stanford	$10,000	
Documentation, software, office supplies	$2,000	55,500
TRAVEL		
Domestic conferences (6x) @ $2,250	$13,500	
Lab & NCBI visits by wranglers, domestic (6x) @$1,250	$7,500	
Consortium meetings (8x) @$2,250	$18,000	
International conferences (4x) @$3,000	$12,000	
Lab & EBI visits by wranglers, international (3x) @$2,500	$7,500	58,500
PATIENT CARE COSTS	INPATIENT	-
	OUTPATIENT	-
ALTERATIONS AND RENOVATIONS *(Itemize by category)*		
OTHER EXPENSES *(Itemize by category)*		-
CONSORTIUM/CONTRACTUAL COSTS	DIRECT COSTS	1,957,968
SUBTOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD *(Item 7a, Face Page)*		$ 3,372,027
CONSORTIUM/CONTRACTUAL COSTS	FACILITIES AND ADMINISTRATION COSTS	813,312
TOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD		**$ 4,185,339**

DETAILED BUDGET FOR INITIAL BUDGET PERIOD DIRECT COSTS ONLY	FROM	THROUGH
	07/01/12	06/30/13

List PERSONNEL *(Applicant organization only)*
Use Cal, Acad, or Summer to Enter Months Devoted to Project
Enter Dollar Amounts Requested *(omit cents)* for Salary Requested and Fringe Benefits

NAME	ROLE ON PROJECT	Cal. Mnths	Acad. Mnths	Summer Mnths	INST.BASE SALARY	SALARY REQUESTED	FRINGE BENEFITS	TOTAL
To Be Named	Web Designer	12.00				[redacted]	[redacted]	[redacted]
To Be Named	Project Coordinator	12.00				[redacted]	[redacted]	[redacted]
To Be Named	Software Programmer	12.00				[redacted]	[redacted]	[redacted]
To Be Named	Software Programmer	12.00				[redacted]	[redacted]	[redacted]
Benjamin C. Hitz	Sr. Scientific Programmer	6.00				[redacted]	[redacted]	[redacted]
Matthew Simison	Systems Administrator	6.00				[redacted]	[redacted]	[redacted]
		-						
Personnel totals on previous page	**SUBTOTALS** -				►			

CONSULTANT COSTS		-
EQUIPMENT *(Itemize)*		
SUPPLIES *(Itemize by category)*		
TRAVEL		
PATIENT CARE COSTS	INPATIENT	-
	OUTPATIENT	-
ALTERATIONS AND RENOVATIONS *(Itemize by category)*		
OTHER EXPENSES *(Itemize by category)*		-
CONSORTIUM/CONTRACTUAL COSTS	DIRECT COSTS	
SUBTOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD *(Item 7a, Face Page)*		**$**
CONSORTIUM/CONTRACTUAL COSTS	FACILITIES AND ADMINISTRATION COSTS	
TOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD		**$**

BUDGET FOR ENTIRE PROPOSED PROJECT PERIOD
DIRECT COSTS ONLY

BUDGET CATEGORY TOTALS		INITIAL BUDGET PERIOD *(from Form Page 4)*	2nd ADDITIONAL YEAR OF SUPPORT REQUESTED	3rd ADDITIONAL YEAR OF SUPPORT REQUESTED	4th ADDITIONAL YEAR OF SUPPORT REQUESTED	5th ADDITIONAL YEAR OF SUPPORT REQUESTED
PERSONNEL: *Salary and fringe benefits. Applicant organization only.*		1,240,059	1,277,262	1,315,575	1,355,042	
CONSULTANT COSTS						
EQUIPMENT		60,000	30,000	45,000	20,000	
SUPPLIES		55,500	57,165	58,880	60,646	
TRAVEL		58,500	60,255	62,063	63,925	
PATIENT CARE COSTS	INPATIENT					
	OUTPATIENT					
ALTERATIONS AND RENOVATIONS						
OTHER EXPENSES						
CONSORTIUM / CONTRACTUAL COSTS	DIRECT	1,957,968	2,047,558	2,165,322	2,258,615	
SUBTOTAL DIRECT COSTS *(Sum = Item 8a, Face Page)*		3,372,027	3,472,240	3,646,840	3,758,228	
CONSORTIUM / CONTRACTUAL COSTS	F & A	813,312	846,258	882,412	917,735	
TOTAL DIRECT COSTS		4,185,339	4,318,498	4,529,252	4,675,963	
TOTAL DIRECT COSTS FOR ENTIRE PROPOSED PROJECT PERIOD						**$ 17,709,052**

JUSTIFICATION. Follow the budget justification instructions exactly. Use continuation pages as needed.

Budget Justification – Stanford University

PERSONNEL

Per our negotiated rate agreement with the Office of Naval Research for FY11, the budgeted salary amount for staff includes 8.8% vacation accrual/disability sick leave (DSL) for exempt employees and 7.6% for non-exempt employees. This amount does not exceed total salary. The vacation accrual/DSL rates will be charged at the time of the salary expenditure. No salary will be charged to the award when the employee is on vacation, disability or worker's compensation. The fringe benefit rate is 31.1% for faculty and staff, 19.8% for postdoctoral affiliates, 4.4% for graduate students, and 7.9% for temporary casual personnel. A cost of annual living increase of 3% was assumed for salaries, according to guidelines approved by Stanford University.

PRINCIPAL INVESTIGATOR:

J. Michael Cherry, Ph.D. Principal Investigator, 40% effort (4.8 calendar months). Dr. Cherry will be responsible for the day-to-day management of the DCC. In addition he is PI for the *Saccharomyces* Genome Database (SGD) and PI of the Gene Ontology Consortium (GOC) efforts at Stanford. The design, implementation, and scheduling of data and software projects are managed by Dr. Cherry. He is also responsible for the organization and maintenance of the Internet access systems, production databases and distribution media, and a variety of computing systems that provide services to the public and the project staff. We anticipate significant changes during the next four years and thus Dr. Cherry will maintain a flexible architecture and experienced staff to implement new features as needed. Dr. Cherry will integrate the DCC project staff within the Center for Genomics and Personalized Medicine in office space that is currently available for his use. The project management portion of this proposal describes in detail the process by which the Stanford and UCSC sites will be managed allowing the integration of both sites into one project.

DATA WRANGLERS:

The majority of the DCC project staff have the job title Data Wrangler. Much of the data validation, verification, and integration will be conducted using automatic software pipelines for fast and accurate report generation and data file archiving. However, there are a significant number of issues that need to be addressed manually. The current ENCODE and modENCODE DCC Data Wranglers require on average 1 FTE day for the complete integration of each dataset (data track). The Data Wranglers are scientists who have the technical skills to proficiently handle large-scale genomic datasets and have excellent communication skills that allow them to effectively communicate with staff at the production labs, the DCC, and the DAC. This expertise is also needed in addressing issues that are identified by advanced users, such as members of the AWG, and students needing assistance identifying data they need for a project. In our experience, staff members with a diverse skill set are an asset to a project such as this. In the course of working out issues with the data providers they become more knowledgeable in the experimental methods and biological samples used by each lab. They understand the interactions of the software components as they help test and troubleshoot the pipeline. This knowledge allows them to appropriately answer queries to the Help Desk, create outreach documentation and tutorials, and assist in the specification of software enhancements to both the Portal user interface and internal DCC applications. Another critical component of the position is training our contacts at the data provider sites in the use of DCC pipeline software.

To provide continuity of the new DCC with the current DCC at UCSC we will allow the four current UCSC Wranglers to remain and work from Santa Cruz. However, the primary location for the Data Wranglers will be Stanford; Stanford will do any new hires and new Wranglers will be required to work on-site at Stanford. The Wranglers currently based in Santa Cruz will be required to travel to Stanford once a week to maintain a high level of consistency, build close working relationships with other Wranglers, have an in-depth understanding of operations, and provide candid assessment of issues through small face-to-face discussions and meetings with the PI and Dr. Hong, the Lead for the Data Wrangler group. There are currently four Data Wranglers working from UCSC; Cricket Sloan, Venkat Malladi, Matt Wong and Dr. Ruihua Fang. The Stanford staff has already developed a rapport with the existing ENCODE Data Wranglers.

Eurie L. Hong, Ph.D. Lead Data Wrangler, 100% effort (12.0 calendar months). Dr. Hong will direct this critical and invaluable component of the day-to-day operations of the DCC. She will be responsible for managing the daily work prioritization of the Data Wrangler staff both at Stanford and Santa Cruz. The prioritization of tasks will be preformed with consultation with Dr. Cherry. Dr. Hong's responsibilities will include assessing the progress of the Data Wranglers, identifying issues with the manual tasks, and collaborating with the engineering staff to specify new tools that would enhance productivity. Dr. Hong is fluent in all aspects of the data management for the DCC. She has a complete understanding not only of the meta-database schema but also what data or controlled vocabulary is used in a specific column of a specific table. Dr. Hong is thus an essential person at all discussions of the interaction between software and the database, the editing tools and user interfaces that access the database, and of course the testing and troubleshooting for these software and database components. Dr. Hong will work with Drs. Cherry, Kent, Hitz, and Rosenbloom to define interfaces including the Portal web site and chromosomal region summary pages. She will work with the Quality Assurance Engineers to help define the tests and procedures they use when processing the ENCODE data. Dr. Hong has proven herself to be an outstanding group leader directing the 10 person biocurator group at SGD, setting priorities, defining processes for literature curation, and working with Drs. Cherry and Hitz and Ms. Binkley to coordinate the development, testing, and release of new interfaces and software at SGD. Dr. Hong is a Senior Research Scientist at Stanford University School of Medicine.

Esther Chan, Ph.D. Data Wrangler, 100% effort (12.0 calendar months). Dr. Chan will be a Data Wrangler and work with the production laboratories to define metadata, verify datasets, resolve problems at many levels, and be a member of the DCC Help Desk. Data Wranglers also have basic programming skills, often in Perl or Python, to allow them to manipulate file formats and to explore issues present within a submitted data file. Dr. Chan has three years experience working on high-throughput datasets with the SGD project where she built a pipeline for the evaluation and incorporation of published datasets into the SGD databases. She defined the specifications for the processing several HTP data types including RNA-seq, ChIP-seq, DNase-chip, FAIRE among others. Dr. Chan obtained her Ph.D. from the University of Toronto with Dr. Timothy Hughes in 2010 and is an expert in the evolutionary analysis of conserved genes and their expression in diverse vertebrate species. Her expertise includes analysis of data from microarray and next-gen sequencing technologies. She is skilled at computational techniques for the analysis of experimentally determined binding preferences of regulatory proteins from full genomic sequences.

Three TBN, Data Wrangler, 100% effort (12.0 calendar months). We will hire additional Data Wranglers that will have the same job description as Dr. Chan and the Data Wranglers currently located at UCSC. These new project staff will work with the production laboratories to define metadata, verify datasets, resolve problems at many levels, and be a member of the DCC Help Desk. They are expected to possess basic programming skills, Perl or Python, to allow them to manipulate file formats and to explore issues present within a submitted data file. They should have Ph.D. level training with experience working with high-throughput datasets such as RNA-seq, ChIP-seq, DNase-chip, FAIRE among others. They are critical components of the team that specifies the data and metadata submission pipelines. Their communication skills must be very high, as they must effectively work verbally and electronically with the staff at data production laboratories. The individuals will also communicate with the AWG (experienced bioinformaticians) and the general scientific community that contacts the DCC for assistance.

SCIENTIFIC PROGRAMMERS:

Programmers at SGD typically have some education in biology that allows them to successfully communicate with and understand the needs of the curation staff. Dr. Cherry is fortunate that three of the project's programming staff have Ph.D. level training in biology as well as appropriate training in software engineering. At Stanford we use the title Software Programmer while at UCSC they use the title Software Engineer. The difference in title only represents the difference between the Stanford School of Medicine and the UCSC School of Engineering. All the programming staff monitors the software issue tracking system Redmine with project coordination and prioritization handled by Dr. Hitz and the rest of the management staff. The programming staff participate in a weekly project meeting, as well as a separate programmers meeting. The majority of the software maintained by SGD is written in Perl and JavaScript, however there is a small amount written in Java, C, or PHP. Software used by the DCC will use Perl, Python, JavaScript and C. In addition to Web and User interface development, the programming team will create the metadata database application interface software and middleware to import and export the necessary files and web services needed by the project. The programming team works with Dr. Hong and the Data Wranglers to define processes and interfaces necessary to improve the productivity and fidelity of the data pipeline as operated by both the Wranglers and the data providers themselves. The USCS Genome Browser is the primary avenue for users to directly view the data created and analyzed by the entire ENCODE project, and the DCC programming team provides tools and extensions to the browser and ancillary search engines to enhance the user experience and handle data types yet-to-be-determined. The programmers and engineers at the DCC will work closely with Drs. Cherry, Kent, Hitz, Rosenbloom and Ms. Binkley on the interactions with the projects software and all appropriate databases at UCSC, Stanford and other sites participating in the ENCODE project. A Web Designer will work on the Portal web site as defined by Drs. Cherry and Hitz, in concert with Dr's Kent and Rosenbaum. The Web Designer position will work in Python and Perl and develop JavaScript, HTML and other aspects of the web site as needed.

Ben Hitz, Ph.D., Senior Scientific Programmer, 50% effort (6.0 calendar months). Dr. Hitz will manage the software engineering at Stanford. He also manages the software and database development and maintenance efforts for SGD and GOC, including the exchange of data between LBNL and Stanford for the GOC. Dr. Hitz is the chief architect of the web interface for SGD and brings this expertise to the GOC production environment. For SGD he manages the installation, interconnections, and testing of software developed by third-party sources, such as GBrowse, SPELL, InterMine, TextPresso, and Pathway Tools. He brings a wealth of technical scientific programming experience to the DCC. Drs. Hitz and Cherry supervise the efforts of the TBN Software Programmers and Web Designer. He will define the software architecture for the DCC project through coordination with Drs. Kent and Rosenbloom at UCSC. Drs. Cherry, Kent, Rosenbloom Hitz and Ms. Binkley have worked well together to specify the development of a new DCC submission pipeline and ENCODE Portal.

Two TBN, Software Programmer, 100% effort (12.0 calendar months). These Programmers will work with Dr. Hitz and the UCSC Software Engineering team to build the data submission pipeline, verification processes and data access via the ENCODE Portal. These individuals will have expertise in Perl, Python and C programming languages as well as relational databases and modern web service applications. Their tasks will be specified by Drs. Cherry and Kent and managed by Dr. Hitz.

TBN, Web Designer, 100% effort (12.0 calendar months). This web interface builder will implement JavaScript, CSS and Python components of the Portal web site. This person will also create forms to allow access to the real-time tracking system used by the consortium. Dr. Hitz will supervise the Web Designer. This person will work with the UCSC QA engineers to assess the user interface at the ENCODE Portal and make adjustments as needed.

DATABASE ADMINISTRATOR:

Gail Binkley, M.S., Principal Database Administrator, 25% effort (3.0 calendar months). Ms. Binkley is currently the DBA for all databases used for the SGD and GOC projects. She has more than eleven years experience maintaining RDBMS software including Oracle, MySQL and PostgreSQL. She is the primary responsible for database operations in Dr. Cherry's group and defines all systems level detail of the database environments. Ms. Binkley directs the data backups and maintains security and access control to the databases. Ms. Binkley works closely with Drs. Hitz and Mr. Simison to maintain 24/7 availability of all Internet resources. She will work closely with the UCSC engineering staff for the design, implementation, and maintenance of the metadata-processing pipeline. Ms. Binkley's skills include the careful specification of database environments, including the table schema, SQL triggers and reporting, software maintenance and file system configuration. She will also continue supervision of the system administrators and is involved with setting priorities within the DCC project as she currently does for the SGD and GOC projects.

SYSTEMS ADMINISTRATION:

Matthew Simison, Systems Administrator, 50% effort (6.0 calendar months). Mr. Simison is an experienced Unix Systems Administrator. He has over 12 years experience in systems administration. He has worked in several large computing facilities including Sun Microsystems, Netscape Communications, and Postini Corporation. He has experience in the administration of MySQL and PostgreSQL databases, Perl, Linux, TCP/IP networking and server components needed to maintain an integrated Internet resource with 24/7 availability. For the DCC he will be responsible for the daily operations of the Stanford computing infrastructure and web servers. Server software such as Perl, Python, Java and Tomcat will be maintained, in addition to software modules needed for applications used by the DCC Data Wranglers and any additional tools that are deemed necessary to provide the best resource possible for the project staff. Mr. Simison works closely with Mr. Stuart Miyasato the Senior Systems Administrator for Dr. Cherry's group, in addition to other computational staff on campus. He will work closely with the systems staff at UCSC on emergency planning in the event of small or large systems failures. The Stanford site currently operates with no single point of computing hardware failure. All hardware has multiple power supplies, network interfaces, disk controllers and disk drives. To enhance our reliability we are migrating some of our systems to different computing facilities. This minimizes downtime as the result of fire or power outages. A new campus computing facility will be located at the SLAC National Accelerator Laboratory site. SLAC as a DOE facility is on a separate power grid and thus does not typically loose power at the same time as the Stanford campus.

PROJECT SUPPORT:

TBN, Project Coordinator, 100% effort (12.0 calendar months). We will hire an individual to coordinate the meetings and reports required to make this a productive and smooth-running project. Many telephone conferences will be scheduled between the DCC staff, between the EDCAC staff, between the U54 PIs and the EDCAC, and with other Consortium members. Regular reporting will be prepared for the NIH as well as for the PI and co-PIs of the EDCAC. These reports will include tables and graphs constructed by the Portal software and will also require integration of text from the project staff. There will be considerable travel reimbursements to process through the administrative systems at Stanford. In addition to organizing electronic meetings the Stanford and UCSC components of the EDCAC will host meetings of the AWG and potentially workshops on the use of the ENCODE Portal. Coordination of printed outreach materials to be distributed at meetings and workshops, and electronic media including newsletters, posters, lecture presentations, videos and webinars will require a significant use of time.

EQUIPMENT: $60,000 YEAR 1; $30,000 YEAR 2; $45,000 YEAR 3; $20,000 YEAR 4

Equipment	
Small Linux cluster of 12 nodes	$15,000
Backup Server and Disk array	$35,000
Servers for customized local analysis (2 * $5000)	$10,000
Total Supplies	**$60,000**

A small Linux cluster built from rack mounted Dell computers is requested. This cluster will be an all-purpose resource that will serve many needs within the DCC. This RHEL cluster will provide a development environment for the new data and metadata submission pipelines. It will be used for a variety of tasks necessary for troubleshooting issues with submitted data and exploring possible solutions. Periodically other clusters available to the DCC will be used over capacity. Having a local set of machines that are available for significant tasks will enhance the group's successful operation. A large variety of important files and databases will be maintained by the DCC. This not only includes files making their way to the Big Data Hub but also source code repositories, tracking databases, development databases, and the multiple versions of files retained until a data file can be successfully processed. A robust and secure backup process will handle all these data. We use Simpana software from CommVault to backup files and databases to a Dell DL2200 22 TB disk array. The Simpana software reduces network traffic, minimizes the backup window required for the client, and reduces disk space on the backup server through de-duplication of redundant data. The cost of the server used by this backup server and the large amount of associated disk will cost $35,000.

In the second year, the cluster will be expanded, associated disk arrays and servers will be purchased. An additional high-end server will be purchased for jobs that require the fastest possible single CPU processor. In the third and fourth years we anticipate the disk requirements to increase, some of the servers will need to be replaced, and the total number of nodes to the cluster will be increased to 24. For years 2, 3 and 4 we request $30000, $45000 and $20000, respectively.

TRAVEL: TOTAL: $69,150. (DOMESTIC: $36,200, INTERNATIONAL: $30,150)

Domestic Travel	
Domestic Conferences (6 * $2250)	$13,500
Lab & NCBI visits by Wranglers, domestic (6 * $1250)	$7,500
Consortium Meeting (8 * $2250)	$18,000
Total	$39,000

International Travel	
International Conferences (4 * $3000)	$12,000
Lab & EBI visits by Wranglers, international (3 * $2500)	$7,500
Total	$19,500

The requested funds will allow the PI and Data Wranglers to attend domestic scientific meetings and workshops. We request funding for six trips at $2,250 each. Scientific conferences will educate the scientific research community on the use and usefulness of ENCODE for their research programs. Funds are also requested to allow Data Wranglers to visit production labs for training and troubleshooting, six at $1,250 each. These trips will be more important in the beginning of the project as the pipeline for the new DCC will require more operational sophistication by the production laboratory staff than currently required. Funds for eight members of the DCC are requests to attend the yearly ENCODE Consortium Meeting at $2,250 each. International meetings are also critical for the DCC to attend, four each for $3,000. This is to educate our international colleagues about the ENCODE data, how to obtain access and how others are using the data. Four trips are requested for travel to the EBI and international data providers, if any are funded, requesting three at $2,500. The travel budget is increased by 3% in each subsequent year.

SUPPLIES: $65,500

Supplies	
Desktop computers (7 * $3000)	$21,000
Hardware & Software Maintenance	$22,500
Cloud Backup via Stanford	$10,000
Documentation, Software, and office supplies	$2,000
Total Supplies	**$55,500**

We request funds to purchase seven (7) desktop computers (Apple iMac with second display $3000) for each of the new staff members. Maintenance is an important component of a resource such as the DCC. We request funds for both hardware and software maintenance and support. The high level of hardware support is appropriate to minimize downtime, even though we will have redundant services. We are also including funds for replacement of server and desktop disk drives, mouse, keyboards, and other small but significant items necessary for the functioning of a Bioinformatics Resources. A total of $22,500 is estimated to be required for all these items in the first year. We have stopped relying on tape backups. Tape is expensive and problematic. Oftentimes it is difficult to retrieve old files from tape, tapes can go bad or are defective from the start, and offsite tape storage for disaster recovery is costly and inconvenient for easy access to backed-up data. Thus we have determined it is both cost effective and expedient to only use disk and cloud backups. For desktop computers we used the Enterprise Carbonite solution. This service monitors specified directories and uploads changed files to a secure data cloud. We also use the Apple TimeMachine software that takes a snapshot of the disk every hour. TimeMachine provides immediate access to all previous files while Carbonite provides a remote copy of current files, providing disaster recovery. Files on the Unix servers are backed up to a separate system using software called Simpana from CommVault. This backup server has 22 TB of useable disk. Simpana provides immediate access to previous versions of all UNIX files. This software/hardware solution is also used to backup all of our databases. In addition, critical files and archived snapshots of the databases are backed up to a remote disk supported by the Stanford IT department. The cost of this service is based on the amount of disk space used. An additional feature of the IT department's remote disk storage is that it can be migrated to a data cloud, either a private or commercial cloud of our choice. Funds for these backup solutions are $10,000 in the first year. Finally, a variety of documentation, software, printer cartridges, printer paper, office supplies, other expendable items necessary for a production office environment are requested. Funds for these other supplies will be $2,000 in the first year. The supply budget is increased by 3% in each subsequent year.

As of August 11, 2011, the F&A rate used is the on-campus rate of 57%.

The PI has determined that this is a major project, as defined by OMB Circular A-21, and it meets A-21 requirements for direct charging of administrative expenses. All effort and expenses charged to this project will be for services specific to the project, and not for general support of the academic activities of the faculty or department. In addition, effort charged to this project can be specifically identified to the project.

DETAILED BUDGET FOR INITIAL BUDGET PERIOD DIRECT COSTS ONLY	FROM	THROUGH
	7/1/2012	6/30/2013

List PERSONNEL *(Applicant organization only)*
Use Cal, Acad, or Summer to Enter Months Devoted to Project
Enter Dollar Amounts Requested *(omit cents)* for Salary Requested and Fringe Benefits

NAME	ROLE ON PROJECT	Cal. Mnths	Acad. Mnths	Summer Mnths	INST.BASE SALARY	SALARY REQUESTED	FRINGE BENEFITS	TOTAL
W.J. Kent	PD/PI	2.4			[redacted]	[redacted]	[redacted]	[redacted]
K. Rosenbloom	Sr. Project Engineer	12			[redacted]	[redacted]	[redacted]	[redacted]
D. Karolchik	Program Manager	2.4			[redacted]	[redacted]	[redacted]	[redacted]
A. Zweig	Engineering Manager	1.2			[redacted]	[redacted]	[redacted]	[redacted]
V. Malladi	Data Wrangler	12			[redacted]	[redacted]	[redacted]	[redacted]
C. Sloan	Data Wrangler	12			[redacted]	[redacted]	[redacted]	[redacted]
M. Wong	Data Wrangler	12			[redacted]	[redacted]	[redacted]	[redacted]
R. Fang	Data Wrangler	12			[redacted]	[redacted]	[redacted]	[redacted]
B. Raney	Software Engineer	6			[redacted]	[redacted]	[redacted]	[redacted]
G. Barber	Software Engineer	6			[redacted]	[redacted]	[redacted]	[redacted]
T. Dreszer	Software Engineer	6			[redacted]	[redacted]	[redacted]	[redacted]
L. Meyer	Software Engineer	6			[redacted]	[redacted]	[redacted]	[redacted]
K. Learned	QA Engineer Lead	12			[redacted]	[redacted]	[redacted]	[redacted]
V. Swing	QA Engineer	12			[redacted]	[redacted]	[redacted]	[redacted]
TBH	QA Engineer	12			[redacted]	[redacted]	[redacted]	[redacted]
K. Hayden	Postdoctoral Scholar	12			[redacted]	[redacted]	[redacted]	[redacted]
G. Moro	System Administrator	12			[redacted]	[redacted]	[redacted]	[redacted]
TBD	Grad Student Researcher			1.5	[redacted]	[redacted]	[redacted]	[redacted]
TBD	Grad Student Researcher		4.5		[redacted]	[redacted]	[redacted]	[redacted]
TBD	Grant Admin.	12			[redacted]	[redacted]	[redacted]	[redacted]
Work study students	Admins	12			[redacted]	[redacted]	[redacted]	[redacted]
	SUBTOTALS →					1,000,680	313,020	1,313,701

CONSULTANT COSTS		
EQUIPMENT *(Itemize)* Central file server storage ($25,000); backup storage ($20,000); download server ($20,000); RAID disk enclosure (750TB/year) ($80,000); storage servers ($6,800); download servers ($8,000); GPFS MetaData servers ($11,500); network switches ($10,000); backup hardware ($8,000)		189,300
SUPPLIES *(Itemize by category)* Software contract for GPFS ($15,000); utility/monitor servers($950); firewall servers ($950); transfer servers ($3,200); GPFS manager node ($1,000); project specific materials and supplies ($10,000)		31,100
TRAVEL Domestic ($33,522); International ($23,450);		56,972
INPATIENT CARE COSTS		
OUTPATIENT CARE COSTS		
ALTERATIONS AND RENOVATIONS *(Itemize by category)*		
OTHER EXPENSES *(Itemize by category)* Co-location rental: racks ($6,500); UPS service ($1,400); generator service ($600); power use ($7,000) Grad student registration fees ($4,317) Publication page and open access costs ($4,000) Hosting research meetings ($500) new employee relocation costs ($2,000)		26,317
CONSORTIUM/CONTRACTUAL COSTS	DIRECT COSTS	
SUBTOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD *(Item 7a, Face Page)*		**$** 1,617,390
CONSORTIUM/CONTRACTUAL COSTS	FACILITIES AND ADMINISTRATIVE COSTS	725,261
TOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD		**$** 2,342,651

BUDGET FOR ENTIRE PROPOSED PROJECT PERIOD
DIRECT COSTS ONLY

BUDGET CATEGORY TOTALS	INITIAL BUDGET PERIOD *(from Form Page 4)*	2nd ADDITIONAL YEAR OF SUPPORT REQUESTED	3rd ADDITIONAL YEAR OF SUPPORT REQUESTED	4th ADDITIONAL YEAR OF SUPPORT REQUESTED	5th ADDITIONAL YEAR OF SUPPORT REQUESTED
PERSONNEL: *Salary and fringe benefits. Applicant organization only.*	1,313,701	1,370,801	1,430,161	11,491,877	
CONSULTANT COSTS					
EQUIPMENT	189,300	214,300	239,300	264,300	
SUPPLIES	31,100	31,000	31,000	31,000	
TRAVEL	56,972	59,821	62,509	65,331	
INPATIENT CARE COSTS					
OUTPATIENT CARE COSTS					
ALTERATIONS AND RENOVATIONS					
OTHER EXPENSES	26,317	27,867	29,572	31,448	
DIRECT CONSORTIUM/ CONTRACTUAL COSTS					
SUBTOTAL DIRECT COSTS *(Sum = Item 8a, Face Page)*	1,617,390	1,703,890	1,792,643	1,884,056	
F&A CONSORTIUM/ CONTRACTUAL COSTS	725,261	756,135	788,090	821,326	
TOTAL DIRECT COSTS	2,342,651	2,460,025	2,580,733	2,705,382	
TOTAL DIRECT COSTS FOR ENTIRE PROPOSED PROJECT PERIOD					**$** 10,088,791

JUSTIFICATION. Follow the budget justification instructions exactly. Use continuation pages as needed.

We are requesting $1,617,390 of direct costs in year one, allocating approximately 81.2% to personnel, 11.7% for equipment, 1.9% for project specific supplies, 3.6% for travel, and 1.5% for other costs including graduate student fees, co-location rental costs, publications, project related supplies, and expenses of hosting project-related meetings.

Justification is provided on the following continuation pages.

UCSC ENCODE DCC Budget Justification

PERSONNEL: Direct costs for salary and benefits ($1.314M) account for approximately 80% of total direct costs on the proposed 2012-13 budget. The roles of each position in the UCSC ENCODE DCC are described below. See also the organizational structure and staff responsibilities of the research section of the main grant for more details of the individual roles. PI Kent will devote 20% effort to the project. DCC personnel will include 11.5 FTE of existing staff, 1.0 staff FTE to be hired, plus one postdoctoral scholar and one graduate student researcher. Salaries for existing personnel are projections of 3% over their current year salaries. Benefit rates are projected to increase 2% annually in years 2 through 4.

PI (Kent, 0.20 FTE): Assembles and hires ENCODE project team. Trains programmers. Develops and oversees software architecture. Performs scientific review. Coordinates between the ENCODE group and the Genome Browser group at UCSC. Coordinates import of data from previous grant cycle ENCODE DCC. Presents highlights of ENCODE project at selected conferences.

Senior Project Engineer (Rosenbloom, 1.0 FTE): In conjunction with colleagues at Stanford, designs and implements the ENCODE portal web site, including information aimed at biomedical end-users and online real-time reports for the funding agency. Coordinates reuse of software from previous grant cycle of ENCODE DCC. Serves as liaison between Stanford engineering group and UCSC engineering group. Develops training materials for data wranglers and production labs on DCC software.

Program Manager (Karolchik, 0.2 FTE): Provides personnel management for programming staff, including leading the hiring of new staff, insuring that new staff are trained in engineering best practices and standards that are applied across UCSC genome informatics projects, helping resolve communication problems and other conflicts among the staff, managing performance, and assembling yearly performance reviews.

Engineering Manager (Zweig 0.1 FTE): Works with ENCODE DCC project manager to coordinate management of engineering staff partly funded by other grants and to insure that software intended for integration into the Genome Browser undergoes a proper testing and release cycle.

Data Wrangler (Malladi, 1.0 FTE): Works with production labs to create data agreements, facilitate data submission, and troubleshoot problems during automated processing. Handles data generated on mouse genome.

Data Wrangler (Sloan, 1.0 FTE): Works with production labs to create data agreements, facilitate data submission, and troubleshoot problems during automated processing. Assists lead data wrangler Hong on project management. Provides consultation for production labs that need help in organizing internal data tracking.

Data Wrangler (Wong, 1.0 FTE): Works with production labs to create data agreements, facilitate data submission, and troubleshoot problems during automated processing. Creates scripts to help automate repetitive aspects of wrangling.

Data Wrangler (Fang, 1.0 FTE): Works with production labs to create data agreements, facilitate data submission, and troubleshoot problems during automated processing. Assists labs with creating readable documentation as needed.

Software Engineer (Meyer 0.5 FTE): Develops new UCSC Genome Browser displays to highlight the ENCODE data, focusing on tracks that synthesize information on regulatory sequences. Integrates third-party tools such as dynamic node/edge graphs into DCC web works. Because much of the software developed for this project will be useful in other contexts at UCSC as well, Larry is funded 50% by this grant and 50% by the UCSC Genome Browser grant.

Software Engineer (Raney 0.5 FTE): Develops new UCSC Genome Browser displays to highlight the ENCODE data, focusing in particular on 5C, ChiaPet, and other data that bring together regions that may be widely separated and cross chromosomes, and displays that involve comparisons between multiple species. Because much of the software developed for this project will be useful in other contexts at UCSC as well, Brian is funded 50% by this grant and 50% by the UCSC Genome Browser grant.

Software Engineer (Dreszer 0.5 FTE): Provides JavaScript programming expertise. Improves search tools and other tools for locating relevant data among the thousands of ENCODE experiments. Develops tools to help integrate data into archives such as those at NCBI. Because much of the software developed for this project will be useful in other contexts at UCSC as well, Tim is funded 50% by this grant and 50% by the UCSC Genome Browser grant.

Software Engineer (Barber 0.5 FTE): Extends UCSC Track Data Hubs to accommodate new types of data generated by ENCODE. Develops and integrates methods for fast parallel data transfer between production labs, the DCC main data storage at the UCSD SDSC, and other places that mirror the entire ENCODE data set. Because much of the software developed for this project will be useful in other contexts at UCSC as well, Galt is funded 50% by this grant and 50% by the UCSC Genome Browser grant.

Quality Assurance Lead (Learned 1.0 FTE): Recruits, trains, and manages the performance of QA personnel. Tests software to insure it does not hang, crash, lose data or otherwise have problems. Reviews ENCODE data and documentation with a critical eye for accuracy, consistency, and completeness. Runs automated quality-checking tools and analyzes the results.

Quality Assurance Engineer (Swing 1.0 FTE): Tests software to insure it does not hang, crash, lose data or otherwise have problems. Reviews ENCODE data and documentation with a critical eye for accuracy, consistency, and completeness. Runs automated quality-checking tools and analyzes the results.

Quality Assurance Engineer (TBH 1.0 FTE): Tests software to insure it does not hang, crash, lose data or otherwise have problems. Reviews ENCODE data and documentation with a critical eye for accuracy, consistency, and completeness. Runs automated quality-checking tools and analyzes the results.

Post-doctoral Scholar (Karen Hayden, 1.0 FTE): Develops new algorithms for the display and analysis of ENCODE data, focusing primarily on mapping issues. Interfaces with the ENCODE Analysis Working Group. Facilitates communication between UCSC staff and ENCODE scientists working in these areas. Participates in data integration, validation, and presentation.

Grad Student Researcher (0.5 FTE): Computer graphics engineering student whose thesis project is to develop computer graphic displays for ENCODE data.

System Administrator (Gary Moro, 1.0 FTE): Builds and maintains computer clusters and networks. Installs workstations and software. Troubleshoots problems.

Grant Administration (1.0 FTE): (Various administrators that assist with grant management) Provide oversight and coordination of proposals, reporting, financial management, resources and personnel.

Work-study Students (TBH, 1.0 FTE): (Several part-time undergraduates) Perform basic clerical functions associated with ENCODE staff and space.

CAPITAL EQUIPMENT: Major equipment ($189,300) represents 8% of the total direct costs on this budget. As sequencing technology continues to improve and data generation accelerates, supporting the ENCODE project will require augmentation of our existing equipment resources. Data growth is estimated to double each year, which will require expansion of the ENCODE DCC file servers, disk storage, data backup facilities, and network capabilities. The cost per TB of expansion is estimated at approximately $1,000 for our central file server and approximately $800 for the other servers.

Our projected costs include:

- $25,000 -- growth of our central file server (“hive”)
- $20,000 -- additional backup capacity (“encodebackup”)
- $20,000 -- additional capacity for the download server (“hgdownload”)
- $80,000 -- RAID disk enclosures (750 TB per year)
- $6,800 -- storage servers
- $8,000 -- download servers
- $11,500 -- GPFS metadata servers
- $10,000 -- network switches
- $8,000 -- backup hardware (tape drive/tapes)

TRAVEL: Due to the collaborative needs of this project, significant travel is required for attendance at ENCODE administrative and scientific meetings, meetings between ENCODE DCC staff and contributing labs, and coordination between the UCSC and Stanford-based personnel. In the first year, funds are requested for the following travel (trip costs are projected to increase 5% annually in years 2 through 4):

- 2 roundtrips per week between UCSC and Stanford ($5,772@0.555/mi)
- 5 domestic trips to ENCODE–related scientific and collaborative meetings ($11,250)
- 4 international trips to ENCODE-related scientific and collaborative meetings ($13,400)
- 1 four-person trip to the ENCODE Consortium meeting in DC ($9,000)
- 6 domestic trips for lab wranglers ($1,250)
- 3 international trips for lab wranglers ($10,050)

OTHER:

Co-location rental costs: Due to the limited capacity for future growth in the UCSC campus computer room facilities, some of the ENCODE DCC equipment has been relocated to the UCSD SDSC colocation facility, which provides sufficient room for expansion and more reliable power and cooling. The colocation costs include a rack rental fee of $6,500 per rack (increasing 5% per year), UPS service (7 kW @ $200/kW) -- $1,400, generator service (1.5 kW @ $400/kW) -- $600, and power usage (7 kW @ 1000/kW) -- $7,000. Power costs are projected to increase 10% annually in years 2 through 4.

Central file server software: 50% of costs for software contract for GPFS ($15,000). Because this software is also used for the Genome Browser project, the cost is split between the 2 grants.

Minor Equipment: Includes funds for laptop and desktop workstations for 4 project staff ($10,000) and other project-related non-inventorial computing equipment: utility/monitor servers ($950), firewall servers ($950), transfer servers ($3,200) and GPFS manager node ($1,000).

Graduate Student Fees: Registration and health insurance fees projected at 10% and 5% (respectively) over FY11 rates, per UCSC-provided escalation factors.

Supplies: Miscellaneous project-related supplies and materials ($1,000).

Local meeting expenses: Meals/refreshments and other host expenses for ENCODE project-related meetings ($500).

The indirect cost rate is 51% according to the agreement with DHHS from 8/17/05.

DETAILED BUDGET FOR INITIAL BUDGET PERIOD DIRECT COSTS ONLY	FROM	THROUGH
	07/01/12	06/30/13

List PERSONNEL *(Applicant organization only)*
Use Cal, Acad, or Summer to Enter Months Devoted to Project
Enter Dollar Amounts Requested *(omit cents)* for Salary Requested and Fringe Benefits

NAME	ROLE ON PROJECT	Cal. Mnths	Acad. Mnths	Summer Mnths	INST.BASE SALARY	SALARY REQUESTED	FRINGE BENEFITS	TOTAL
Ewan Birney	PD/PI	1.8				0	0	0
TBA	Bioinformatician	12				[redacted]	[redacted]	[redacted]
TBA	Bioinformatician	12				[redacted]	[redacted]	[redacted]
SUBTOTALS →						163,724	24,826	188,550

CONSULTANT COSTS	
EQUIPMENT *(Itemize)*	
SUPPLIES *(Itemize by category)* Storage discs - 10 Fast I/O terabytes ($10,000) 2xLaptops ($3,869) Consumables ($565)	14,434
TRAVEL 3xTravels to US project meetings	8,222
INPATIENT CARE COSTS	
OUTPATIENT CARE COSTS	
ALTERATIONS AND RENOVATIONS *(Itemize by category)*	
OTHER EXPENSES *(Itemize by category)*	

CONSORTIUM/CONTRACTUAL COSTS	DIRECT COSTS	
SUBTOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD *(Item 7a, Face Page)*		**$** 211,206
CONSORTIUM/CONTRACTUAL COSTS	FACILITIES AND ADMINISTRATIVE COSTS	16,896
TOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD		**$** 228,102

Program Director/Principal Investigator (Last, First, Middle): Cherry, J. Michael

BUDGET FOR ENTIRE PROPOSED PROJECT PERIOD
DIRECT COSTS ONLY

BUDGET CATEGORY TOTALS	INITIAL BUDGET PERIOD *(from Form Page 4)*	2nd ADDITIONAL YEAR OF SUPPORT REQUESTED	3rd ADDITIONAL YEAR OF SUPPORT REQUESTED	4th ADDITIONAL YEAR OF SUPPORT REQUESTED	5th ADDITIONAL YEAR OF SUPPORT REQUESTED
PERSONNEL: *Salary and fringe benefits. Applicant organization only.*	188,550	201,628	212,772	224,404	
CONSULTANT COSTS					
EQUIPMENT					
SUPPLIES	14,434	564	14,434	564	
TRAVEL	8,222	8,222	8,222	8,222	
INPATIENT CARE COSTS					
OUTPATIENT CARE COSTS					
ALTERATIONS AND RENOVATIONS					
OTHER EXPENSES					
DIRECT CONSORTIUM/ CONTRACTUAL COSTS					
SUBTOTAL DIRECT COSTS *(Sum = Item 8a, Face Page)*	211,206	210,414	235,428	233,190	
F&A CONSORTIUM/ CONTRACTUAL COSTS	16,896	16,833	18,834	18,656	
TOTAL DIRECT COSTS	228,102	227,247	254,262	251,846	
TOTAL DIRECT COSTS FOR ENTIRE PROPOSED PROJECT PERIOD					**$** 961,457

JUSTIFICATION. Follow the budget justification instructions exactly. Use continuation pages as needed.

Budget Justification for EBI.

Personnel

Ewan Birney

The PI, Ewan Birney, will contribute oversight, experience and ideas. He will dedicated 0.6 Calendar months to this, and no requests are made for his funds.

1 Senior Engineer, Computational Background Grade 6, EBI

The EBI requests the budget of a skilled bioinformatics engineer with experience in both large scale data processing and solid engineering credentials over the project timescale. The individual will have an extensive programming background, with a track record in working in large-scale data volumes in genomics. (EMBL Grade 6 - 12 cal. months)

1 Senior Engineer, Biological Background Grade 6, EBI

The EBI requests the budget of a skilled bioinformatics engineer with experience in both large scale data processing and a good understanding of biological cell types over the project timescale. The individual will have a cell biology background , and ideally have a track record in working in large scale data volumes in genomics. (EMBL Grade 6 - 12 cal. months)

The EBI will provide the necessary large-scale disk and compute resources for exploring large dataset requirements.

The IDC rate 8% for EBI as for a foreign organization is determined by the NOT-OD-01-028.

DETAILED BUDGET FOR INITIAL BUDGET PERIOD DIRECT COSTS ONLY	FROM	THROUGH
	07/01/12	06/30/13

List PERSONNEL *(Applicant organization only)*
Use Cal, Acad, or Summer to Enter Months Devoted to Project
Enter Dollar Amounts Requested *(omit cents)* for Salary Requested and Fringe Benefits

NAME	ROLE ON PROJECT	Cal. Mnths	Acad. Mnths	Summer Mnths	INST.BASE SALARY	SALARY REQUESTED	FRINGE BENEFITS	TOTAL
James Taylor	PD/PI		.45		[redacted]	[redacted]	[redacted]	
James Taylor	PD/PI			.15	[redacted]	[redacted]	[redacted]	[redacted]
TBD	Engineer	12			[redacted]	[redacted]	[redacted]	[redacted]
SUBTOTALS →						101,628	27,744	129,372

CONSULTANT COSTS	
EQUIPMENT *(Itemize)*	
SUPPLIES *(Itemize by category)*	
TRAVEL	
INPATIENT CARE COSTS	
OUTPATIENT CARE COSTS	
ALTERATIONS AND RENOVATIONS *(Itemize by category)*	
OTHER EXPENSES *(Itemize by category)*	

CONSORTIUM/CONTRACTUAL COSTS	DIRECT COSTS	
SUBTOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD *(Item 7a, Face Page)*		**$** 129,372
CONSORTIUM/CONTRACTUAL COSTS	FACILITIES AND ADMINISTRATIVE COSTS	71,155
TOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD		**$** 200,527

BUDGET FOR ENTIRE PROPOSED PROJECT PERIOD
DIRECT COSTS ONLY

BUDGET CATEGORY TOTALS	INITIAL BUDGET PERIOD *(from Form Page 4)*	2nd ADDITIONAL YEAR OF SUPPORT REQUESTED	3rd ADDITIONAL YEAR OF SUPPORT REQUESTED	4th ADDITIONAL YEAR OF SUPPORT REQUESTED	5th ADDITIONAL YEAR OF SUPPORT REQUESTED
PERSONNEL: *Salary and fringe benefits. Applicant organization only.*	129,372	133,254	137,251	141,369	
CONSULTANT COSTS					
EQUIPMENT					
SUPPLIES					
TRAVEL					
INPATIENT CARE COSTS					
OUTPATIENT CARE COSTS					
ALTERATIONS AND RENOVATIONS					
OTHER EXPENSES					
DIRECT CONSORTIUM/ CONTRACTUAL COSTS					
SUBTOTAL DIRECT COSTS *(Sum = Item 8a, Face Page)*	129,372	133,254	137,251	141,369	
F&A CONSORTIUM/ CONTRACTUAL COSTS	71,155	73,290	75,488	77,753	
TOTAL DIRECT COSTS	129,372	133,254	137,251	141,369	
TOTAL DIRECT COSTS FOR ENTIRE PROPOSED PROJECT PERIOD					**$** 838,931

JUSTIFICATION. Follow the budget justification instructions exactly. Use continuation pages as needed.

Budget Justification Emory University

Personnel

James Taylor, Ph.D. (PI; 0.45 Academic and 0.15 summer months)

Dr. Taylor will supervise the Emory portion of the proposed project, will oversee software design and implementation, as well as work with collaborators to develop analysis approaches for driving biological projects. In addition, Dr. Taylor will contribute to the development of training materials, documentation and preparation of manuscripts.

Software Engineer (12 Calendar months)

This person will deploy and maintain a Galaxy instance for access to and analysis of ENCODE data, including integration of new analysis tools as relevant, and work with collaborators to enable direct integration with DCC database(s) and browser(s).

The indirect cost rate is 55% according to the agreement with DHHS from 8/24/11.

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME J. Michael Cherry	POSITION TITLE Associate Professor (Research)
eRA COMMONS USER NAME (credential, e.g., agency login) cherry.mike	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Purdue University, West Lafayette, Indiana	B.S.	1979	Biological Sciences
Purdue University, West Lafayette, Indiana	B.S.	1979	Biochemistry
University of California, Berkeley, California	Ph.D.	1985	Molecular Biology
Harvard University, Cambridge, Massachusetts		1985-1988	Genetics

A. Personal Statement

The primary focus of my research is the integration of experimental results into an environment that promotes access and enhances biomedical discovery through the creation of databases, tools and web interfaces. My group is accomplished at process of manual literature curation, including techniques for finding and extracting the fine details of published experimental results into a robust database of biological information. My experience includes the analysis of high-throughput datasets and the use of high quality curated information to annotation these results. Specifically my group provides bioinformatic resources that serve biological research communities with information and tools that provide service to the scientists and educators. My work is also focused on the development of the Gene Ontology a precise structured encyclopedia of terms used for consistent annotation of gene products and its application to curation of gene products and the subsequent public distribution via the Internet. I direct a group of Senior Biocurators, Bioinformatic Analysts and Scientific Programmers that serve the biological research and teaching communities via our public resources. I have been active in bioinformatics since 1990 and Director of the *Saccharomyces* Genome Database (SGD) a major model organism database since 1993. My current work facilitates research by others through our collection, curation, integration and dissemination of experimental results for *Saccharomyces cerevisiae*. My accomplishments over the past year, as detailed below, have been to maintain SGD and Gene Ontology as premier resources providing manually curated "gold standard" annotations. I continue to provide my expertise to other database projects as an advisor.

B. Positions and Honors.

Positions and Employment

1985 – 1988	Research Fellow in Genetics, Harvard University
1988 – 1993	Director of Computing, Molecular Biology, Massachusetts General Hospital
1991 – 1993	Research Associate in Genetics, Harvard University
1991 – 1993	Project Manager, AAtDB (An *Arabidopsis thaliana* Database). Massachusetts General Hospital & Department of Genetics, Harvard University (PI: Howard Goodman)
1993 – 1996	Head, Computing. Stanford DNA Sequence & Technology Center, NIH grant to Stanford University (PI: Ron Davis)
1993 – 2001	Chief Curator & Director, *Saccharomyces* Genome Database, NIH grant to Stanford University (PI: David Botstein)
1995 – 1999	Principal Investigator, *Arabidopsis thaliana* Database, NSF grant to Stanford University.
1998 – 2000	Director, Stanford Microarray Database, Stanford University, NIH grant to Stanford University (co-PIs: Pat Brown & David Botstein)

1999 – 2001	Co-Principal Investigator, *Arabidopsis* Functional Genomics Consortium, NSF grant to Michigan State University (PI: Pam Green)
2001 – 2012	Co-Principal Investigator, Gene Ontology Consortium, NIH grant to The Jackson Laboratory, Bar Harbor, Maine (PI: Judith Blake)
2001 –	Associate Professor (Research) Genetics, Stanford University School of Medicine
2001 –	Principal Investigator, *Saccharomyces* Genome Database, NIH grant to Stanford University
2003 – 2006	Principal investigator, *Tetrahymena* Genome Database, NIH grant to TIGR, (subcontract to Stanford; PI: Jonathan Eisen)
2009 – 2011	co-Principal Investigator, YeastMine at SGD, NIH grant to University of Cambridge, UK (PI: Gos Micklem)
2012 –	Principal Investigator, Gene Ontology Consortium, NIH grant to Jackson Laboratory (PI: Judith Blake; Multi-PI award)
2011 –	Principal Investigator, YeastMine at SGD, NIH grant to University of Cambridge, UK (PI: Gos Micklem; Multi-PI award)

Other Experience and Professional Memberships

2002 –	Member Advisory Committee, Wormbase, *Caenorhabditis* database, Caltech PI: Paul Sternberg
2002 – 2006	Member, NIH Genome Research Review Committee (GRRC)
2003 – 2005	Member of Scientific Advisory Group, The Blueprint Initiative, Toronto PI: Chris Hogue
2004 – 2007	Member of Scientific Advisory Board, dictyBase, *Dictyostelium* database, PI: Rex Chisholm
2004 – 2008	Member Advisory Committee, TIGR Rice Genome Annotation Project, PI: Robin Buell
2004 – 2008	Member Advisory Committee, EcoCyc Project at SRI. PI: Peter Karp.
2004 – 2008	Chair, Advisory Committee, Medicago Genome Sequencing, PI: Nevin Young, U. Minnesota
2005 – 2007	Scientific Advisory Panel, NHGRI ENCODE Project
2007 – 2011	External Consultants Panel, NHGRI for ENCODE and modENCODE Projects
2007 – 2009	Scientific Advisory Board, Integrative Biology Project, University of Toronto, PI: B. Andrews
2008 –	Scientific Advisory Board, FlyBase, Harvard University PI: W. Gelbart
2008 –	Member, NIH Genomics, Computational Biology and Technology study section (GCAT)
2010 –	Member Executive Board, International Society of Biocuration
2010 –	Associate Editor, Database - The Journal of Biological Databases and Curation
2011 –	Associate Editor, G3: Genes | Genomes | Genetics

C. Selected peer-reviewed publications (Selected from 76 peer-reviewed publications)

Most relevant to the current application

1. Chan, E.T. and Cherry, J.M. (2012) Considerations for creating and annotating the budding yeast Genome Map at SGD: a progress report. Database (Oxford), 10.1093/database/bar057. Epub ahead of print.
2. Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E.T., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hirschman, J.E., Hitz, B.C., Karra, K., Krieger, C.J., Miyasato, S.R., Nash, R.S., Park, J., Skrzypek, M.S., Simison, M., Weng, S., & Wong, E.D. (2011). *Saccharomyces* Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* 10.1093/nar/gkr1029. Epub ahead of print.
3. Balakrishnan, R., Park, J., Karra, K., Hitz, B.C., Binkley, G., Hong, E.L., Sullivan, J., Micklem, G., Cherry, J.M. (2012) YeastMine - An integrated data warehouse for *S. cerevisiae* data as a multi-purpose tool-kit. Database (Oxford), 10.1093/database/bar062. Epub ahead of print.
4. Gaudet, P., Bairoch, A., Field, D., Sansone, S.A., Taylor, C., Attwood, T.K., Bateman, A., Blake, J.A., Bult, C.J., Cherry, J.M., et al. Towards BioDBcore: a community-defined information specification for biological databases. Nucleic Acids Res. 2011; 39 (Database issue): D7-10
5. Rangarajan, A., Schedl, T., Yook, K., Chan, J., Haenel, S., Otis, L., Faelten, S., DePellegrin-Connelly, T., Isaacson, R., Skrzypek, M.S., Marygold, S.J., Stefancsik, R., Cherry, J.M., Sternberg, P.W., and Muller, H.M. (2011) Toward an interactive article: integrating journals and biological databases. BMC Bioinformatics. 12; 175.
6. Botstein, D. and J.M. Cherry. (1997) Molecular Linguistics: Extracting Information from gene and protein sequences. *Proc. Natl. Acad. Sci. USA* **94**:5506-5507 PMCID: PMC34160

7. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and G. Sherlock. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genetics* **25**:25-29 PMCID: PMC3037419
8. Gaudet, P., Chisholm, R., Berardini, T., Dimmer, E., Engel, S.R., Fey, P., Hill, D.P., Howe, D., Hu, J.C., Huntley, R., Khodiyar, V.K., Kishore, R., Li, D., Lovering, R.C., McCarthy, F., Ni, L., Petri, V., Siegele, D.A., Tweedie, S., Van Auken, K., Wood, V., Basu, S., Carbon, S., Dolan, M., Mungall, C.J., Dolinski, K., Thomas, P., Ashburner, M., Blake, J.A., Cherry, J.M., Lewis, S.E. (2009) The Gene Ontology's Reference Genome Project: A Unified Framework for Functional Annotation across Species. *PLoS Comput. Biol*. v.5(7): e1000431 PMCID: PMC2699109
9. Tian, W., Zhang, L.V., Tasan, M., Gibbons, F.D., King, O.D., Park, J., Wunderlich, Z., Cherry, J.M. and F.P. Roth. (2008) Combining guilt-by-association and guilt-by-profiling to predict *Saccharomyces cerevisiae* gene function. *Genome Biol*., **9**: Suppl 1:S7 PMCID: PMC2447541

Additional recent publications of importance to the field (in chronological order)

1. Weng, S., Dong, Q., Balakrishnan, R., Christie, K., Costanzo, M., Dolinski, K., Dwight, S.S., Engel, S., Fisk, D.G., Hong, E., Issel-Tarver, L., Sethuraman, A., Theesfeld, C., Andrada, R., Binkley, G., Lane, C., Schroeder, M., Botstein, D., Cherry, J.M. (2003) *Saccharomyces* Genome Database (SGD) provides biochemical and structural information for budding yeast proteins. *Nucleic Acids Res.* **31**:216-218 PMCID: PMC165501
2. Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M., and G. Sherlock. (2004) GO::TermFinder – open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*. **20**:3710-3715 PMCID: PMC3037731
3. Hirschman, J.E., Balakrishnan, R., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hong, E.L., Livstone, M.S., Nash, R., Park, J., Oughtred, R., Skrzypek, M., Starr, B., Theesfeld, C.L., Williams, J., Andrada, R., Binkley, G., Dong, Q., Lane, C., Miyasato, S., Sethuraman, A., Schroeder, M., Thanawala, M.K., Weng, S., Dolinski, K., Botstein, D. and J.M. Cherry. (2006) Genome Snapshot: a new resource at the *Saccharomyces* Genome Database (SGD) presenting an overview of the *Saccharomyces cerevisiae* genome. *Nucleic Acids Res* **34**:D442-D445 PMCID: PMC1347479
4. Nash, R., Weng, S., Hitz, B., Balakrishnan, R., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hirschman, J.E., Hong, E.L., Livstone, M.S., Oughtred, R., Park, J., Skrzypek, M., Theesfeld, C.L., Binkley, G., Dong, Q., Lane, C., Miyasato, S., Sethuraman, A., Schroeder, M., Dolinski, K., Botstein, D., and J.M. Cherry (2007) Expanded protein information at SGD: new pages and proteome browser. *Nucleic Acids Res*., **35**:D468-D471 PMCID: PMC1669759
5. Hong, E.L., Balakrishnan, R., Dong, Q., Christie, K.R., Park, J., Binkley, G., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hirschman, J.E., Hitz, B.C., Krieger, C.J., Livstone, M.S., Miyasato, S.R., Nash, R.S., Oughtred, R., Skrzypek, M.S., Weng, S., Wong, E.D., Zhu, K.K., Dolinski, K., Botstein, D., and J.M. Cherry. (2008) Gene Ontology Annotations at SGD: New Data Sources and Annotation Methods. *Nucleic Acids Res.,* **36**: Database Issue PMCID: PMC2238894
6. Costanzo, M.C., Skrzypek, M.S., Nash, R., Wong, E., Binkley, G., Engel, S.R., Hitz, B., Hong, E.L., and Cherry, J.M. (2009) New mutant phenotype data curation system in the *Saccharomyces* Genome Database. *Database, The Journal of Biol. Databases and Curation,* 2009: bap001 PMCID: PMC2790299
7. Christie, K.R., Hong, E.L., and Cherry, J.M. (2009) Functional annotations for the *Saccharomyces cerevisiae* genome: the knowns and the known unknowns. *Trends in Microbiology,* **17**:286-294 PubMed PMID: 19577472
8. Costanzo, M.C., Park, J., Balakrishnan, R., Cherry, J.M. and Hong, E.L. (2011) Using computational predictions to improve literature-based Gene Ontology annotations: a feasibility study. *Database, (Oxford)* bar004. PMCID: PMC3067894

D. Research Support

Ongoing Research Support

NIH 5 P41 HG001315 (PI: J.M. Cherry) 09/23/2011 – 02/29/2016
NIH NHGRI
Genomic Database for the Yeast *Saccharomyces*
The primary goal of this grant is to design, implement, and make generally available a database containing correlated information about the genome of the budding yeast *Saccharomyces cerevisiae*.
Role: Principal Investigator

5 P41 HG002273 (PI: J. Blake @ The Jackson Labs) 03/01/07 – 02/28/12
NIH NHGRI
Gene Ontology Consortium
The Gene Ontology (GO) Consortium is a collaboration among model organism database groups, genome annotation centers and others who are actively engaged in the annotation of genomes, genes and gene products. By sharing in the development and adoption of semantic standards, by contributing to a common data repository, and by disseminating our resources freely to the scientific community, we promote and enhance the ability of scientists to use all available information to further the understanding of human health and disease.
Role: Co-Investigator

1 R01 HG004834 (PI: J.M. Cherry; Contact PI: G. Micklem) 06/01/11-5/31/14
NIH NHGRI
InterMOD: integrated data and tools to support model organism research
Role: Principal Investigator

Completed Research Support

5 R01 GM67012 (PI: J. Eisen @ TIGR) 04/01/03 – 03/31/06
NIH NIGMS
Tetrahymena Genome Sequencing Project
The goal of this project is to create a Web accessible resource for the *Tetrahymena* community that uses the database and software resources developed by the *Saccharomyces* Genome Database (SGD). This will be accomplished by reusing the components of the SGD resource.
Role: PI of Stanford subcontract

5 R25 RR17381 (PI: E. Yuan @ The Tech Museum) 07/01/02 – 06/30/07
NIH NCRR
Life's New Frontier: Public Health Genetics
The primary goal is to generate and solicit innovative ideas for designing, maintaining and expanding the museum's current and future genetics exhibits, workshops, and program in ways that will appeal to a range of school age children as well as the general public.
Role: PI of Stanford subcontract

2 R01 HG002432 Boeke (PI) 09/06/06 – 06/30/09
NIH NHGRI
Genetic Interaction Map of Yeast
To facilitate deposition of budding yeast genetic interactions into SGD and international databases.
Role: Co-Investigator

1 R01 HG004834 Micklem (PI) 01/01/09-2/28/11
NIH NHGRI
Extending InterMine to yeast, rat and zebrafish model organism databases
Role: Co-Principal Investigator

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors in the order listed on Form Page 2.
Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME W. James Kent	POSITION TITLE Research Scientist Director, UCSC Genome Browser Project
eRA COMMONS USER NAME (credential, e.g., agency login) JKENT123	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
University of California, Santa Cruz	BA	1981	Mathematics
University of California, Santa Cruz	MA	1983	Mathematics
University of California, Santa Cruz	PhD	2002	Biology

NOTE: The Biographical Sketch may not exceed four pages. Follow the formats and instructions below.

A. Personal Statement

My experience as the PI in the last iteration of the ENCODE Data Coordinating Center (DCC) gives me a solid understanding of the issues that are likely to face the DCC in this iteration, and would help provide continuity between the old DCC and the new DCC. In this iteration I would be focusing on my core strengths of industrial strength software development, the visualization and interpretation of full genome data setsIn the last iteration of ENCODE, while these skills were important, an even larger part of the DCC effort turned out to be in the areas of data curation, and in helping the production labs successfully navigate the rather complex data and metadata requirements of ENCODE. It Is for this reason I'm stepping aside as PI for Michael Cherry, who's interest and experience in these areas is greater than my own.

Nonetheless the role of myself and my group at UCSC will be substantial. As the developer and in the next grant iteration the PI of the UCSC Genome Browser I am uniquely positioned to insure that the ENCODE DCC makes good use of that resource, and that that resource integrates data from ENCODE. My 16 years of consumer software development experience, followed by a PhD in Biology and an active research career give me an excellent background to develop scientific software and information resources in general. Outside of this resource project my own research interests are in, the regulation of mRNA transcription, the analysis of alternative and antisense transcripts, and comparative genomics – all areas of active and fruitful data production by ENCODE. My research has included substantial time in the wet lab as well as developing and applying tools for computational analysis, so I share some of the perspectives of the ENCODE data producers, both from my own direct experience, and from listening to them during the course of the previous grant cycle. From my many years of programming I have a good idea of how to build reliable software, and what is possible to build in a reasonable amount of time. From my management experience both in industry and in a research setting, I know how to recognize and develop software engineering skills in other people, and how to organize and motivate people to build something larger than one person could build alone. In summary I have the background and interest to successfully serve as the coPI on this important genomics resource project, particularly overseeing the software development efforts and providing continuity with the previous generation of the ENCODE DCC.

B. Positions and Honors

Positions and Employment

1983-1985	Software Engineer, Island Graphics Inc.
1985-1993	Owner and Principle Programmer, Dancing Flame Software
1993-1995	Software project manager, Autodesk Inc.
1995-1997	Owner and Principle Programmer, Dancing Flame Software
1998-2001	Computer Consultant
1999-2000	Biology Teaching Assistant, UC Santa Cruz
2002-2004	Assistant Research Scientist, UC Santa Cruz
2004-2008	Associate Research Scientist, UC Santa Cruz
2008-	Research Scientist, UC Santa Cruz
2008-	Associate Director, Center for Biomolecular Science & Engineering, UC Santa Cruz

Other Experience and Professional Memberships

2006-	Member WormBASE Scientific Advisory Board
2007-2009	Member ENSEMBL Scientific Advisory Board
2007-2010	Coorganizer (1 of 3) for Cold Spring Harbor/Wellcome Trust Genome Informatics Conference

Honors

1989	PC Magazine Award for Technical Excellence, Graphics Category (for Autodesk Animator)
2002	Benjamin Franklin Award for Openness in Bioinformatics, Bioinformatics.org
2002	GT All-Star, Genome Technology Magazine
2003	Tech Award Laureate, Health Category, San Jose Tech Museum of Innovation
2003	Overton Prize, International Society for Computational Biology
2009	Curt Stern Award, American Society of Human Genetics

C. Selected Peer-reviewed Publications (Selected from 61 peer reviewed publications)

NIH encourages applicants to limit the list of selected peer-reviewed publications or manuscripts in press to no more than 15. Do not include manuscripts submitted or in preparation. The individual may choose to include selected publications based on recency, importance to the field, and/or relevance to the proposed research. When citing articles that fall under the Public Access Policy, were authored or co-authored by the applicant and arose from NIH support, provide the NIH Manuscript Submission reference number (e.g., NIHMS97531) or the PubMed Central (PMC) reference number (e.g., PMCID234567) for each article. If the PMCID is not yet available because the Journal submits articles directly to PMC on behalf of their authors, indicate "PMC Journal - In Process." A list of these Journals is posted at: http://publicaccess.nih.gov/submit_process_journals.htm. Citations that are not covered by the Public Access Policy, but are publicly available in a free, online format may include URLs or PMCID numbers along with the full reference (note that copies of publicly available publications are not accepted as appendix material.)

1. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Research 2002;12:996-1006. PMC186604
2. Kent WJ. BLAT – the BLAST-like alignment tool. Genome Research 2002;12:656-664. PMC187518
3. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A. 2003 Sep 30;100(20):11484-9. PMC208784

4. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ. The UCSC Genome Browser Database. Nucleic Acids Res. 2003 Jan 1;31(1):51-4. PMC165576
5. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. The UCSC Table Browser data retrieval tool. Nucleic Acids Research 2004;32:D493-496. PMC308837
6. Kent WJ, Hsu F, Karolchik D, Kuhn RM, Clawson H, Trumbower H, Haussler D. Exploring relationships and mining data with the UCSC Gene Sorter. Genome Res. 2005 May;15(5):737-41. Bioinformatics. 2006 May 1;22(9):1036-46
7. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. The UCSC Known Genes.
8. Miller W, Rosenbloom K, Hardison RC, Hou M, Taylor J, Raney B, Burhans R, King DC, Baertsch R, Blankenberg D, Kosakovsky Pond SL, Nekrutenko A, Giardine B, Harris RS, Tyekucheva S, Diekhans M, Pringle TH, Murphy WJ, Lesk A, Weinstock GM, Lindblad-Toh K, Gibbs RA, Lander ES, Siepel A, Haussler D, Kent WJ. 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. Genome Res. 2007 Dec;17(12):1797-808. PMC2099589
9. Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, Kober KM, Miller W, Pedersen JS, Pohl A, Raney BJ, Rhead B, Rosenbloom KR, Smith KE, Stanke M, Thakkapallayil A, Trumbower H, Wang T, Zweig AS, Haussler D, Kent WJ. The UCSC Genome Browser Database: 2008 update. Nucleic Acids Res. 2008 Jan;36(Database issue):D773-9. PMC2238835
10. Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M, Meyer L, Hsu F, Hinrichs AS, Harte RA, Giardine B, Fujita P, Diekhans M, Dreszer T, Clawson H, Barber GP, Haussler D, Kent WJ. The UCSC Genome Browser Database: Update 20009. Nucleic Acids Res. 2009 Jan; 37(Database issue):D755-61. PMC2686463
11. Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, Pohl A, Pheasant M, Meyer LR, Learned K, Hsu F, Hillman-Jackson J, Harte RA, Giardine B, Dreszer TR, Clawson H, Barber GP, Haussler D, Kent WJ**.** The UCSC Genome Browser Database: Update 2010. Nucleic Acids Res. 2010 Jan;38(Database issue):D613-9. PMC2808870
12. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsering of large distributed datasets. Bioinformatics. 2010 Sep 1;26(17):2204-7. PMC2922891
13. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, Diekhans M, Dreszer TR, Giardine BM, Harte RA, Hillman-Jackson J, Hsu F, Kirkup V, Kuhn RM, Learned K, Li CH, Meyer LR, Pohl A, Raney BJ, Rosenbloom KR, Smith KE, Haussler D, Kent WJ. The UCSC Genome Browser Database: Update 2011. Nucleic Acids Res. 2011 Jan;39(Database issue):D876-82. PMC Journal – In Process.
14. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, Diekhans M, Dreszer TR, Giardine BM, Harte RA, Hillman-Jackson J, Hsu F, Kirkup V, Kuhn RM, Learned K, Li CH, Meyer LR, Pohl A, Raney BJ, Rosenbloom KR, Smith KE, Haussler D, Kent WJ. ENCODE whole-genome data in the UCSC genome browser (2011 update). Nucleic Acids Res. 2011 Jan;39(Database issue):D951-9. PMC3013705
15. ENCODE Project Consortium, Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, Bernstein BE, Gingeras TR, Kent WJ, Birney E, Wold B, Crawford GE. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 2011 Apr;9(4) PMC3079585

D. Research Support

List both selected ongoing and completed research projects for the past three years (Federal or non-Federally-supported). *Begin with the projects that are most relevant to the research proposed in the application.* Briefly indicate the overall goals of the projects and responsibilities of the key person identified on the Biographical Sketch. Do not include number of person months or direct costs.

2 P41 HG002371-08 (Haussler) 07/01/01- 07/31/12
National Human Genome Research Institute
Title: UCSC Center for Genomic Science and Minority Outreach Program
Project Goals: Genome annotation and web browser for the human and other genomes; outreach and user support; participation in the NISC Comparative Sequencing Program. Supplemental funding for a minority outreach program to increase involvement of underrepresented minorities in genome research.
Role: Co-PI with David Haussler

1 U41 HG004568-01 Kent (PI) 09/29/07- 06/30/11
National Human Genome Research Institute
Project Title: The UCSC ENCODE Data Coordination Center
Project Goals: Collect, organize, store, and provide access to data from the ENCODE project and other related projects through the UCSC Genome Browser.
Role: PI

1U41HG004269-01 Subcontract (L. Stein) 03/01/07- 02/28/10
NIH/NHGRI
Title: A Data Coordinating Center for ModENCODE
Project Goals: Assist the consortium in setting up and using the UCSC liftover service to map features to new builds; help create the database structures, parsers, and validators for multiple sequence alignment data; create genome-wide alignments of Caenorhabditis and/or Drosophila sequences if such alignments are not available; provide guidance on the submission workflow management system. Provide support as new data are added. Role: Co-PI with David Haussler

SC#20070152 Kent (PI) 11/20/06-11/19/09
Global Solutions for Infectious Disease
Subcontract title: GSID Clinical and Sequence Database Project
Subcontract Goals: Develop a relational database that will contain all significant clinical trial data from two Phase-III HIV vaccine clinical trials (VAX 003 and VAX 004) sponsored by VaxGen, Inc. Once developed, the database will be made available to the HIV vaccine academic and research community via the World Wide Web. Role: PI

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors in the order listed on Form Page 2.
Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME Ewan Birney	POSITION TITLE Team Leader, PANDA
eRA COMMONS USER NAME (credential, e.g., agency login) ebirney	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Balliol College, Oxford	BA	1992-1996	Biochemistry
St. John’s College, Cambridge	PhD	1996-2000	Genetics

Personal Statement.

I have had a long career in bioinformatics, starting from before I was an undergraduate where I programmed my first sequence analysis tools at Cold Spring Harbor the summer before going to Oxford. As well as algorithm production (Genewise, Velvet, CRAM) I have participated in many genomics projects – in particular the human genome, mouse genome and other genome projects. I was part of the original concept group for the ENCODE project, and lead the analysis of the both the pilot project (leading to a successful large Nature paper and many companion papers) and the first ENCODE scale up (leading to the submission of 8 Nature papers and >60 companion papers in November/December 2011). In April of 2012 I take on a larger scale, Associate Director role at the EBI, and do not have the capacity to play the same central analysis coordination role, as well as there being many talented individuals in the genomics community who can provide this leadership. However I do have an agreement to dedicate ~20% of my time on “personal” research projects and believe I can still offer an important personal contribution to ENCODE. I have worked with Dr Kent for over 10 years, often under high pressure, and have many co-authored publications; I have known Professor Cherry for around 8 years and although this would be our first co-grant, I have no concerns about working collaboratively with him.

A Positions and Honors.

2003: Awarded Francis Crick Lectureship from the Royal Society, UK (Prize for outstanding young molecular biology researcher)
2003: Promoted to Senior Scientists (equivalent of Senior Faculty) in EMBL
1999- Present: Head of DNA Data, European Bioinformatics Institute, Hinxton UK

B. Selected peer review publications
(153 Peer reviewed publications, 20 in Nature or Science, H-index 60, >40,000 Citations, >25,000 since 2006)

The ENCODE Project Consortium. Initial Analysis of the Encyclopedia of DNA Elements in the Human Genome (submitted November 2011)

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, Alkan C, Prüfer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Höber B, Höffner B, Siegemund M, Weihmann

A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PL, Eichler EE, Falush D, **Birney E**, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Pääbo S. A draft sequence of the Neandertal genome. Science. 2010 May 7;328(5979):710-22.

Atanur SS, Birol I, Guryev V, Hirst M, Hummel O, Morrissey C, Behmoaras J, Fernandez-Suarez XM, Johnson MD, McLaren WM, Patone G, Petretto E, Plessy C, Rockland KS, Rockland C, Saar K, Zhao Y, Carninci P, Flicek P, Kurtz T, Cuppen E, Pravenec M, Hubner N, Jones SJ, **Birney E**, Aitman TJ. The genome sequence of the spontaneously hypertensive rat: Analysis and functional significance. Genome Res. 2010 Jun;20(6):791-803.

McDaniell R, Lee BK, Song L, Liu Z, Boyle AP, Erdos MR, Scott LJ, Morken MA, Kucera KS, Battenhouse A, Keefe D, Collins FS, Willard HF, Lieb JD, Furey TS, Crawford GE, Iyer VR, **Birney E**. Heritable individual-specific and allele-specific chromatin signatures in humans. Science. 2010 Apr 9;328(5975):235-9.

Hoffman MM, **Birney E**. An effective model for natural selection in promoters. Genome Res. 2010 May;20(5):685-92

Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, Lin ML, Beare D, Lau KW, Greenman C, Varela I, Nik-Zainal S, Davies HR, Ordoñez GR, Mudie LJ, Latimer C, Edkins S, Stebbings L, Chen L, Jia M, Leroy C, Marshall J, Menzies A, Butler A, Teague JW, Mangion J, Sun YA, McLaughlin SF, Peckham HE, Tsung EF, Costa GL, Lee CC, Minna JD, Gazdar A, **Birney E**, Rhodes MD, McKernan KJ, Stratton MR, Futreal PA, Campbell PJ. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature. 2010 Jan 14;463(7278):184-90

Paten B, Herrero J, Fitzgerald S, Beal K, Flicek P, Holmes I**, Birney E**. Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome Res. 2008 18:1829-43.

Paten B, Herrero J, Beal K, Fitzgerald S, **Birney E**. Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Res. 2008 18:1814-28.

O'Reilly PF, **Birney E**, Balding DJ. Confounding between recombination and selection, and the Ped/Pop method for detecting selection. Genome Res. 2008 Aug;18(8):1304-13.

Zerbino DR, **Birney E.** Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 200818:821-9.

Warren etal (one of 54 authors) Genome analysis of the platypus reveals unique signatures of evolution Nature. 2008 May 8;453(7192):175-83.

Ettwiller L, Paten B, Ramialison M, **Birney E**, Wittbrodt J. Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. Nat Methods. 2007 Jul;4(7):563-5.

The ENCODE Project Consortium (**Lead Author** with 308 authors). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007 Jun 14;447(7146):799-816.

C. Research Support

Active

081979/Z/07/Z (PI: Birney) 07/01/07 – 06/30/12
Wellcome Trust, UK
The nomenclature of human genes
To provide a single naming authority for human genes to facilitate interchange between researchers

062023/A/00/B (PI: Birney) 10/01/06 – 09/30/11
Wellcome Trust, UK
Ensembl: a current, complete and consistent annotation of large scale genome sequence
Annotation of vertebrate Genomes, in particular Human and Mouse. Provision of a user friendly web interface and information infrastructure.

2P41HG003345-04 (PI: Birney) 09/19/07 – 06/30/12
National Institutes of Health
The Nomenclature of Human Genes.
To provide for a complete set of human readable names for each human gene.

1 P41 HG003751-01 A2 (Birney – co-Investigator) 04/01/07 – 02/28/2011
National Institutes of Health
Reactome: An Open Knowledgebase of Human Pathways.
To generate an accurate representation of known human molecular pathways.

5 U01 HG004695 (Birney – Coordinator) 04/01/09 – 03/31/10
National Institutes of Health
EDAC: ENCODE Data Analysis Center.
To provide a complete inventory of all functional elements of the human genome.

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors in the order listed on Form Page 2.
Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME James Taylor	POSITION TITLE Assistant Professor Department of Biology; Department of Math & Computer Science, Emory University
eRA COMMONS USER NAME (credential, e.g., agency login) jpt4.nyu	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
University of Vermont, Burlington, VT	B.S.	1997-2000	Computer Science
Penn State University, University Park, PA	Ph.D.	2003-2006	Computer Science

A. Personal Statement

I am currently one of the leaders of the Galaxy project, a framework for making data analysis tools accessible to the biomedical research community. Prior to my PhD work I spent several years in industry as a software engineer specializing in design and architecture. I applied this expertise as one of the original developers of Galaxy, and I continue to lead the software architecture and development aspects of the Galaxy project. Galaxy now has thousands of regular users who perform hundreds of thousands of analysis every month through its primary public instance, as well as dozens of other instances around the world. As a result, I have extensive experience with large-scale data management, particularly in the context of heterogeneous data intensive analysis. I have a long standing collaboration with Jim Kent to ensure tight integration between the UCSC genome browser and its associated databases with Galaxy, allowing users to analyze data stored in the UCSC database (including the current ENCODE data) through Galaxy, and display results in the context of other ENCODE derived annotation data (this work was published in the context of the ENCODE pilot; Blankenberg et al. 2007). Since 2009 my group has been working on ways to leverage "cloud computing" to make large-scale data analysis more accessible. We have developed a platform for dynamically scalable analysis on infrastructure cloud resources, which can be coupled to Galaxy for ease of use.

Outside of Galaxy, my research interests are in genomic and epigenomic aspects of the regulation of transcription, and my group uses comparative and functional genomic approaches to understand regulatory element architecture and chromatin organization. As part of my previous work in this area I was heavily involved in the analysis of the ENCODE pilot project, particularly in the areas of transcriptional regulation, comparative genomics, and variation. I am currently most involved in the Mouse focused activities of ENCODE, but have been a close observer of the integrated analysis of the Human data. As a result, I have a detailed understanding of how ENCODE data has been analyzed in the past, and the infrastructure necessary to support this analysis.

B. Positions and Honors.

1999-2001	Senior Software Engineer, The NYBOR Corporation
2001-2003	Senior Software Engineer, 4Lane Digital
2003-2006	Research Assistant, Penn State University
2006-2008	Visiting Member / Courant Instructor, New York University
2008-	Assistant Professor, Emory University

C. Selected peer-reviewed publications

Afgan E, Baker D†, Coraor C, Nekrutenko A, Taylor J. “Harnessing cloud-computing for biomedical research with Galaxy Cloud”, Nature Biotechnology. (in press)

Afgan E, Baker D, Coraor N, Chapman B, Nekrutenko A, Taylor J. “Galaxy CloudMan: delivering cloud compute clusters”. BMC Bioinformatics. 2011(Suppl 12):S4.

Goecks J, The Galaxy Team, Nekrutenko A, Taylor J. “A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences”. Genome Biology 2010 Sep; 11:R86.

Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. "Galaxy: a web-based genome analysis tool for experimentalists". Current Protocols in Molecular Biology. 2010 Jan; Chapter 19:Unit 19.10.1-21.

Koskovsky Pond S, Wadhawan S, Chiaromonte F, Ananda G, Chung W, Taylor J, and Nekrutenko A. "Windshield splatter analysis with the Galaxy metagenomic pipeline". Genome Research. 2009 Nov; 19(11):2144-53.

King DC, Taylor J, Zhang Y, Cheng Y, Lawson HA, Martin J, ENCODE groups for Transcriptional Regulation and Multispecies Alignment, Chiaromonte F, Miller W, Hardison RC. “Finding cis-regulatory elements using comparative genomics: some lessons from ENCODE data”. Genome Research. 2007 Jun; 17(6):775-86.

Blankenberg D, Taylor J, Schenk I, He J, Zhang Y, Ghent M, Veeraraghavan N, Albert I, Miller W, Makova K, Hardison RC, Nekrutenko A. “A framework for collaborative analysis of ENCODE data: Making large-scale analyses biologist-friendly”. Genome Research. 2007 Jun; 17(6):960-4.

The ENCODE Project Consortium. “Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project”. Nature. 2007 Jun 14;447(7146):799-816.

Taylor J, Tyekucheva S, King DC, Hardison RC, Miller W, and Chiaromonte F. “ESPERR: Learning strong and weak signals in genomic sequence alignments to identify functional elements”. Genome Research. 2006 Dec; 16(12):1596-604.

Taylor J, Tyekucheva S, Zody M, Chiaromonte F, Makova KD. “Strong and Weak Male Mutation Bias at Different Sites in the Primate Genomes: Insights from the Human-Chimpanzee Comparison”. Molecular Biology and Evolution. 2006 Mar; 23(3):565-573.

D. Research Support.

NSF DBI-0850103 (Nekrutenko; Taylor Co-PI) 8/1/2009-7/31/2012
Cyberinfrastructure for Accessible and Reproducible Research in Life Sciences

NIH R01HG004909 (Nekruntenko, Taylor Co-PI) 1/1/09-12/31/13
An efficient lightweight environment for biomedical computation
Development of workflow and library functionality within the Galaxy framework

NIH 2R01DK065806-06 (Hardison PD; Taylor MPI) 2/1/2009-3/31/2014
Global Predictions and Tests of Erythroid Regulation
Identify important cis elements responsible for erythroid gene expression and study their features through functional testing.

NIH 1R21HG005133-01 (Taylor PI) 7/1/2009-6/30/2011
A Turnkey Solution for Next Generation Sequence Data Analysis
Extending Galaxy to provide an integrated software framework and tools for analysis of data generated by high-throughput sequencing.

NIH 1RC2HG005542-01 (Taylor PD) 10/1/2009-9/30/2011
Dynamically scalable, accessible analysis for next generation sequence data
Creation of infrastructure for performing NGS analysis using cloud-computing resources.

NIH 1RC2HG005573-01 (Hardison PI; Taylor Co-PI) 10/1/2009-9/30/2011
Enhance human ENCODE by functional comparisons to mouse
Generate experimental datasets in mouse cell lines that complement those produced by ENCODE, perform integrated comparative analysis.

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors in the order listed on Form Page 2. Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME Eurie L. Hong	POSITION TITLE Senior Research Scientist
eRA COMMONS USER NAME (credential, e.g., agency login)	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Stanford University, Stanford, CA	B.S. with honors	06/1996	Biological Sciences
University of Chicago, Chicago, IL	Ph.D.	03/2002	Molecular Genetics and Cell Biology

A. Personal Statement

In this proposal, we describe the tasks and processes by which the Data Coordination component of the EDCAC will manage the data generated by the ENCODE project. This project will continue to maintain and develop pipelines, interfaces, and software that allow production labs to submit protocols, meta-data, and experimental results from high-throughput studies and allow the larger scientific community to query, browse, and view these data. My experiences at the *Saccharomyces* Genome Database (SGD) and with the GO Consortium (GOC) have prepared me to serve as Senior Data Wrangler of this project. Through various roles and responsibilities, my skills describing and analyzing scientific results using controlled vocabularies, integrating diverse types of data, defining and implementing novel work flows, leading teams to implement web-based interfaces and tools to view scientific data, and interacting with an international scientific community will be directly applicable to the tasks of a Sr. Data Wrangler with the ENCODE project.

As a Scientific Curator at the *Saccharomyces* Genome Database (SGD), I identified key scientific results from peer-reviewed literature in a wide range of biological topics and contributed to the development of major areas in the Gene Ontology (GO). During my time as Head Curator, I successfully led the development and implementation of novel literature curation workflows and web-based curation interfaces to review all literature published about *S. cerevisiae*. In addition to literature-based curation efforts, I collaborated with researchers to integrate their gene expression data and high-throughput transcription data at SGD. As Head Curator and as Co-Manager of User Advocacy with the GO Consortium (GOC), I worked closely with a team of scientific curators, database administrators, software engineers, and system administrators to prioritize and implement tasks during the development of new search and analysis tools. Currently as a Senior Research Scientist, I evaluate the accuracy of annotations generated using controlled vocabularies, which provides insight into how data can be better stored and modeled. These roles have trained me to quickly understand new fields of research and provided me with opportunities to collaborate with an international group of curators and scientists to define curation workflows to improve efficiency and accuracy as well as develop and implement new user interfaces.

B. Positions and Honors.

Positions and Employment

2002 – 2003 Scientific Curator, *Saccharomyces* Genome Database, Stanford University, Stanford, CA, NIH grant to Stanford University (PI: J. Michael Cherry)

2003 – 2004 Lead Scientific Curator, *Saccharomyces* Genome Database, Stanford University, Stanford, CA, NIH grant to Stanford University (PI: J. Michael Cherry)

2004 – 2010 Head Curator, *Saccharomyces* Genome Database, Stanford University, Stanford, CA, NIH grant to Stanford University (PI: J. Michael Cherry)

2006 – 2007 Co-Manager, User Advocacy, Gene Ontology Consortium, NIH grant to The Jackson Laboratory, Bar Harbor, Maine (PI: Judith Blake)

2010 – Senior Research Scientist, Department of Genetics, Stanford University, Stanford, CA

Other Experience and Professional Memberships

1999 – Participant, Howard Hughes Medical Institute Ask A Scientist program

2005 – Member, Genetics Society of America

2010 Area committee member, Databases & Ontologies, International Conference on Intelligent Systems for Molecular Biology

Honors

1997 Predoctoral Fellowship, National Science Foundation

1997 – 2002 Predoctoral Fellowship, Howard Hughes Medical Institute

C. Selected Peer-reviewed Publications (Selected from 20 peer-reviewed publications).

Most relevant to the current application

1. Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E.T., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hirschman, J.E., Hitz, B.C., Karra, K., Krieger, C.J., Miyasato, S.R., Nash, R.S., Park, J., Skrzypek, M.S., Simison, M., Weng, S., & Wong, E.D. (2011). *Saccharomyces* Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* Epub ahead of print.
2. Christie, K.R., Hong, E.L., & Cherry, J.M. (2009). Functional annotations for the *Saccharomyces cerevisiae* genome: the knowns and the known unknowns. *Trends Microbiol,* **17**(7), 286-294. PMCID: PMC3057094.
3. Hong, E.L., Balakrishnan, R., Dong, Q., Christie, K.R., Park, J., Binkley, G., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hirschman, J.E., Hitz, B.C., Krieger, C.J., Livstone, M.S., Miyasato, S.R., Nash, R.S., Oughtred, R., Skrzypek, M.S., Weng, S., Wong, E.D., Zhu, K.K., Dolinski, K., Botstein, D., & Cherry, J.M. (2008). Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res,* **36**(Database issue), D577-581. PMCID: PMC2238894.
4. Costanzo, M.C., Park, J., Balakrishnan, R., Cherry, J.M., & Hong, E.L. (2011) Using computational predictions to improve literature-based Gene Ontology annotations: a feasibility study. *Database (Oxford).* bar004. PMCID: PMC3067894.
5. Costanzo, M.C., Skrzypek, M.S., Nash, R., Wong, E., Binkley, G., Engel, S.R., Hitz, B., Hong, E.L., & Cherry, J.M. (2009). New mutant phenotype data curation system in the *Saccharomyces* Genome Database. *Database (Oxford),* bap001. PMCID: PMC2790299.

Additional recent publications of importance to the field (in chronological order)

1. Weng, S., Dong, Q., Balakrishnan, R., Christie, K., Costanzo, M., Dolinski, K., Dwight, S.S., Engel, S., Fisk, D.G., Hong, E., Issel-Tarver, L., Sethuraman, A., Theesfeld, C., Andrada, R., Binkley, G., Lane, C., Schroeder, M., Botstein, D., & Michael Cherry, J. (2003). *Saccharomyces* Genome Database (SGD) provides biochemical and structural information for budding yeast proteins. *Nucleic Acids Res,* **31**(1), 216-218. PMCID: PMC165501.

2. Christie, K.R., Weng, S., Balakrishnan, R., Costanzo, M.C., Dolinski, K., Dwight, S.S., Engel, S.R., Feierbach, B., Fisk, D.G., Hirschman, J.E., Hong, E.L., Issel-Tarver, L., Nash, R., Sethuraman, A., Starr, B., Theesfeld, C.L., Andrada, R., Binkley, G., Dong, Q., Lane, C., Schroeder, M., Botstein, D., & Cherry, J.M. (2004). Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. *Nucleic Acids Res,* **32**(Database issue), D311-314. PMCID: PMC308767.
3. Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G.M., Blake, J.A., Bult, C., Dolan, M., Drabkin, H., Eppig, J.T., Hill, D.P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J.M., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S., Fisk, D.G., Hirschman, J.E., Hong, E.L., Nash, R.S., Sethuraman, A., Theesfeld, C.L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S.Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E.M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., & White, R. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res,* **32**(Database issue), D258-261. PMCID: PMC308770.
4. Dwight, S.S., Balakrishnan, R., Christie, K.R., Costanzo, M.C., Dolinski, K., Engel, S.R., Feierbach, B., Fisk, D.G., Hirschman, J., Hong, E.L., Issel-Tarver, L., Nash, R.S., Sethuraman, A., Starr, B., Theesfeld, C.L., Andrada, R., Binkley, G., Dong, Q., Lane, C., Schroeder, M., Weng, S., Botstein, D., & Cherry, J.M. (2004). *Saccharomyces* genome database: underlying principles and organisation. *Brief Bioinform,* **5**(1), 9-22. PMCID: PMC3037832.
5. Balakrishnan, R., Christie, K.R., Costanzo, M.C., Dolinski, K., Dwight, S.S., Engel, S.R., Fisk, D.G., Hirschman, J.E., Hong, E.L., Nash, R., Oughtred, R., Skrzypek, M., Theesfeld, C.L., Binkley, G., Dong, Q., Lane, C., Sethuraman, A., Weng, S., Botstein, D., & Cherry, J.M. (2005). Fungal BLAST and Model Organism BLASTP Best Hits: new comparison resources at the Saccharomyces Genome Database (SGD). *Nucleic Acids Res,* **33**(Database issue), D374-377. PMCID: PMC539977.
6. Fisk, D.G., Ball, C.A., Dolinski, K., Engel, S.R., Hong, E.L., Issel-Tarver, L., Schwartz, K., Sethuraman, A., Botstein, D., & Cherry, J.M. (2006). *Saccharomyces cerevisiae* S288C genome annotation: a working hypothesis. *Yeast,* **23**(12), 857-865. PMCID: PMC3040122.
7. Hirschman, J.E., Balakrishnan, R., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hong, E.L., Livstone, M.S., Nash, R., Park, J., Oughtred, R., Skrzypek, M., Starr, B., Theesfeld, C.L., Williams, J., Andrada, R., Binkley, G., Dong, Q., Lane, C., Miyasato, S., Sethuraman, A., Schroeder, M., Thanawala, M.K., Weng, S., Dolinski, K., Botstein, D., & Cherry, J.M. (2006). Genome Snapshot: a new resource at the Saccharomyces Genome Database (SGD) presenting an overview of the Saccharomyces cerevisiae genome. *Nucleic Acids Res,* **34**(Database issue), D442-445. PMCID: PMC1347479.
8. Carbon, S., Ireland, A., Mungall, C.J., Shu, S., Marshall, B., Lewis, S., AmiGO Hub, & Web Presence Working Group. (2009). AmiGO: online access to ontology and annotation data. *Bioinformatics,* **25**(2), 288-289. PMCID: PMC2639003.
9. Gene Ontoogy Consortium. (2010). The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res,* **38**(Database issue), D331-335. PMCID: PMC2808930.
10. Engel, S.R., Balakrishnan, R., Binkley, G., Christie, K.R., Costanzo, M.C., Dwight, S.S., Fisk, D.G., Hirschman, J.E., Hitz, B.C., Hong, E.L., Krieger, C.J., Livstone, M.S., Miyasato, S.R., Nash, R., Oughtred, R., Park, J., Skrzypek, M.S., Weng, S., Wong, E.D., Dolinski, K., Botstein, D., & Cherry, J.M. (2010). *Saccharomyces* Genome Database provides mutant phenotype data. *Nucleic Acids Res,* **38**(Database issue), D433-436. PMCID: PMC2808950.

D. Research Support

None

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors in the order listed on Form Page 2.
Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME Hitz, Benjamin Cisco	POSITION TITLE Senior Software Developer
eRA COMMONS USER NAME (credential, e.g., agency login)	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
University of California, Los Angeles	B.S.	06/92	Biochemistry
Columbia University, New York	Ph. D.	09/99	Molecular Biophysics

Please refer to the application instructions in order to complete sections A, B, C, and D of the Biographical Sketch.

A. Personal Statement

I am an experienced software developer and manager with a broad understanding of modern genomics, bioinformatics and biochemistry. I have served as the leader and manager of the programming group of 4 at the *Saccharomyces* Genome Database, as well as interacted extensively with the curation staff at SGD and all components of the Gene Ontology Consortium. I have lead efforts to develop, maintain, and expand web applications and scientific databases of a heterogeneous and detailed nature. I am always on the lookout to adopt technologies that have progressed beyond "bleeding edge" into tools and techniques that improve the productivity of my software development groups. I am extremely flexible in technology choices and work equally well with systems written in Perl, Python, Java, Javascript, C or any other language, system, or framework that is needed or encountered. As a manager, I empower my reports to solve problems in their own way, providing guidance, suggestions, and code review along the process.

I maintain contacts in the open source development word and GMOD ("Generic Model Organism Database") project and attend a variety of scientific conferences to keep myself abreast of new developments in biology, genomics, informatics, and biotechnologies. I have met extensively with Jim Kent and Kate Rosenbloom and feel there is much common ground for a long and fruitful collaboration.

B. Positions and Honors. List in chronological order previous positions, concluding with your present position. List any honors. Include present membership on any Federal Government public advisory committee.

Positions and Employment

1999-2002	Scientific Applications Manager, ProCeryon Biosciences Inc., NY NY
2002-2004	Bioinformatics Research Scientist, Exelixis Inc, South San Francisco, CA
2005-Present	Lead Software Developer, Saccharomyces Genome Database, Stanford University, CA

C. Selected peer-reviewed publications (in chronological order). Do not include publications submitted or in preparation.

1. Bu DF, Erlander MG, Hitz BC, Tillakaratne NJ, Kaufman DL, Wagner-McPherson CB, Evans GA, Tobin AJ. Two human glutamate decarboxylases, 65-kDa GAD and 67-kDa GAD, are each encoded by a single gene. Proc Natl Acad Sci U S A. 1992 Mar 15;89(6):2115-9. PMCID: PMC48607

2. Yang AS, Hitz B, Honig B. Free energy determinants of secondary structure formation: III. beta-turns and their role In protein folding. J Mol Biol. 1996 Jun 21;259(4):873-82. PubMed PMID: 8683589

3. Nayal M, Hitz BC, Honig B. GRASS: a server for the graphical representation and analysis of structures. Protein Sci. 1999 Mar;8(3):676-9. PMCID: PMC2144291.

4. Nash, R., Weng, S., Hitz, B., Balakrishnan, R., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hirschman, J.E., Hong, E.L., Livstone, M.S., Oughtred, R., Park, J., Skrzypek, M., Theesfeld, C.L., Binkley, G., Dong, Q., Lane, C., Miyasato, S., Sethuraman, A., Schroeder, M., Dolinski, K., Botstein, D., & Cherry, J.M. (2007). Expanded protein information at SGD: new pages and proteome browser. Nucleic Acids Res., 35(Database issue), D468-71. PMCID: PMC1669759.

5. Hong, E.L., Balakrishnan, R., Dong, Q., Christie, K.R., Park, J., Binkley, G., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hirschman, J.E., Hitz, B.C., Krieger, C.J., Livstone, M.S., Miyasato, S.R., Nash, R.S., Oughtred, R., Skrzypek, M.S., Weng, S., Wong, E.D., Zhu, K.K., Dolinski, K., Botstein, D., & Cherry, J.M. (2008). Gene Ontology annotations at SGD: new data sources and annotation methods. Nucleic Acids Res. 36(Database issue), D577-81. PMCID: PMC2238894.

6. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, Ireland A, Lomax J, Carbon S, Mungall C, Hitz B, Balakrishnan R, Dolan M, Wood V, Hong E, Gaudet P. AmiGO: online access to ontology and annotation data. Bioinformatics. 2009 Jan 15;25(2):288-9. PMCID: PMC2639003

7. Costanzo, M.C., Skrzypek, M.S., Nash, R., Wong, E., Binkley, G., Engel, S.R., Hitz, B., Hong, E.L., Cherry, J.M., & the *Saccharomyces* Genome Database Project. (2009). New mutant phenotype data curation system in the *Saccharomyces* Genome Database. Database, bap001. PMCID: PMC2790299.

8. Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements. Nucleic Acids Res. 2010 Jan;38(Database issue):D331-5. PMCID: PMC2808930.

9. Engel, S.R., Balakrishnan, R., Binkley, G., Christie, K.R., Costanzo, M.C., Dwight, S.S., Fisk, D.G., Hirschman, J.E., Hitz, B.C., Hong, E.L., Krieger, C.J., Livstone, M.S., Miyasato, S.R., Nash, R., Oughtred, R., Park, J., Skrzypek, M.S., Weng, S., Wong, E.D., Dolinski, K., Botstein, D., & Cherry, J.M. (2010). *Saccharomyces* Genome Database provides mutant phenotype data. Nucleic Acids Res., 38(Database issue), D433-6. PMCID: PMC2808950.

C. Research Support. List selected ongoing or completed (during the last three years) research projects (federal and non-federal support

NONE

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors in the order listed on Form Page 2.
Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME Rosenbloom, Kate Rose	POSITION TITLE Bioinformatics Software Developer, Technical Project Manager
eRA COMMONS USER NAME (credential, e.g., agency login)	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Stanford University, California	BA with Distinction	1974	Human Biology
University of California, Berkeley		1976	Neurobiology (Graduate program)

A. Personal Statement

My eight years of experience with ENCODE, both production and pilot phases, in engineering and technical project management roles, places me in a unique position to understand the requirements and play a senior role in the design and development of engineering solutions for the next phase of ENCODE. Prior to joining UCSC, my 25+ year career as a professional developer of scientific, technical, and systems software in both industry and research settings has given me broad experience in methodologies and best practices for creating reliable and usable software that can last the test of time.

Throughout my career I have devoted special attention to the user experience of the software and scientific resources with which I have been associated. During the ENCODE project, I held primary responsibility for the public ENCODE portal website, as well as ENCODE DCC outreach. I will apply this experience in the next phase of ENCODE to maximizing utility of the ENCODE data to the broader biomedical community.

As an early member of the UCSC Genome Browser software developer group and responsible engineer for several organism genome browsers (Fugu and Chimp) and comparative genomics tracks and displays, I have a solid understanding of issues of genome data visualization, and a long-standing working relationship with the Genome Browser engineers who will integrate ENCODE data into the browser and related sites.

B. Positions and Honors

Positions and Employment

1974-1975	Research Assistant, Cellular Immunology, Stanford Hospital
1975-1976	Programming Intern, Xerox Palo Alto Research Center
1976-1981 and 1985-1987	Life Sciences Programmer; UNIX Systems Manager, Informatics Inc. at NASA/Ames Research Center
1981-1984	Software Engineer, ROLM Corp.
1984-1994	Software Consultant (Clients: IBM, Weitek, Seagate, Objectivity)
1995-2000	Senior Software Engineer, Sage Instruments
2000-2003	Senior Software Engineer, RapidMoney Corp.
2003-2007	Bioinformatics Software Developer, UC Santa Cruz
2008-present	ENCODE Data Coordination Center Technical Project Manager, UC Santa Cruz

Other Experience and Professional Memberships

2011 (pending) Member Human Proteome Project Scientific Advisory Board

C. Selected Peer-reviewed Publications

1. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, et al. **Genome sequence of the Brown Norway rat yields insights into mammalian evolution**. Nature. 2004;428(6982):493-521. Epub 2004/04/02. doi: 10.1038/nature02426. PubMed PMID: 15057822.
2. **The ENCODE (ENCyclopedia Of DNA Elements) Project**. Science. 2004;306(5696):636-40. PubMed PMID: 15499007.
3. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, **Rosenbloom K**, et al. **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes**. Genome research. 2005;15(8):1034-50. Epub 2005/07/19. doi: 10.1101/gr.3715005. PubMed PMID: 16024819; PubMed Central PMCID: PMC1182216.
4. Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, et al. **Forces shaping the fastest evolving regions in the human genome**. PLoS genetics. 2006;2(10):e168. Epub 2006/10/17. doi: 10.1371/journal.pgen.0020168. PubMed PMID: 17040131; PubMed Central PMCID: PMC1599772.
5. Pedersen JS, Bejerano G, Siepel A, **Rosenbloom K**, Lindblad-Toh K, Lander ES, et al. **Identification and classification of conserved RNA secondary structures in the human genome**. PLoS computational biology. 2006;2(4):e33. Epub 2006/04/22. doi: 10.1371/journal.pcbi.0020033. PubMed PMID: 16628248; PubMed Central PMCID: PMC1440920.
6. Thomas DJ, **Rosenbloom KR**, Clawson H, Hinrichs AS, Trumbower H, Raney BJ, et al. **The ENCODE Project at UC Santa Cruz**. Nucleic Acids Res. 2007;35(Database issue):D663-7. Epub 2006/12/15. doi: gkl1017 [pii] 10.1093/nar/gkl1017. PubMed PMID: 17166863; PubMed Central PMCID: PMC1781110.
7. Karolchik D, Bejerano G, Hinrichs AS, Kuhn RM, Miller W, **Rosenbloom KR**, et al. **Comparative genomic analysis using the UCSC genome browser**. Methods Mol Biol. 2007;395:17-34. Epub 2007/11/13. doi: 1-59745-514-8:17 [pii]. PubMed PMID: 17993665.
8. Margulies EH, Cooper GM, Asimenos G, Thomas DJ, Dewey CN, Siepel A, et al. **Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome**. Genome Res. 2007;17(6):760-74. Epub 2007/06/15. doi: 17/6/760 [pii] 10.1101/gr.6034307. PubMed PMID: 17567995; PubMed Central PMCID: PMC1891336.
9. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, et al. **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** Nature. 2007;447(7146):799-816. Epub 2007/06/16. doi: 10.1038/nature05874. PubMed PMID: 17571346; PubMed Central PMCID: PMC2212820.
10. Miller W, **Rosenbloom K**, Hardison RC, Hou M, Taylor J, Raney B, et al. **28-way vertebrate alignment and conservation track in the UCSC Genome Browser**. Genome Res. 2007;17(12):1797-808. Epub 2007/11/07. doi: gr.6761107 [pii] 10.1101/gr.6761107. PubMed PMID: 17984227; PubMed Central PMCID: PMC2099589.
11. **Rosenbloom K,** Taylor J, Schaeffer S, Kent J, Haussler D, Miller W. **Phylogenomic resources at the UCSC Genome Browser.** Methods Mol Biol. 2008;422:133-44. Epub 2008/07/17. doi: 10.1007/978-1-59745-581-7_9. PubMed PMID: 18629665.
12. Pollard KS, Hubisz MJ, **Rosenbloom KR**, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20(1):110-21. Epub 2009/10/28. doi: 10.1101/gr.097857.109. PubMed PMID: 19858363; PubMed Central PMCID: PMC2798823.
13. **Rosenbloom KR**, Dreszer TR, Long JC, Malladi VS, Sloan CA, Raney BJ, et al. **ENCODE whole-genome data in the UCSC Genome Browser: update 2012**. Nucleic Acids Res. 2011. Epub 2011/11/15. doi: 10.1093/nar/gkr1012. PubMed PMID: 22075998.

14. Churakov G, Sadasivuni MK, **Rosenbloom KR**, Huchon D, Brosius J, Schmitz J. **Rodent evolution: back to the root**. Mol Biol Evol. 2010;27(6):1315-26. Epub 2010/01/27. doi: 10.1093/molbev/msq019. PubMed PMID: 20100942.
15. Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, Kuhn RM, et al. **The UCSC Genome Browser database: extensions and updates 2011**. Nucleic Acids Res. 2011. Epub 2011/11/17. doi:10.1093/nar/gkr1055. PubMed PMID: 22086951.

D. Research Support

1 U41 HG004568-01 Kent (PI) 09/29/07- 06/30/11
National Human Genome Research Institute
Project Title: The UCSC ENCODE Data Coordination Center
Project Goals: Collect, organize, store, and provide access to data from the ENCODE project and other related projects through the UCSC Genome Browser.
Role: Technical Project Manager

RESOURCES

Follow the 398 application instructions in Part I, 4.7 Resources.

Please see the following pages for Resources/Facilities and Equipment for the Stanford University, plus three (3) other cooperation institutions (i.e. UCSC, EBI and Emory).

STANFORD RESOURCES

Follow the 398 application instructions in Part I, 2.7 Resources.

Dr. Cherry's group owns five Sun Microsystems T2000 systems that are used to provide production, curation, staging/testing, and development for the *Saccharomyces* Genome Database (SGD) resource. These Solaris servers have disk arrays providing 3.6TB of usable disk space. In addition we have seven Red Hat Enterprise Linux computers, all Dells, that are used by SGD for third-party software implemented as specialized components of the production web site. SGD also has a handful of older Sun Microsystems servers that provide various support roles. These servers collectively provide 26 TB of disk space for SGD. The Cherry group also provides the production environment for the Gene Ontology (GO) project with four Red Hat Enterprise Linux Dell systems with a total disk capacity of 3.6TB. Three older servers provide non-production support roles. Both SGD and GO production web servers are provided to the Internet via two Foundry ServerIronXL load balancer appliances. All production servers and their boot disks are configured with no single point of failure. Unix backups have moved to a disk-based system involving Simpana. Unix filesystems and databases are backed from a dedicated Dell DL2200 with 22 TB of disk space for the backups. This client server model is similar to our old tape setup except with disk instead of tape. In addition, database dumps and important files are copied to a central campus facility for archiving where they are backed up and can be also stored on a data cloud via a campus negotiated arrangement. The Cherry group owns 20 Macintosh iMac desktop systems, most with a second display. The group also has Windows and Apple laptops that are used by the staff. There are two HP Laser printers, one color and one gray scale, for use by the group. All of this equipment is used at no direct cost to the sponsor of this proposal.

The SGD and GOC staff offices are at a School of Medicine building located two blocks south of the Stanford campus (1501 S. California Ave). Transit time to the Medical Center is 10 minutes via bicycle, 25 minutes via free campus shuttle, and 10 minutes via a GEM electric vehicle maintained by the Cherry Lab. Dr. Cherry is allocated cubicle space at the Center for Genomics and Personalized Medicine. The total assigned cubicle space averages ~60 sq. ft. per person. The current staff for the SGD and GOC projects totals 16 Ph.D. level scientist and a staff of 3 with over 35 years experience between them that provides database and Unix systems administration. There are a sufficient number of cubicles currently available to provide homes for the additional staff associated with this proposal. Dr. Cherry is assigned space in two campus computer rooms within the Department of Genetics campus location (85 and 50 sq. ft.). The Medical Center computer rooms have a HVAC system separate from building HVAC. The Department of Genetics provides a 50KVA UPS battery system that conditions power for both computer rooms. The smaller room contains servers for a number of labs and is also the networking and telecommunications hub for the Department.

The Center for Genomics and Personalized Medicine is connected via high-speed optical fiber to the main routing center for Stanford University's Internet and Internet2 access points. Stanford University provides email and associated disk space for all staff. Stanford Networking maintains several co-localization facilities. The new $41.2 million Stanford Research Computing Facility encompassing approximately 22,000 sq. ft. will be located at the SLAC campus one mile west of campus. This new computing facility should be operational by mid-2013.

The Cherry group will be moving to a new School of Medicine building in the summer of 2013. The School is creating the Technology and Innovation Campus just south of our current location. This Porter Drive campus and will be the new home of the Center for Genomics and Personalized Medicine (Mike Snyder, Director), the Stanford Genome Technology Center (Ron Davis, Director), as well as laboratories from the Departments of Genetics, Radiology, Medicine, and Radiation Oncology. All labs and centers to be located at this new state of the art campus will be focused around biomedical technologies, genomics and molecular biological applications to medicine. The new facilities will provide Dr. Cherry plenty of room for growth. He will be assigned several offices for senior staff, and as many cubicles as needed. Each cubicle will provide ~80 sq. ft. per person. The new center will include a much larger computer room (~400 sq. ft.) with modern power and HVAC appropriate for a computing facility. Dr. Cherry's group will manage this machine room. Thus there will be ample office and computing space for expansion of Dr. Cherry's group.

UCSC RESOURCES

Follow the 398 application instructions in Part I, 2.7 Resources.

Dry Laboratory Facilities:

Four computational research labs (from 540 to 1450 ft2), divided into cubicles, and a row of private offices on the 5th floor east wing of the Engineering 2 building provide seating for 51 research and technical staff, postdocs, graduate students, telecommuters, & visiting researchers. Computers are either Dell desktop workstations or IBM, Apple, or Dell laptops; some staff have both desktop setups and laptops.

Computing Resources:

Below we have listed the most important major equipment items already available for this project, noting the location and pertinent capabilities of each. We then describe the overall computational infrastructure available to this project.

- 256 quad core Intel Xeon compute nodes (8 GB memory each, 1024 cores total); rm 213 Baskin Engineering
- 198 dual AMD Opteron processor compute nodes (4 GB memory each, 396 processors total); rm 594 Engineering 2
- 8 dual dual-core AMD processors (32 GB memory each); rm 213 Baskin Engineering
- 2 computers (1 TB memory and 64 processing cores each, 24 TB local disk space) for software and database development, rm 213 Baskin Engineering
- 2 computers (500 GB memory and 64 processing cores each, 20 TB local disk space) for high-availability virtual machine hosting, rm 213 Baskin Engineering
- 32 file servers (24 data servers, 8 metadata servers, 240 terabytes of networked data storage total); rm 213 Baskin Engineering
- 8 dual quad core Intel Xeon processor web servers (64 GB memory each, 64 processors total); UCSC ITS Data Center
- 15 BLAT servers (16-64 GB memory each); UCSC ITS Data Center
- Download server; UCSC ITS Data Center
- Redundant/Load-balancing download and networked file server; UCSD SDSC Colocation Center
- 36 computers (288 cores, connected with 10G network) with access to 500 TB of data; UCSD SDSC Colocation Center
- 49 desktop workstations and 27 laptops for staff, postdoctoral researchers, graduate students, telecommuters, and visitors; Engineering 2, 5th floor.

Computational clusters: To support UCSC's Genome Browsers and associated tools, databases, and genome research, the Center for Biomolecular Science and Engineering (CBSE) runs two parallel processor facilities, currently managed and supported by 3.25 FTE system administrators. The newest of these, the Swarm, consists of 256 quad core Intel Xeon compute nodes, each with 8 gigabytes of memory, totaling 1024 cores and running on a Linux operating system. The second, the PitaKluster, has 198 dual AMD Opteron processor compute nodes, each having 4 gigabytes of memory. Both systems were designed to provide an exceptional amount of inexpensive computing power in minimal space. For memory intensive jobs, CBSE employs a cluster of 8 machines with dual dual-core AMD processors and 32 gigabytes of memory each. These computational clusters are supported by a group of file servers (24 data servers and 8 metadata servers running IBM GPFS), providing almost 240 terabytes of usable, replicated network storage.

UCSC RESOURCES

Follow the 398 application instructions in Part I, 2.7 Resources.

The clusters are connected with at least gigabit ethernet to our network infrastructure.

Network infrastructure: We currently have a Cisco based network backbone, with the entire backbone being linked at 10Gb/s. In our main datacenter, we utilize a Cisco 6509-E and and Cisco 6506-E for private and firewalled area switching, as well as several 3750-E switches for edge switching to the UCSC core. We have about a dozen end-hosts connected at 10Gb/s as well to the 6506-E for high speed access to our firewalled space as well as our GPFS filesystem and cluster environment. The internal backbone in our main datacenter extends to two satellite locations in shared datacenters around campus, in which we have a small presence. These satellite locations are connected at 10Gb/s through 3750-E switches.

Development computers: CBSE employs 2 main computers with 64 computing cores and 1TB of memory each for software and database development. These development servers attach to 24 TB of local disk space. We also provide a high-availability connected computer setup for virtual machine hosting. These redundant machines have 64 computing cores, 500 GB of memory, and access to 20 TB of local disk space.

Web servers: The web servers for the UCSC Genome Browser consist of 8 dual quad core Intel Xeon processors that each offer 64 GB of internal solid state storage and 64 gigabytes of memory. These machines have access to a central file server that provides up to 10 extra TB of shared disk area, and a central Mysql database server that holds up to 13 TB of genomic data. Fifteen additional servers provide web access to BLAT (Blast-like alignment tool) software. Each of these machines have 16 to 64 gigabytes of memory, since BLAT is a memory-intensive application. Some of the other services we provide include a genome-preview server that allows the public to access our raw data before it has gone through QA, a public mysql server that hosts all our mysql data, a custom track server to store user-generated custom tracks, and a wiki server that holds public information and can keep track of named sessions. We are also working on providing systems in different geographical areas to reduce the access penalty, and the first European system will soon be deployed. We are currently migrating and taking advantage of cloud technology and have 1 cloud server (8 gigabytes of memory) in use for our BLAT software. Finally, a local download server allows users to download our data; it serves nearly 2 TB of data every day. We also house 1 identical download server and one additional filesever at the UCSD SDSC colocation center for load-balancing and redundancy. All of the machines serving the genome browser data are housed in a data center that is designed to function 24/7, 365 days a year.

We are quickly outgrowing our computer room facilities on campus, so we're planning on relocating part or all of our operations to a colocation facility in UCSD SDSC. This facility provides plenty of room for growth, as well as reliable 24/7 power and cooling to maximize the uptime of our servers. We currently have 4 racks holding 36 computers (288 cores, connected with 10G network) with access to 500 TB of data located in the SDSC datacenter.

Office:
Offices and cubicles for 8 administrative staff and several student assistants, also on the 5th floor east wing of the Engineering 2 building, provide administrative support to the UCSC Genome Browser staff and other programs and efforts managed by the CBSE

UCSC RESOURCES

Follow the 398 application instructions in Part I, 2.7 Resources.

Other:
Ten conference and meeting rooms (10–100 seats) are available in both the Baskin Engineering and Engineering 2 buildings; 6 of these have teleconferencing and videoconferencing facilities (webcam or videocam; three with plasma screens); 3 have built-in video/data projection, sound systems, Mac and PC computers with web and presentation software; 1 has a computer-controlled 4-screen projection system. Several small informal meeting areas are available in office suites and attached to the larger work areas throughout the 5th floor of the Engineering 2 building

EBI RESOURCES

Follow the 398 application instructions in Part I, 2.7 Resources.

For the proposed project, no laboratory, clinical or animal facilities will be needed.

As for the office space, the Ensembl team at the EBI is located on the Wellcome Trust Genome Campus in Hinxton, Cambricgeshire, UK. The EBI employs ca. 490 staff in two connected, purpose-built buildings containing both office space and machine rooms. One of the buildings was only completed in 2007, and the other has been renovated in 2009. The usual office infrastructure is available at a high standard. Adminsitrative tasks are shared between the on-site EBI adminstration and the EMBL headquarters in Heidelberg, Germany. Adequate office space is available for current and personnel to be hired.
The EBI shares a central library facility with the adjacent Sanger Institute.

Powerful computers for data storage and database development and high-speed internet connections are crucial for this project. The EBI is Europe's largest bioinformatics centre, providing access to its databases and application services to the international research community through a 10Gbps internet connection via London, supplemented by a 1Gbps backup via Cambridge. The EBI uses 4 machine rooms; 1 on campus at Hinxton, one nearby but off campus and two in London. The London datacentres have theoretical bandwidth of up to 100 Gbit to the main JANET/GEANT/INTERNET2 routers in London; the bandwidth is capped by agreement with the UK Academic network at 2 Gbit/second, but with provision for occassional spiking up to 10Gbit/second. The EBI has a high speed UDP based Aspera server protocol, allowing with very efficient use of the JANET/GEANT/INTERNET2 network, and we have observed sustained 3 Gbit/second transfers to both NCBI and UCSC using the Aspera network; this is faster than many National or even internal to some campuses transfer.

Database services are hosted by 10 dedicated Sun Solaris and 60 Linux servers running Orale 9i, Oracle 11g and MySQL. Serving of static and dynamic web pages is managed by an array of 100+ dedicated web servers (multi-core linux servers) in a redundant load-balanced fail-over configuration.
For compute-intensive tasks, such as sequence similarity searches, the web servers are supported by 8000 CPU core Linux farms. All computational tasks are managed in an institute-wide LSF queuing system. About 10+ Petabyte of disk storage is centrally managed and backed up,comprising SAN (Storage Area Network) and NAS (Network Attached Storage) systems.
The desktops are standard Dell/IBM/Compaq PC's, replaced every three years, or SunRays.

MAJOR EQUIPMENT: List the most important equipment items already available for this project, noting the location and pertinent capabilities of each.

The following major resources are shared EBI resources, available for this project:

Database Servers:	10 Sun Solaris Servers (32 CPU, 32GB RAM), 60 Redhat Linux Servers (4-32 CPU, 16-256GB RAM)
Servers:	Linux farms: 8000 CPU Cores, 4-8GB per core
Central Storage:	10+ Petabyte (4608 TB) (SAN + NAS)
Internal Network:	40 Gbps backbone, 1000/100 Mbps desktop connections
External Network:	10 Gbps main link to JANET at London, 1Gbps backup link (via Cambridge), Capped outflow/inflow at 2 Gbps with provision to spike to 10Gps to the other major networks

EMORY RESOURCES

Follow the 398 application instructions in Part I, 2.7 Resources.

The Taylor lab is located on the second floor of the O. Wayne Rollins Research building at Emory University. Dr. Taylor has his office there, along with an “open plan” software development lab and private office space for senior personnel. Facilities include a dedicated compute and development server (32 core), cluster (144 core, 12 node) and storage array (48 TB), as well as shared cluster resources located in server facilities at the Emory Math and Sciences building (connected by a high speed private network).

CHECKLIST

TYPE OF APPLICATION *(Check all that apply)*

☒ NEW application. *(This application is being submitted to the PHS for the first time.)*

☐ RESUBMISSION of application number: ______
(This application replaces a prior unfunded version of a new, renewal, or revision application.)

☐ RENEWAL of grant number: ______
(This application is to extend a funded grant beyond its current project period.)

☐ REVISION to grant number: ______
(This application is for additional funds to supplement a currently funded grant.)

☐ CHANGE of program director/principal investigator:
Name of former program director/principal investigator: ______

☐ CHANGE of Grantee Institution. Name of former Institution: ______

☐ FOREIGN application ☐ Domestic Grant with foreign involvement List Country(ies) Involved:

INVENTIONS AND PATENTS (*Renewal appl. Only*) ☐ No ☐ Yes

If "Yes," ☐ Previously reported ☐ Not previously reported

1. PROGRAM INCOME ***(See instructions.)***
All applications must indicate whether program income is anticipated during the period(s) for which grant support is requested. If program income is anticipated, use the format below to reflect the amount and source(s).

Budget Period	Anticipated Amount	Source(s)

2. ASSURANCES/CERTIFICATIONS ***(See instructions.)***
In signing the application Face Page, the authorized organizational representative agrees to comply with the policies, assurances and/or certifications listed in the application instructions when applicable. Descriptions of individual assurances/certifications are provided in Part III and listed in Part 1, 4.1 under item 14. If unable to certify compliance, where applicable, provide an explanation and place it after this page.

3. FACILITIES AND ADMINISTRATIVE COSTS (F&A)/ INDIRECT COSTS. See specific instructions.

☐ DHHS Agreement dated: ______ ☐ No Facilities And Administration Costs Requested.

☐ DHHS Agreement being negotiated with ______ Regional Office.

☒ No DHHS Agreement, but rate established with Office of Naval Research Date 08/05/11

CALCULATION* *(The entire grant application, including the Checklist, will be reproduced and provided to peer reviewers as confidential information)*

a. Initial budget period:	Amount of base $	1,429,059	x Rate applied	57%	% = F&A costs	$	814,564
b. 02 year	Amount of base $	1,394,682	x Rate applied	57%	% = F&A costs	$	794,969
c. 03 year	Amount of base $	1,436,518	x Rate applied	57%	% = F&A costs	$	818,815
d. 04 year	Amount of base $	1,479,613	x Rate applied	57%	% = F&A costs	$	843,379
e. 05 year	Amount of base $	-	x Rate applied		% = F&A costs	$	
					TOTAL F&A Costs	$	3,271,727

*Check appropriate box(es):

☐ Salary and wages base ☒ Modified total direct cost base ☐ Other base *(Explain)*

☐ Off-site, other special rate, or more than one rate involved *(Explain)*

Explanation *(Attach separate sheet, if necessary.)*:

4. DISCLOSURE PERMISSION STATEMENT: If this application does not result in an award, is the Government permitted to disclose the title of your proposed project, and the name, address, telephone number and e-mail address of the official signing for the applicant organization, to organizations that may be interested in contacting you for further information (e.g., possible collaborations, investment)? ☒ YES ☐ No

A. SPECIFIC AIMS

This project will serve the data production labs by specifying the accurate and complete submission of data, serve the greater ENCODE (Encyclopedia of DNA Elements) project by verifying and maintaining accuracy of the data, and serve the diverse user communities of skilled bioinformatics researchers, biologists, and educators by providing access to this collection of valuable genomic data. The collection includes the results of an unsurpassed set of genome-wide of chromatin and transcription factor assays that for the first time is allowing scientists to create a reliable, well annotated, and near complete set of transcriptional regulatory elements. It also includes results from a diverse set of innovative assays performed at the RNA and protein level. The DCC will define user-friendly interfaces and outreach media for connecting the scientific community to the experimental data and the results of their analysis. This community and project focused resource will guarantee ENCODE as the human genome portal for biomedical research by integrating the annotations of regulatory elements with the functional annotations of gene products.

Aim 1. The DCC will specify and process the data from the ENCODE data producers. The flow of data from the data producers to the ENCODE Portal will be specified and maintained. The submission pipeline includes controlled vocabulary registration, specification of metadata, validation and verification of the data.

Aim 2. The DCC will develop new and adopt existing technologies to facilitate access and integration of ENCODE data by biomedical research communities. New and existing software technologies for data access, integration and analysis will be integrated within the various pipelines and analysis tools leading to the ENCODE Portal. Huge data management and storage will be combined with metadata driven file registry.

Aim 3. The DCC will design, develop and maintain the ENCODE Portal. The multifunctional data portal will be designed, developed, and maintained for searching, filtering, and downloading data, metadata, and tracking project progress. The ENCODE Portal will become the central hub for access to the ENCODE data via a wide range of search options and providing connections to related resources for further in-depth investigation.

Aim 4. The DCC will provide enhanced access to data and computing environments via the cloud and Galaxy. All levels of data will be provided via the Internet from a state of the art computing facility. The data will be geographically distributed and available through conduits appropriate for differing requirements of bioinformatics experts through the public. Computing resources will be provided for expert and novice users via the Galaxy cloud computational back end.

Aim 5. The DCC will ensure accurate deposition of project results into the appropriate public archival repositories. A pipeline for submission of all necessary data and metadata will be prepared and modified as needed to meet the changing requirements of public sites such as NCBI databases, UCSC Genome Browser, EBI databases, and the appropriate community and model organism databases.

Aim 6. The DCC will incorporate and maintain previously produced data from ENCODE, modENCODE, and other genomic projects. This effort will require transferring data and converting existing metadata into a new specification, integrating the metadata within the newly developed AnnoDB database, and providing continuous access to these data. This task will include translation of *Homo*, *Mus*, *Drosophila* and *Caenorhabditis* metadata. Other large collections of data will be incorporated as defined by the NHGRI.

Aim 7. The DCC will be integrated with the DAC to form the unified ENCODE Data Coordination and Analysis Center. The DCC PI and co-Investigators will work with the Data Analysis Component's PI and co-Investigators to create and maintain a cohesive and dynamic organization for the optimal functioning of the ENCODE project. The DCC will provide the data management and user outreach to complement the integrative analysis of the DAC.

Aim 8. The DCC will maintain service to and interactions with the research community. The DCC will educate and assist the expert and novice in the use of ENCODE results and resources. A variety of services will be provided, including a Help Desk, community discussions, tutorials, presentations and updates via social media outlets, to assist the public in understanding of ENCODE data and the project in general.

B. RESEARCH STRATEGY

B.1 OVERVIEW

B.1.a Rationale

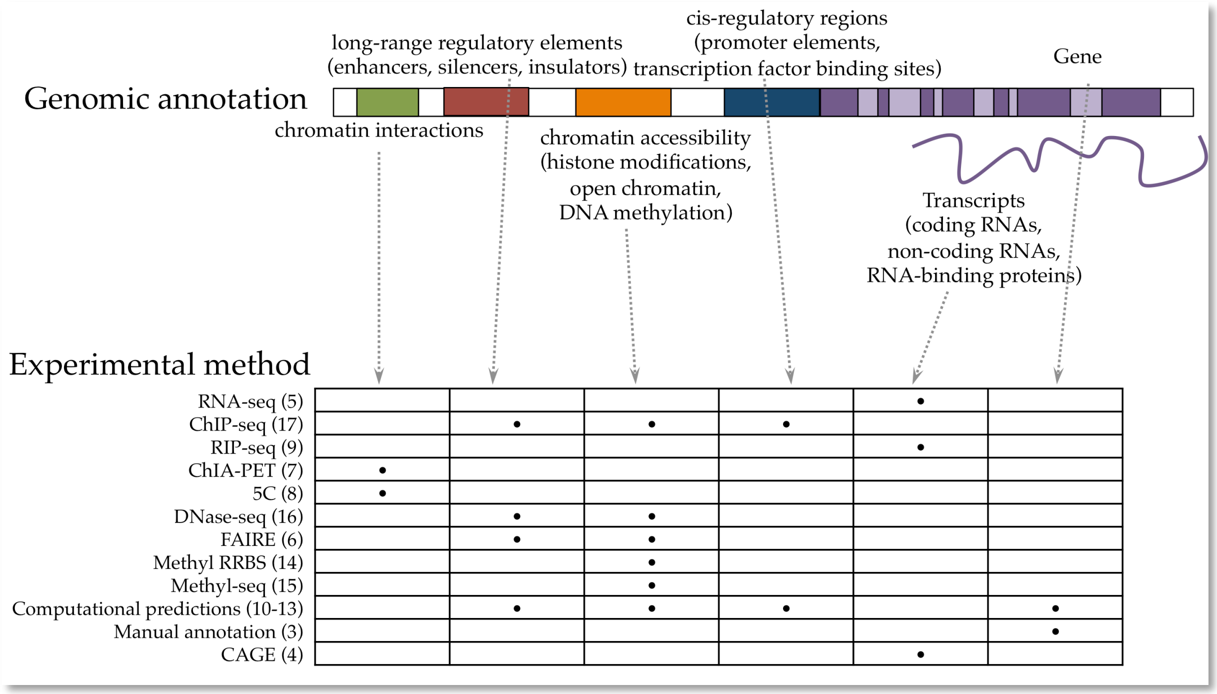
The ENCODE project was established shortly after the sequence of the nucleotides in the human genome was first established [1, 2]. The sequence information alone, three billion bases of A's, C's, G's, and T's, is a necessary starting point, but without additional annotations is of little use in understanding how a cell functions, or how genetic variants influence the phenotype of an organism, including genetic components of disease states. The overall strategy of the ENCODE project is to identify the functional elements of the genome, separating them from the large fraction of the genome that are figuratively speaking just along for the ride, including molecular parasites such as transposons and viruses, regions duplicated by errors in meiosis, and relics of genes once used by ancestral species, but which have decayed into pseudogenes in our species. The functional elements ENCODE seeks to identify include the exons of protein coding genes, transcribed regions such as piRNAs and lincRNAs that function at the RNA level, promoters that specify where genes start, enhancer, repressor and insulator regions that determine in what cell types and what environmental contexts a promoter will fire, splicing regulators that control how introns are spliced out of RNA, and signaling patches within RNAs that affect where in the cell they are localized, how frequently they are translated into protein, and how quickly they are degraded.

In addition to identifying functional elements, ENCODE seeks to annotate them, both directly and in collaboration with separately funded projects. Within ENCODE the large collection of cell lines, tissues, and experimental treatments allow us to specify where in the body and in what situation an element is used. A large collection of antibodies and other specific molecular probes allows us to define what molecules interact with an element. By analyzing the chromosomal elements in the context of human genome diversity projects such as 1000 Genomes, we are able to say what elements have variant alleles in a population. By intersecting ENCODE regions with the results of medically oriented genomic projects, we are able to identify alleles likely to have health consequences. One of the great successes of the current ENCODE project is the co-occurrence of notable regions defined by genome wide association studies with ENCODE-defined regulatory elements. The proposed extension of the ENCODE project will extend this success, and will be essential in interpreting the results of medical sequencing projects.

In order to achieve these ends, ENCODE employs a wide range of techniques. To determine the transcribed elements including mRNAs, the GENCODE subproject employs a team of biocurators who synthesize data from a wide range of assays to build accurate gene models that address the full complexity of vertebrate transcription including alternative promoters, alternative splicing, and RNA editing [3]. These assays include sequencing G-cap selected RNAs (CAGE) [4] for promoter characterization, next-generation RNA sequencing and assembly methods [5] for detecting alternative splicing and for quantizing the level of RNA in different cellular contexts. To determine transcriptional regulatory regions ENCODE uses DNase I, FAIRE [6], and nucleosome positioning assays to locate regions of the chromosome accessible to regulatory elements, DNA methylation assays and ChIP-seq of modified histones to define the overall chromatin architecture, and ChIP-seq of transcription factors to determine the players involved in the regulatory interactions. To determine which regulatory elements interact with each other, ENCODE uses ChIA-PET [7] and 5-C [8]. To identify regulatory elements that operate on the RNA rather than the DNA level, ENCODE employs RIP-seq [9] as well as a variety of computational assays [10-13] (Figure 1).

The variety of assays used by ENCODE necessitates involvement of a large number of labs, each of which is constantly striving to improve and develop new assays. Collecting the results of these diverse assays, along with necessary metadata needed to understand and reproduce the assays, and then organizing and presenting the resulting information in such a way that it is easy to browse, easy to search, easy to analyze, and ultimately possible to synthesize into not a mere catalog, but an *encyclopedia* of DNA elements is the core job of the Data Coordinating Center (DCC) and is the subject of this grant proposal. This job is an enormous challenge, one that will require all of the experience and expertise of the existing ENCODE DCC as well as significant new efforts as the ENCODE project expands in scope and aims for tighter integration with other major biomedical resource projects. To this end in this proposal we are bringing together the expertise of Jim Kent's group at UCSC who manage the current DCC with Michael Cherry's group who have extensive experience working with biomedical metadata, ontologies, and model organism gene and genome databases.

***Figure 1: Summary of the types of features captured by ENCODE**. Experimental methods used by ENCODE that are able to ascertain components of the chromosomal features shown. See references [3-17] for details about the experimental method.*



RNA-seq (5)					•	
ChIP-seq (17)		•	•	•		
RIP-seq (9)					•	
ChIA-PET (7)	•					
5C (8)	•					
DNase-seq (16)		•	•			
FAIRE (6)		•	•			
Methyl RRBS (14)			•			
Methyl-seq (15)			•			
Computational predictions (10-13)		•	•	•		•
Manual annotation (3)						•
CAGE (4)					•	

Integrating a diverse set of experimental data is the only way to start understanding the experiments in the context of a working human cell and to use the experiments to define a set of functional elements. The cell somehow integrates these elements in an elegant way to combine regulatory inputs in a combinatorial fashion to produce precise context-dependent gene expression outputs. To start to tease apart the complex inputs and outputs really requires smart and intuitive methods to house, slice and view the data internally and as an outside user. This is what (hopefully) will be described in the rest of the document.

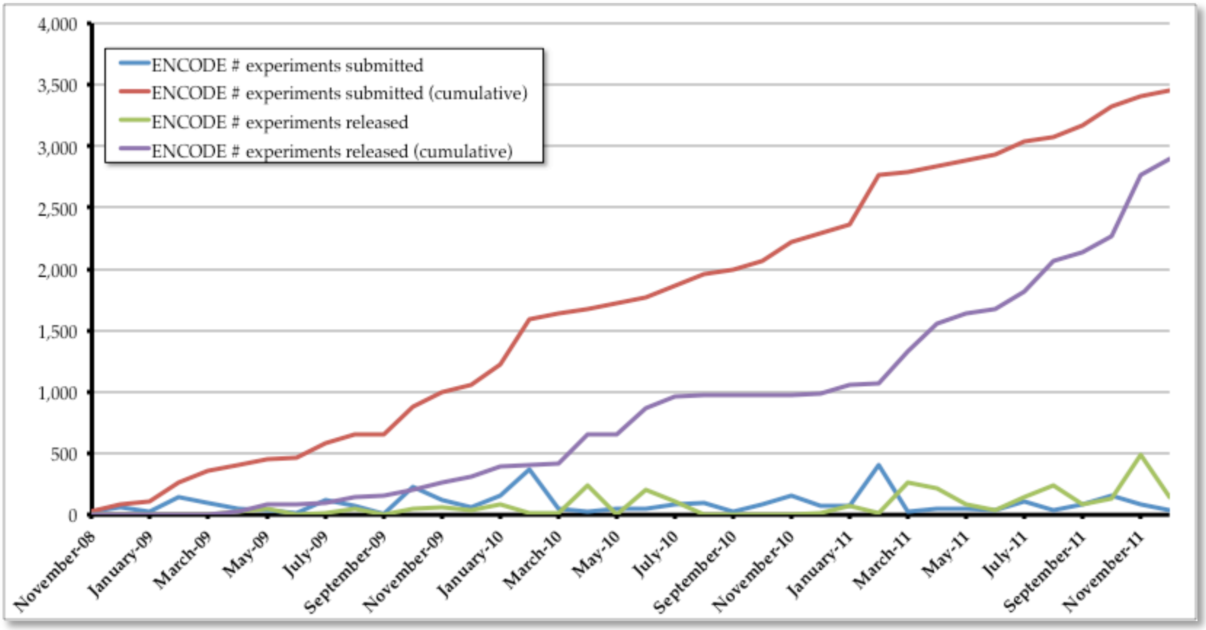
In 2003 the ENCODE (Encyclopedia of DNA Elements) project [18] began to investigate 1% of the human genome to determine the feasibility of exploring the complex, dynamic and transitive landscape of the human genome. In 2007 after determining the success of the pilot project [19] a full-scale effort began, identifying a range of functional elements in the human genome using a variety of high-throughput methods (Figure 1). Over the past four years 15.2 trillion base pairs of sequence data for 183 transcription factors from 248 human cell lines have been reported and submitted to the Data Coordination Center at UC Santa Cruz [20]. In 2009 The ENCODE project expanded to include similar data from the model organism *M. musculus*. The PI of the ENCODE DCC, Jim Kent, and his team at UC Santa Cruz developed a pipeline for the submission, validation, quality assessment, search and display of these complex data.

Concurrent with the ENCODE effort, the modENCODE project [21], focused on the model organisms *Drosophila* and *Caenorhabditis*, was created with similar goals in 2007. Since then, 316 experiments for *C. elegans* and 896 experiments for *D. melanogaster*, covering a similar range of experimental methods and biological interpretations as the ENCODE project, have been submitted to the modENCODE DCC maintained by Lincoln Stein (Ontario Institute for Cancer Research, Canada), along with Suzanna Lewis (Lawrence Berkeley National Laboratory, Berkeley, California), and Gos Micklem (University of Cambridge, UK).

B.1.b. Description of current ENCODE DCC

The ENCODE data submission, processing, and tracking pipeline has been in production use at the UCSC DCC since Summer 2008, providing a stable resource for accepting, processing, and validating datasets from the ENCODE data providers. The submission process begins with the establishment of a data agreement for each type of data that is generated by the production lab. The data agreement includes documentation of the experimental methods, description of the experimental conditions that will become the metadata for the submission, and expected file formats. Web-based interfaces developed by the DCC provide user-friendly pages for data providers to upload and submit ENCODE data and for production labs and DCC staff to monitor progression of these submissions through the pipeline (see Appendix 1.1 for a screenshot of the current ENCODE submission interface). This pipeline has been flexible enough to accommodate the range of experimental methods listed in Figure 1, for both human and mouse experiments. As of November 2011, the submission pipeline has processed and released almost 2700 experiments in the human and mouse genomes, approximately 230 of which are mouse (Figure 2, Appendix 1.2, and Appendix 1.3).

Figure 2: Summary of number of experiments in the human genome submitted to the ENCODE DCC and released.



For specific support of ENCODE, the DCC maintains an early access 'Preview' Browser (http://genome-preview.ucsc.edu). This Browser provides pre-release public access to all ENCODE data shortly after the data is submitted to the DCC while the data is undergoing quality assurance review.

For public access to ENCODE data, the DCC developed a dedicated ENCODE public website ("Portal"), at http://encodeproject.org. The portal features sections for Human and Mouse ENCODE projects. For each project, the portal hosts a summary of all experiments submitted, links to both the Genome Browser and the Preview browser for the appropriate genome builds (hg19 for human, mm9 for mouse), and tables of metadata from the ENCODE controlled vocabulary. The portal also features a session gallery, in which specific ENCODE tracks and genomic regions are highlighted with descriptive text, and a hyperlink that allows the user to launch a pre-configured Genome Browser session for further exploration of the highlighted data. The Portal provides link to pages describing details of metadata, tools for searching for datasets of interest, outreach and education resources, data standards and policy documents, and lists of publications.

ENCODE data users have access to all tools provided by the UCSC Genome Bioinformatics group – the Genome Browser, Table Browser, Gene Sorter, Blat, amongst others [22]. Specific tools developed for the ENCODE needs include sophisticated track configuration controls, track and file search using ENCODE metadata terms or free text, and sortable index pages for file downloads. Usage of ENCODE data by the scientific community has steadily grown. In the October and November of 2011, 62% of the 107,000 users viewing the human genome at UCSC Genome Browser had at least one ENCODE track in their display.

Software Environment: The Portal software platform relies on the Ruby on Rails (http://rubyonrails.org) web development framework, a MySQL (http://www.mysql.com) database, and a flexible back-end suite of file validation and database load utilities. The ENCODE DCC maintains all data tables in a responsive and reliable MySQL database on which the Genome Browser and other UCSC bioinformatics tools are based. FTP and HTTP file downloads are supported from a dedicated download server. A secondary download server is at the San Diego Supercomputer Center (SDSC), to alleviate congestion and provide redundancy.

The DCC manages projects using the Redmine (http://www.redmine.org) project management web application, and source code versioning is managed using the Git (http://git-scm.com) distributed software version control system.

Outreach: In addition to the Portal and the tools developed by the UCSC Genome Bioinformatics group, the ENCODE DCC supports the broader biomedical community by exporting data to other established repositories and by outreach and educational efforts. The DCC regularly submits ENCODE data to NCBI GEO after quality review is complete. The list of ENCODE data with accession numbers can be viewed at GEO [23]. Both Human and Mouse ENCODE now have GEO BioProject ID's (http://www.ncbi.nlm.nih.gov/bioproject).

The DCC announces new data releases and other ENCODE developments to a dedicated "encode-announce" (encode-announce@soe.ucsc.edu) mailing list, and maintains a news section for this information on the Portal. User questions are handled by another mailing list for ENCODE-specific questions, while general questions about use of resources and tools at UCSC are handled by the Genome Browser user help mailing list.

In addition, the DCC presents ENCODE resource gives presentations or posters at a minimum of two scientific meetings per year (e.g. CSHL Biology of Genomes and American Society of Human Genetics). The DCC summarizes the ENCODE data available each year in a manuscript published in the *Nucleic Acids Research* (NAR) Database issue (January). The three most recent NAR articles [20, 24, 25] have been cited 119 times (according to Google Scholar).

Coordination and management: In addition to maintaining the data submission pipeline, storing the data and metadata, and providing access to the data, the DCC plays a significant role in organizing and participating in meetings and conference calls. The DCC has taken many steps to foster smooth and effective communications with data production labs, working and analysis groups, the DAC, and Consortium members at large. In addition to teleconferences for data selection after data freezes, the DCC has established periodic conference calls with some of the larger labs. DCC representatives join weekly Analysis Working Group (AWG) and Transcriptome Working Group conference calls, and other Working Group calls when requested. The DCC management works closely with DAC management in a range of activities -- defining, managing, and enforcing analysis data freezes, specifying requirements for data formats and metadata, and importing selected analysis datasets into the DCC. Management calls include monthly NHGRI/DCC and monthly PI group teleconferences. The DCC reviews its status on all monthly ENCODE and Mouse ENCODE consortia call. The DCC has participated with the modENCODE DCC on occasional site visits, meetings, and calls to provide information on our workings to the ENCODE External Consultants Panel. The ENCODE private wiki (http://encodewiki.ucsc.edu) established by the DCC is actively used within the Consortium and Data Analysis Center for direct project communication and coordination.

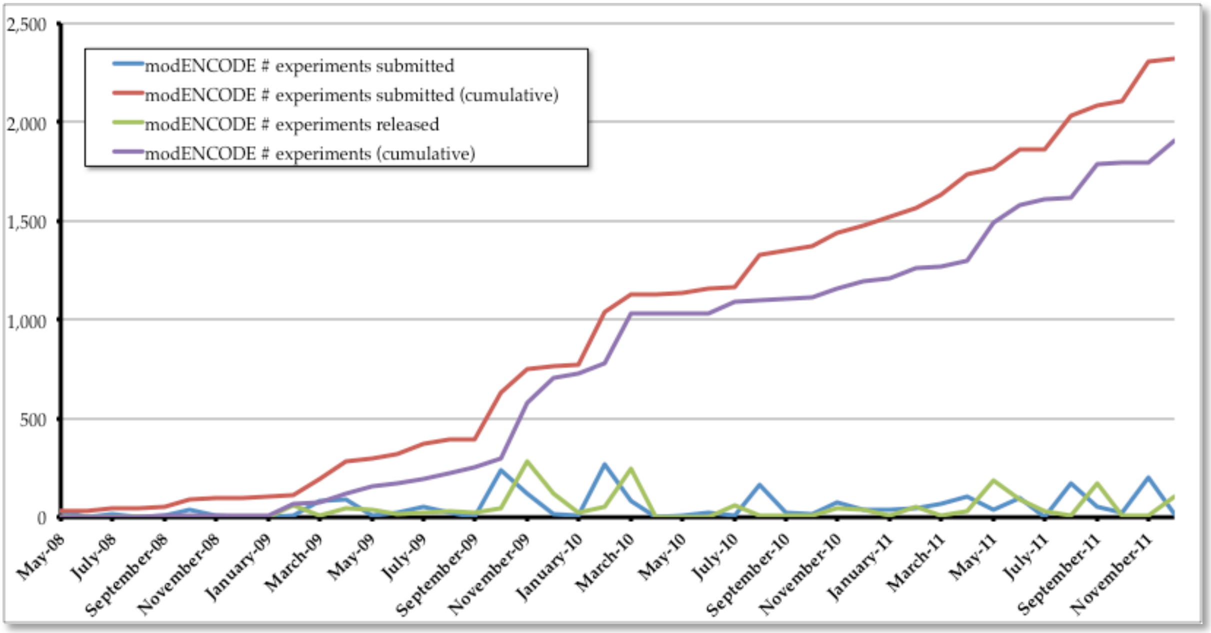
The DCC sends a large contingent to the annual Consortium meeting, sends at least one representative to each AWG workshop, and is represented at all PI group meetings. At consortium and analysis meetings, DCC representatives often host tutorial sessions and hands-on consultations for data providers. DCC Wranglers have initiated lab visits when concentrated efforts are needed to facilitate data submission changes. The DCC has also hosted a meeting of ENCODE working groups at UCSC.

Project staffing for the current ENCODE DCC has evolved over the four years since its inception, with management demands greatly exceeding expectation from the original grant. For the duration of the ENCODE DCC at UCSC, Jim Kent has provided overall direction as PI, with Kate Rosenbloom acting as Scientific/Technical Project Manager.

B.1.C. DESCRIPTION OF CURRENT MODENCODE DCC

The modENCODE data and metadata submission, storage, quality control, and distribution pipeline shares common elements with the ENCODE pipeline. Namely, the DCC works with the production labs to define the scope of the experiments that will be submitted by the lab and the experimental conditions that will be included as metadata, performs quality checks to validate the data, distributes the data to multiple databases and resources (such as GEO NCBI [26], FlyBase [27], and WormBase [28] to encourage its use by the scientific community. Visualization of the experimental data is available through the GBrowse genome browser for each organism [29-31] while querying and downloading of the datasets is available through modMine [32, 33] As of November 2011, approximately 1800 datasets have been processed and released by modENCODE (Figure 3).

Figure 3: Summary of experiments submitted and released by the modENCODE DCC*. The number of datasets represents both fly and worm data.*



B.2 RESEARCH RESOURCE PRODUCTION

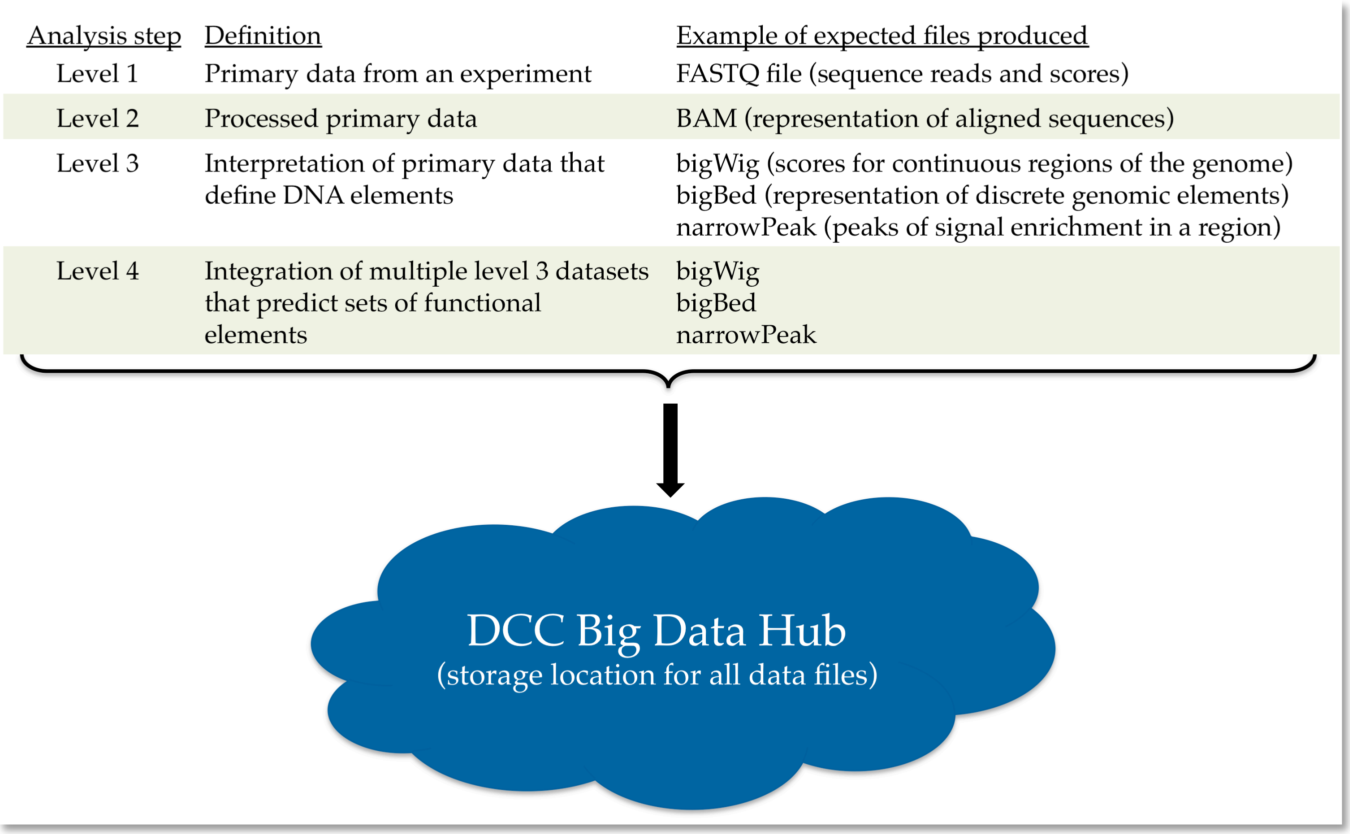
In the next phase of the ENCODE project the data production labs will become larger and will focus on established technologies for the determination of the characteristics of chromosomes during differentiation and development of human cell types. These production sites will produce more data, faster and of higher quality. These results will be processed by analysis centers, with methods that will be accessible to all: computational researchers, laboratory scientists, clinicians, educators, and the inquisitive public. All forms of these data and their analyses, from the primary experimental measurements (Level 1) to predicted sets of functional elements (Level 4), must be stored in such a way that allow regions of the genome to be identified and all the data used in creating these analyses to be available (Figure 4). These analyses will be coordinated by the Data Analysis component of a larger collaboration between the Analysis Working Group and the Data Coordinating component of the ENCODE Data Coordination & Analysis Center (EDCAC). With this phase of ENCODE the virtual EDCAC will act in concert to promote the data acquisition, verification, distribution, application, integration and analysis of ENCODE data. The availability of the analyses and the data underlying the results is the prime motivation for the creation of the EDCAC.

In this proposal we will extend the model defined by Kent, Stein, and colleagues for the next phase of the ENCODE adventure. The new DCC will create software and pipelines that will empower the production labs to validate their data prior to submission to the DCC. Certification and verification of these data by the DCC will be automated through the increased and complete use of controlled vocabularies and ontologies which will capture structured information about the procedures, cell types, and reagents used to create the reported results. The data files will be stored in the DCC Big Data Hub while the experimental data's annotations (metadata) will be stored in a relational database. These data will be made accessible via a data cloud that enhances the ability of scientist to utilize these data using their own, and often, novel computation methods or via web-based interfaces that allow searching and visualization of the results. The Data Coordinating component is the product of UCSC and Stanford joining forces to create a unique team with experience in all aspects required to provide for the four critical needs of the DCC (Figure 5). The Stanford and UCSC modules of this project has many years experience maintaining resource centers that include herding data submissions as well as developing and maintaining production software, powerful visualization tools, and creation of web sites that empower users to explore and learn from the collected information.

B.2.a **Aim 1: The DCC will specify and process the data from the ENCODE data producers**

The flow of data beginning with the data generation at the production labs and ending with its release on the ENCODE Portal will be specified and maintained by the DCC. This process starts with validation of the experimental data, specification of metadata describing the experimental details, which includes controlled vocabulary registration, and submission of both metadata descriptions and experimental data by the production labs. After submission, the metadata and data files will be verified via independent pipelines by the DCC. Software will be created for the production groups (client-side software) to facilitate the certification and validation of data and for the DCC (server-side software) to facilitate the verification of data. The quality of the data will be assessed at each stage.

Figure 4: Summary of data levels expected from the ENCODE production labs and analysis groups. *Examples of expected files, along with their definitions and file formats. All data files will be stored in the new DCC Big Data Hub. A file system housed at SDSC using a high efficiency file system.*



Analysis step	Definition	Example of expected files produced
Level 1	Primary data from an experiment	FASTQ file (sequence reads and scores)
Level 2	Processed primary data	BAM (representation of aligned sequences)
Level 3	Interpretation of primary data that define DNA elements	bigWig (scores for continuous regions of the genome) bigBed (representation of discrete genomic elements) narrowPeak (peaks of signal enrichment in a region)
Level 4	Integration of multiple level 3 datasets that predict sets of functional elements	bigWig bigBed narrowPeak

In addition to providing core functions that collect, store, and distribute the data submitted by the production labs, the DCC will integrate and maintain data previously produced by the ENCODE, modENCODE, and other genomic projects. These data will be available from the ENCODE portal with documentation and software that will provide a complete environment for biological discovery.

1. Certification and Validation of experimental data created by the production labs

As part of the certification process, production labs will submit protocols the labs plan to use as well as providing a description of the reagents that will be used for the experiments. The protocols will include the experimental method, such as ChIP-seq or DNase-seq, and any algorithms or software packages, such as QuEST [34] or bowtie [35, 36], and the software options and setting used for generating or analyzing the data. The description of reagents will include specifics about the experiment performed, the cell type studied, identity of antibodies used, the sequencing platform used, or any drugs that were added to the growth media. These details of the experimental method and descriptions of the reagents, collectively known as metadata, will be submitted using controlled vocabularies and ontology terms. Controlled vocabularies, standardized nomenclature, and ontologies defined and selected by the DCC will be used to populate these metadata fields (Table 2). These data will be translated for export into acceptable file formats for NCBI GEO [26] or an appropriate Model Organism Database, etc. The NCBO collection of public ontologies can be found at http://www.bioontology.org.

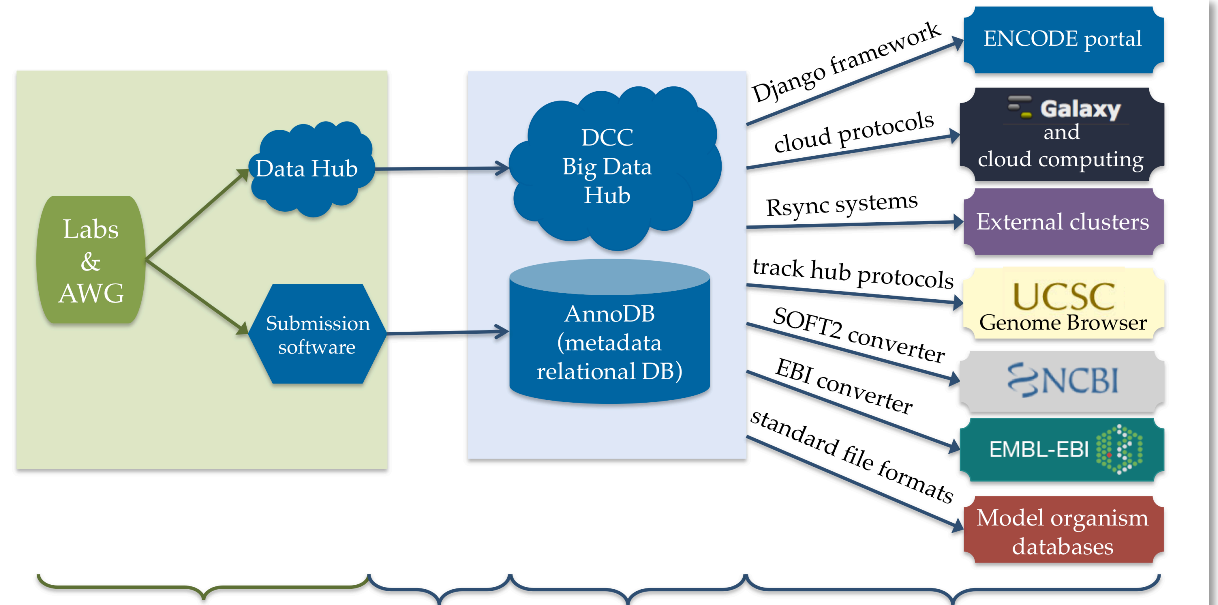
The production labs will determine if the necessary information can be submitted from these pre-determined controlled vocabularies. If not, the production labs will request additions to the metadata fields. The DCC Data Wranglers will handle all these requests, creating a provisional vocabulary term when new terms are required in order to minimize delays in processing. Data Wranglers are skilled individuals trained in data handling, metadata specification, and all aspects of data submission and manipulation, and assistance and education of the ENCODE user community. Once a request for a new term is received, the Data Wranglers will initiate additions and modifications to ENCODE-specific controlled vocabularies and ontology projects that are actively developed, such as the Gene Ontology (GO, [37], Sequence Ontology (SO, [38]), Ontology of Biomedical Investigations (OBI, [39]), Evidence Code Ontology (ECO,[40]), ChEBI ([41]), Cell Ontology (CL, [42]), Brenda Tissue Ontology (BTO, [43]) and MGED Ontology (MO, [44]). The Data Wranglers will reconcile provisional terms from all providers during the verification process, prior to initial release of experimental data. The required metadata fields must be defined in order for a new record to be certified.

Table 1: Glossary of Terms.

AnnoDB	Relational database used at the DCC to store metadata associated with data files	**GPFS**	**General Parallel File System**
AWG	**ENCODE Analysis Working Group**	**IHEC**	**International Human Epigenome Consortium**
Big Data Hub	Storage location for all files maintained by the DCC	**MOD**	Model Organism Database
DAC	ENCODE Data Analysis Center	**modENCODE**	Model Organism Encyclopedia of DNA Elements
Data Wrangler	Personnel which manages the interaction with lab, manipulates, transfers, & shepherds data to the Portal	**NCBO**	National Center for Biomedical Ontology
DC4	DCC Cluster of Computers	**OMIM**	Online Mendelian Inheritance in Man
DCC	ENCODE Data Coordination Center	**OSDC**	Open Science Data Cloud
EC2	Elastic Compute Cloud	**Portal**	ENCODE project public website
EDCAC	ENCODE Data Coordination and Analysis Center = DCC + DAC	**S3**	Simple Storage Service
Galaxy	Web based bioinformatics analysis and pipelining environment for the cloud.	**SDSC**	San Diego Supercomputer Center
GEO	Gene Expression Omnibus, NCBI	**SRA**	Sequence Read Archive, NCBI
GO	Gene Ontology	**UCSC**	University of California Santa Cruz

The current ENCODE and modENCODE DCC projects have used slightly different approaches to obtaining the metadata from the providers (see Appendix 1.4 and 1.5 for list of current metadata fields captured by ENCODE and modENCODE respectively). While the input format, submission tools and metadata storage was different between the two groups, the end result was quite similar, allowing access to the rich datasets using structured information. We propose to develop a client-side application that production labs will use to create dataset records that will be submitted to the DCC. The goal is to acquire precise information describing the details of an experiment via an intuitive interface that does not require unnecessary repetitive mouse clicks or dictionary searches. The submission forms will contain fields defining the required information. These fields will be pre-populated with controlled vocabulary terms but allow for the flexibility of requesting new terms. Production labs that have previously submitted datasets to the DCC, in the previous or current ENCODE project, will have new records pre-populated from their past submissions to expedite subsequent submissions. The use of AJAX style autocomplete will facilitate quick determination of any new vocabulary terms that have not already been used in the data provider's existing metadata. This DCC supplied software will conduct checks of the metadata; examples of these checks and when the checks will be performed are shown in Table 3 and a full list of current and proposed checks are listed in Appendix 2.1.

***Figure 5: Overall data flow and major processes**. The production labs will use programs provided by the DCC to populate web accessible data hubs and submit metadata from their own computer systems. These hubs can be visualized directly on the UCSC Genome Browser, allowing the lab a chance to do visual quality checks, and allowing external scientists to access the data early. The metadata will be written to a file that can be accessed by the DCC. When the labs are ready to submit, the DCC will pull data and the metadata from the labs, applying numerous automatic checks to the experimental data and the metadata. The DCC will use a high-throughput batch process to fetch the data and metadata. The data will go into a cloud file server, while the metadata will go into a relational database that references the data by unique URLs in the cloud. From the DCC the data will be distributed to external databases and directly to users via a number of protocols. See the corresponding aims for additional information.*



Group:	Data producers	DCC	DCC	DCC & External databases
Purpose:	Data certification	Data verification	Data storage	Data distribution
Action:	Certify metadata Validate data Create data hubs Submit data	Pull data files Pull metadata	File storage of data Store metadata	Reformat data Map metadata Provide access to data
Details:	Aim 1	Aim 1	Aim 2	Aims 3,4,5

Once the metadata data record is certified, the production labs can begin the process of validating their experimental data. It is essential that the metadata record be certified because the validation tools will test components of the experimental data as specified by this record. Beginning the certification process in advance of data generation also ensures the correct checks are present in the validation tools. The experiment type determines what files need to be generated, but many experiment types share the same file formats. For example, ChIP-seq and DNase-seq both use the same main data file formats, FASTQ for the short reads, BAM for the short read alignments, bigWig for plots of signal levels across the genome, and the narrowPeak variant of bigBed for defining regions of above-background signal intensity. However in other respects ChIP-seq protocols are significantly different from DNase-seq protocols, not the least because DNase-seq does not require an antibody. The validation software will include modules for each data type and perform the appropriate checks (Table 3).

As part of the data validation process, producers will establish a data hub. Data hubs use an environment that was created by the UCSC Genome Browser team that will be extended to capture ENCODE data in detail. A data hub is essentially a directory system that is provided via an HTTP(s) server. Documentation for the structure and setup of a data hub is available online [23]. The data hubs can be private in that knowledge of a specific URL is required, and an access password may be required as well. The access is defined by the data producers and thus under their control. Strict requirements for directory structure and management of the data will be defined by the DCC. The DCC will provide expert assistance to help the data providers build and maintain their server. The data hubs will allow the metadata and data sets to be downloaded by the DCC when notification of validation is complete.

***Table 2: Ontologies and unique identifiers to be used for metadata submission**. The basic experimental details for a ChIP-seq experiment using antibodies against GATA-1 in K562 cells are used as an abbreviated sample for the types of data that would be captured using existing ontologies or unique database identifiers.*

Metadata field	Relevant ontology or unique identifier	Example entry	Currently used at ENCODE?	Currently used at modENCODE?
Cell type	BTO or CL	K562	Y	Y
Cellular location of input sample	GO	nucleus	Y	Y
Input sample	MGED or OBI or SO or GO	chromatin	N	Y
Antibody target	UniProt ID or gene ID of target protein	GATA-1	N	Y
Assay performed	MGED or OBI or ECO	ChIP-Seq	N	Y/N
Chemical treatments	ChEBI	Sodium butyrate	N	N

In addition to the sample checks and features described in Table 3, the system will retain flexibility to add any new checks than can be defined and implemented. Use of current software by data providers will be mandated by an internal version check and automated download. We anticipate the installation of the local validation software to be straightforward, as we will simply require a Linux computer connected to the Internet. A pre-build software installer will be provided with a defined set of modules and libraries to simplify software installation. If more sophisticated software is required we will build virtual Linux machine images that can be downloaded and used with a minimum of system administration effort. This later appropriate has been used successfully used by Ensembl to distribute analysis tools with remote database access preconfigured. Remote connection software similar to what is used by computer support services will also be used to aid in the debugging process. The new software created for the submission pipeline will be quite different from the current ENCODE or modENCODE system. The DCC staff will facilitate automatic checks and certification of metadata by the data producer. Initially the DCC staff will visit each producer to help in the setup of all required software. This will include the HTTP server for the data hub, automatic log processing and error monitoring so the producer and DCC will be alerted if components at the producer's environment are not functional. This will allow the DCC to be proactive in it support of this critical initial exchange.

Education of the production lab will include complete documentation and online tutorials. Most importantly the DCC staff will use person-to-person videoconference such as provided by WebEx, Skype, and FaceTime and live Internet chat to inform the data provider of the necessary steps in the process. Each production lab will be encouraged to identify lab members who will serve as the primary contact for the DCC. The incentive to devote this level of commitment to data submission is to recognize the need by the labs and the public for powerful metadata search features. This will also increase the quality of the data before it reaches the DCC.

2. Verification of data submitted by the production labs

Once the production labs have submitted all data files and metadata documentation, the Data Wranglers will shepherd the submitted data through the second phase of the submission pipeline, data verification. This process will include final checks of the data files for format and content as well as finalization of all required metadata (Table 3). The result of this step is a well-documented and confirmed set of data that is released to the ENCODE Portal as described in Specific Aim 3.

When the Data Wrangler confirms that the metadata and data files are ready for incorporation into the DCC, the data files will be automatically pulled from the production lab's data hub. The software used by the production labs will write a report in lab's data hub that will also be emailed to an automated mailbox to begin the retrieval of the dataset. This report will provide a complete description of key attributes of the data, any software errors or manual checks that should be addressed by the data provider staff, and hard errors that require the providers attention. The validator can be run iteratively, each time generating a new report, until all errors are resolved. Once pulled from the data producer, and passing quick automatic basic DCC-side verifications, the data file is defined as submitted and the two-week clock for data release starts.

When the producer's data is retrieved, each data file, experiment and study will obtain a unique identifier [ten character hexadecimal string allowing 16**10 unique IDs]. The Big Data Hub will use the unique identifier string as the filename and the full metadata associated with this ID will be in the DCC database. Using the associated metadata, datasets can be virtually renamed if necessary for particular user interfaces. Unlike in previous DCCs the filename will not represent the metadata of the assay. In addition, each data file will be associated with a specific permanent URL and programmatically well defined for access via automated web services.

Table 3: Examples of check to be built into the software and the step at which they are performed. *The same checks are run at multiple steps to ensure data integrity. Specific checks are detailed in Appendix 2.1*

Example of automated check	Performed at certification?	Performed at validation?	Performed at verification?
All terms that are supposed to be in a controlled vocabulary really are in the controlled vocabulary	●	●	●
Submitted metadata values are consistent with each other	●	●	●
Main data files that the metadata says are supposed to exist actually do exist		●	●
Check that file formats are correct		●	●
Check that chromosomes names are part of the reference chromosome set for the assembly and organism		●	●
Check that chromosome coordinates do not exceed the length of the chromosome		●	●
Check that probability values are between 0 and 1		●	●
Check that the signal data is not dominated by huge peaks at regions prone to mapping artifacts such as genomic mitochondrial pseudogenes or alpha-satellite repeats		●	●
Check that the two replicates correlate significantly above chance.		●	●
Check that replicates across labs also agree			●
If there is an updated dataset, do datasets agree with previously submitted dataset? What has changed?			●

Experiences from the previous ENCODE and modENCODE DCC projects has shown that data wrangling is a relatively complex process, requiring an average of one FTE day to completely process one data file into one data track. This step is typically not completed within a calendar day rather it is spread over up to two weeks as interaction with the data provider is often required several times. Even though the data producers will be expected to provide data certification and validation during the submission process, we cannot reasonably propose to reduce this number below 0.5 FTE days per track due to the human nature of communicating with groups that support ontology development, with production labs to resolve errors, with the QA team to fix errors, and with archival and data distribution sites to properly export the data. Although there is every effort to validate data and ensure quality, errors do make their way into the data repository. One of the essential roles of the Data Wrangler is to track those errors once they are found. If an error does make it all the way to the public, for instance the lab mislabeled a cell line, the Data Wrangler follows the correction all the way through the process. This means getting replacement data, marking existing data as revoked, making sure the error is communicated to the data users and the correction is propagated to all of the other repositories such as GEO. In addition the wrangler must work with engineers to create possible improvements in the validation process. This will be a major area of active development for the DCC, creating software to make the process more efficient for the data producer and Data Wrangler.

3. Versioning and data provenance

A critical issue that has arisen in the past is the updating, tracking, and versioning of a given experimental report. The experiences of the previous ENCODE and modENCODE DCC projects have led us to build in a fully functional versioning system for both the data files and associated metadata (which may have to be edited or replaced wholesale due to data entry or other errors). Updates to data files or metadata will trigger a new hexadecimal string to be assigned to the data and metadata. Access to previous versions will be made available via the ENCODE Portal as well as web service queries (each version is accessible with a unique URL).

4. Importance of Data Wranglers to ENCODE

The overall data product that is being submitted by the producing laboratories is complex, dynamic and large. The data product complexity includes raw and processed data in several formats, lab methods, analysis parameters, controlled vocabularies and ontology terms, biological replicates and technical replicate information, versions, corrections, experimental validation, and qualitative descriptions. The data and metadata is being manipulated for display, for raw analysis, for searching and for inserting into several differing repositories. The information is dynamic in that the requirements for the formats are developing, new methods of verification are being discovered, and enhancements of lab techniques are being incorporated. Finally, the volume of data is large, in that as sequencing methods become cheaper, more and more raw data is being generated. As analysis methods are being developed, more versions of the same data seem relevant for comparisons. To move these data products along the process that is itself adapting to the changing needs of the data, the role of the Data Wrangler is essential. Data Wranglers are the integral human element in problem solving how to get the biologically and technically complex data into a more defined structure. In addition, the Data Wranglers are the primary collectors of the qualitative data about each track.

The Wranglers serve several masters and each of these masters requires unique assistance. These tasks are associated with their mission and their quest for ENCODE data. The masters of the Wrangler's services include production labs members, project software engineers, and external users.

4a. Role in QA, technical training, problem solving. The DCC Data Wrangler 's primary focus is to assist the producing labs in submitting a quality data product that complies with the standard of the group and yet reflects any experimental variation that may be relevant for downstream analysis. Many people (lab technicians, graduate students, bioinformaticians, and finally the data submitter) handle the data in the production labs. Each of those individuals makes assumptions about the pieces of data that they are receiving. A change can happen in the lab or bioinformatics processes which is not communicated to the submitter. It is not until the data is collected together and reviewed by the Data Wrangler that key questions come to light for example: The size of your raw data files have significantly changed? Did you change your sequencer or methods mid experiment? In this way, the Data Wrangler serves as a first round of QA for the data product and can assist the lab in developing their quality assurance methods. Additionally labs are expected to comply with formats, standards versions, validation requirements and controlled vocabulary usage. Achieving compliance requires communication, training and assistance on the part of the Data Wrangler. Often the labs need advice about which controlled vocabularies to use.

The DCC Data Wranglers work with the producing labs to uncover the idiosyncratic problems associated with each dataset. A common problem that arises is the discrepancy in the data submitted to the expected data format laid out by the lab in the initial data plan. It is the role of the Data Wranglers to work with the submitters to understand the reasons behind the discrepancy and generate a plan to convey the information effectively to the public and various repositories.

4b. Role in requirements gathering. As the Data Wranglers are intimately involved with the producing labs and the experiments, they are in the position of communicating back to the DCC's software engineers and of the other repositories the future needs or requests of the submitters as well as any errors or complications with existing software. Additionally as submitters to the other repositories, they will need to gather and communicate changing requirements on data standards from the repositories back to the DCC software engineers and the submitting laboratories.

4c. Role in cross lab consistency and data organization. The DCC serves as the central hub for the ENCODE project. It is the Data Wrangler's task to assist in maintaining cross lab consistency and quality standards. The Data Wrangler is uniquely positioned to be able to see the subtle changes and divergence between lab methods. They are also a common point of dissemination of process and standards from the consortia to the actual data producers/submitters. They can see the various issues that come up for the labs regarding the requirements and can collate those issues for the standards working groups. Their expertise and larger picture view is needed in the development of new formats, new metadata terms and new submission processes or requirements. Their goal is to synthesize all ENCODE data into one larger data product. Part of this process is reporting on the progress of labs and the issues that have stalled data submission.

4d. Role in release notes, versioning, track descriptions, and repositories. In addition to experimental metadata, the DCC Wranglers collect more qualitative data on the overall data product, including method changes, relationships to other data products, experimental reasoning, versioning and releases. This information is gathered in the track description pages on the browser and is summarized in release notes and ENCODE Portal news. The Data Wranglers submit the product to the various public repositories and maintain links between them.

B.2.b **Aim 2: The DCC will develop new technologies and adopt existing technologies to facilitate access and integration of ENCODE data by the biomedical research community.**

ENCODE data, created at local provider sites, will be organized into web-accessible distributed data hubs, one for each production lab. These hubs, containing data in standard genomics file formats, will be created by DCC client-side software with DCC Wrangler assistance as needed (Aim 1). After further automatic and manual quality assurance steps, the data will be integrated into the central data repository, the DCC Big Data Hub, where it will be accessible as files in standard formats, via web services, and by visualization tools (such as the UCSC Genome Browser) that support these standard formats (Figure 5). In addition to the genomic primary data (including raw as well as processed data and peaks), metadata sufficient to search, organize, and manage this data will be collected in AnnoDB, our relational metadata database. The primary visualization tool for ENCODE data is the UCSC Genome Browser. The Browser will be extended as necessary to support data formats and visualization methods used by the ENCODE DCC as they arise. Other visualization tools are implicitly supported, by providing the data in commonly used file formats. New and existing software technologies for data access, integration and analysis will be brought into the various pipelines as these technologies become standard. The DCC is committed to using existing, standard data formats (many of which were invented at UCSC for the Genome Browser). If it proves necessary to develop new data formats, any software developed by the DCC will be made freely available via an open source license. Allowances will be made in all software engineering efforts for integration of tools that provide data via new Internet and web standard protocols.

1. File system in the Big Data Hub

Data files that have passed preliminary verification checks (via DCC-provided software; Table 3 and Appendix 2.1) will be incorporated into the DCC Big Data Hub, physically located at the UCSD Supercomputer Center (SDSC) (Figure 4). The complete metadata, including cell lines, antibodies and other reagents, and references to experimental protocols, will be added to AnnoDB. The primary data provided by the ENCODE project is large (estimated to be more than 1 petabyte by the end of the project), but relatively simple and stable. It is organized into large files in one of a half-dozen well-defined formats that essentially represent quantitative data across the three billion nucleotides of the human genome. Since the storage of billions of reads or genome coordinates are ill suited to a relational database, due to unacceptably long indexing times and a substantial increase in storage requirements that are already quite demanding, data from these files will not be stored in a relational database. The AnnoDB system will store only references to the primary data, referencing database unique identifiers to the permanent URL of each resource (i.e., data file on the Big Data Hub), and MD5 checksums to ensure data integrity.

2. AnnoDB – the Metadata relational database

The conditions, parameters, and protocols used for conducting and analyzing experiments are collectively called metadata. The metadata, in contrast to the data files, is of modest size (less than 100 megabytes we estimate) but is complex and can change rapidly. Flexibility in the metadata is a feature; it is designed to be extendable, as experimental protocols improve or are updated, and new types of experimental methods are added. In principle, AnnoDB must allow for the storage of metadata for experiments that have only been designed and not executed. The current DCC records 64 fields of metadata in 2011 (Appendix 1.4), up from 15 at the start of the project in 2006. This combination of modest size, high complexity, and rapid addition of new fields is handled well by a relational database. Currently, ENCODE DCC metadata is stored in a text file format. This format is used by the UCSC Genome Browser, and works passably well in that system as long as the number of data tracks is no more than a few hundred. However this system has become increasingly awkward as the number of experiments grows. Furthermore, because there is no tracking system, it is quite difficult to revert, alter, or modify the metadata for a particular experiment if there is a mistake or update.

Methods for maintaining metadata consistency and integrity, essential for comparing or reproducing experimental results, are fundamental aspects of modern relational databases like Oracle (http://www.oracle.com) or PostgreSQL (http://www.postgresql.org). Using built-in tools, such as triggers and constraints, data validation and verification will be applied every time data is loaded or modified in the database. Formal controlled vocabularies, such as from the OBO Foundry (http://www.obofoundry.org/), as well as internal coded values, for example "status", will be stored in the database and used for validation. The database is also ideal for tracking pipeline steps, QC processes, and data file versions. An audit system will identify when and who inserted, updated, and obsoleted a record from the database. Fine-grained database permissions to access and manipulate data can be assigned on the individual, laboratory, or community level. The Structured Query Language (SQL) provides a powerful method to retrieve complex associations between data. SQL queries can also reveal relationships and inconsistencies between data types that can go unnoticed using other storage methods. Targeted indexes, table partitioning and materialized views are just a few ways that data retrieval can be optimized. The power of a relational database to search selectively on combinations of fields is helpful in many contexts. Many bioinformaticians already know SQL, so little training is required of Data Wranglers and power users. Relational databases are standard enough that interfaces are available in virtually all programming languages, allowing the DCC to create APIs to support the diversity of bioinformatics programming environments as needed. Indeed, for many queries, only a SQL command is needed, to create a table with very little software overhead.

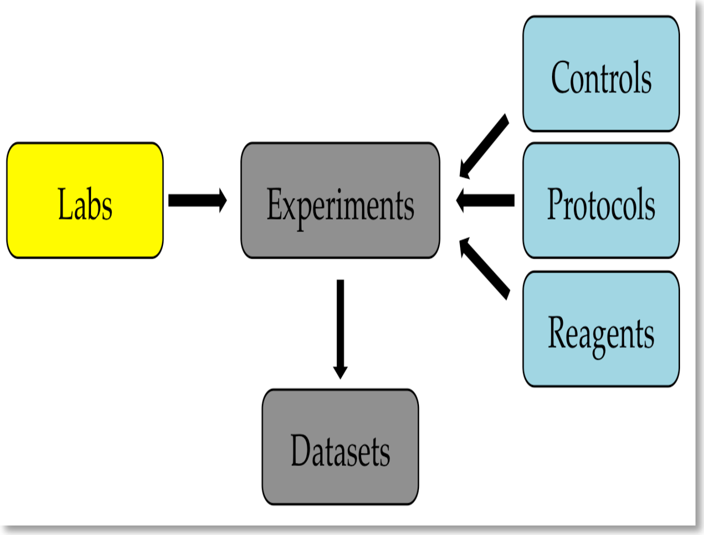


Figure 6: A representation of a possible schema for AnnoDB, demonstrating the ability to capture the metadata for a wide range of data files. *Labs, which can be data production groups or analysis groups, generate experiments. The experiments that are carried out involve controls, protocols, and reagents. Multiple experiments can be grouped together to define a dataset. For a lab experiment, the protocol could be a chromatin immunoprecipitation assay using antibodies to GATA-1 in K562 cells. A dataset could be defined as all the Level 1-3 data files that are generated by this lab experiment (see Fig 4 for definitions). Alternatively, a computational experiment that is performed by an analysis group would define an algorithm as a protocol that uses multiple data files as reagents (or inputs) into their experiment.*

The modENCODE DCC metadata is stored in a relational database using the CHADO schema [45, 46]. Unfortunately the design constraints of CHADO, to be widely used by a variety of biological databases, are such that much of the advantage of a relational database is lost. Instead of having separate fields for each metadata term, in CHADO, the metadata is stored in a generic format where one column specifies the metadata field, and another the value for this field. The result of this is that the user can't know what fields are available simply by consulting the schema. Queries on what is conceptually a field, end up being complex to code and take much longer to execute than they would in a relational schema that was customized to the actual metadata used for the project, rather than a generic one. Therefore, the AnnoDB database will be custom designed and extended specifically to best serve the needs of the ENCODE project and users of its biological data (Figure 6 and Appendix 3.1).

3. Quality Assurance of the data

While the Data Wranglers themselves provide many checks on the production groups, their work involves many interactions with databases, configuration of software. The wrangling work itself is subject to a degree of error. For this reason a separate Quality Assurance (QA) group will also check the final output, using tools, such as the ENCODE Portal and the UCSC Genome Browser, to simulate the user rather than the Wrangler experience. The QA group checks the functionality and content of the new data added to the UCSC Genome Browser, focusing on functionality, content, and performance (full details of checks are described in Appendix 2.2). As the scope and complexity of the data submitted to ENCODE increases, the QA group will coordinate with the software group to define automated summaries, such as the expected number of datasets per cell line or experiment type, to facilitate and expedite the manual review of the data exported into the UCSC Genome Browser. These automated QA checks ensure data integrity in the Big Data Hub and AnnoDB as well as allow specific regions to be visually checked for accurate display of the data and metadata.

4. Software management and quality assurance

All ENCODE software will be Open Source and freely available. Source code will be managed in a repository developed around the Git fast version control system (http://git-scm.com/). Access to the repository will be provided via standard Internet protocols FTP, HTTP, and specialized Git clients that are available for a variety of computer operating systems. The DCC will require all software developed for the EDCAC, and the ENCODE Analysis Working Group (AWG) to be distributed via this software repository. Easy availability of analysis and data processing software is critical to confirm reproducibility of all software-assisted experimental results. This includes essentially all genomics experiments. All DCC software will include a sufficient level of documentation to allow installation, detailed list of all software dependencies, and expected input/output with available options sufficient to reproduce calculations and results on 3rd party computers.

When developing new software the DCC will use models that have worked on commercial software development, utilizing the experience of the UCSC software development team, many of which come from this background. The DCC will follow models for software design and good practices that have succeeded before in large software projects spanning millions of lines of code and decades of maintenance experience. The large user base of the ENCODE project dictates a high level of robustness, higher than is necessary with most academic software.

An incremental approach to development will be followed, in which software is developed in small pieces that incorporate testability as part of the design. This ensures that too much time and other resources are not invested on a project that turns out to be intractable, and prevents simple improvements from being delayed while complex improvements are implemented. A three-week software release cycle will be maintained. Changes that require more time to complete are implemented on a separate branch in the source code control system (Git), and then merged into the main branch when ready, accompanied by additional testing. Our experience in maintaining software for academic and research (approaching decades in some cases) leads us to design software that is both flexible and maintainable, even without availability of the original authors. This flexibility and maintainability is supported by our commitment to professional quality software engineering practices and coding standards.

Three software engineering practices merit particular discussion: an independent quality assurance group, unit testing, and code review. The project will expend significant resources on these, and the payback is well worth the investment in terms of quality and long-term maintainability and productivity, where development is not hindered by the difficulties of building on an unstable foundation.

At UCSC Kent has found, as IBM did 40 years ago, that it is extremely effective to maintain a quality assurance (QA) group that is separate from our development group. This group is staffed by people who are motivated to find bugs, in contrast to software developers who tend to be less enthusiastic about finding problems in their code. By having a separate QA group, targeting hiring of skills specific to software testing, and are able to somewhat reduce our personnel costs since testers tend to entail less budgetary impact than programmers.

Unit testing is the practice of developing automatic tests for a piece of software at the same time that the software itself is being developed. These tests are developed by the programmer working on the code, or by his/her partner when doing pair programming. These tests help ensure that the software works as designed initially. Furthermore, as new functions are added to the application, and as bugs are found and fixed, new regression tests can be added to ensure the same errors do not occur over and over again. Even more important, since these tests are automatic, they can be run regularly, and help ensure that the software keeps working, even as it is changed to extend it to new circumstances. Given the rapid development of the genomics field, we fully expect software requirements for the DCC to change over the 4-year time span of this award. Unit testing is not a substitute for a separate QA group, and in fact tends to catch fairly different classes of bugs. Automatic unit testing tends to catch bugs within a particular programming module, while the QA group will tend to mimic demanding and somewhat chaotic users of a system, and often catches bugs that occur in the interaction between modules. However, bugs found by the QA group can often be turned into tests that are added to the unit test scripts, preventing them from re-occurring.

A third practice that will be rigorously maintained is ubiquitous code review: another programmer reviews each line of code that a programmer commits to our source code control system. This helps catch certain classes of erratic bugs that are not easily caught by the QA staff, who for the most part are not looking at the source code but rather using the programs. Code review also helps diffuse the knowledge of the code throughout the organization, making it more likely that a programmer will reuse a colleague's piece of code rather than reinvent it, and making it easier for one programmer to extend and maintain another programmer's code. Code reviews also help ensure that style, commenting, and naming conventions are applied uniformly through the code, making it more readable and understandable (particularly to people getting started on the project). If a new software task includes user interface elements, or when a major module is added to the system, the programmer is required to review the initial designs with at least two other people charged with representing the user's perspective and maintaining the website's consistency.

The UCSC software development team has be successful in creating important software used by the genomics community. However, the DCC will make use of existing software when possible. For example, initial design work for this proposal indicates extensive use of the Django (https://www.djangoproject.com/) web framework would be appropriate for development of the ENCODE Portal, VisAnt [47] for viewing gene-gene interactions, and modMine [32] as an interface to annotation data. Our QA group will do user-level testing of the tools when integrated, and provide reproducible problem reports as needed to the developers. The DCC will likely not have the resources to contribute significant amounts of new code to these open source projects, of course software created will be shared, and we anticipate that some of these contributions will include fixes to bugs we find in these tools.

The DCC does not have the luxury of hunkering down in a software development bunker for 1-2 years designing and creating a perfect system for handling ENCODE data. ENCODE data has been rolling in continually and we expect it to only increase. We will devise a process to gradually transition the current data submission pipeline and Portal interface to more stable, rigorous and functional versions. The core of this system is a new metadata database AnnoDB (Annotation Database) that is sufficient both for current needs and flexible enough to handle unknown new metadata and data types in the future. The second critical step is the establishment of a secure and robust versioning system to allow all classes of users to track the progress and correct errors in submitted data. Once this database and a common programmer interface is defined and implemented, a myriad of helper applications can be written that access and view the metadata. This includes project specific submission interfaces and Data Wrangler "curation" interfaces to track and modify metadata.

B.2.c **Aim 3: The DCC will design, develop and maintain the ENCODE Portal.**

Data files and metadata that have been certified, validated, verified, and stored in the DCC Big Data Hub and the DDC annotation database, AnnoDB, will be accessible via the ENCODE Portal. The ENCODE Portal will be the primary access point to the EDCAC. This multifunctional data portal will be designed, developed, and maintained for the searching, filtering, and distribution of data, metadata, and tracking information.

1. The ENCODE Portal will provide multiple entry points to the data

Data files (Level 1-3 data), metadata, and analyses of the data (Level 4 data) will be accessible via diverse search queries at the ENCODE Portal. The range of options to query and download data generated by the ENCODE Project will reflect the different scientific communities that will benefit from these data and the diverse biological questions they ask (Table 4).

Using advanced filtering options available at the ENCODE Portal, the metadata used to describe the experiments can be searched and filtered to retrieve a specific set of functional elements identified in the genome as well as the data files that provide the experimental evidence supporting their annotation. A biologist may be interested in only the integrated data (level 4) for a subset of cell types or genomic regions. The researcher may only request the regions in the genome that contain transcription factor binding sites and DNase hypersensitive sites in cardiac cells. Alternatively, a researcher may desire the coordinates of all active promoters that were predicted in specific regions of the genome. These queries would produce a short list of results that can be browsed on the web. A computational biologist, however, may choose to download all the sequence reads generated and aligned to specific regions bound by particular antibodies in a ChIP-seq assay.

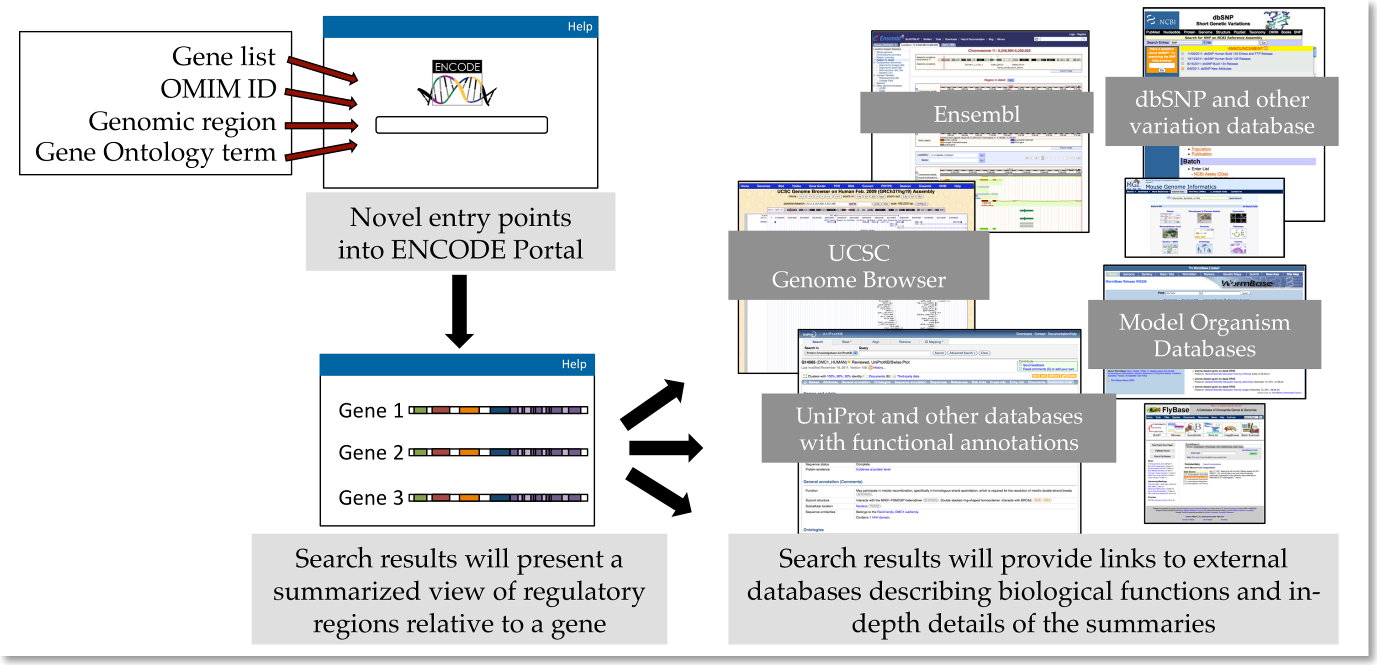
Table 4: Example use cases describing searches that can currently be performed at ENCODE and at modENCODE, along with types of queries that will be implemented at the ENCODE Portal.

Type of query	Can be performed at ENCODE?	Can be performed at modENCODE?	Will be available at ENCODE Portal?
Transcription factor binding sites adjacent to a gene	•	•	•
All ChIP-seq experiments performed in a specific cell type	•	•	•
All experiments performed by a single lab	•	•	•
All genes located within 10kb downstream of where a specific transcription factor binds throughout the genome		•	•
Regions in the genome having co-occurrence of TFBS and DNase hypersensitive sites in a specific cell type		•	•
Producing custom tracks based on filtering and intersecting with other tracks	•		•
Genomic coordinates for all active promoters upstream of a list of genes			•
Download all datasets that were used for a specific publication			•
Transcription factors binding within a defined distance upstream, downstream or within of genes annotated to a specific GO process term			•
Associate distal TF-gene interactions with an OMIM ID			•

Since the identification of regulatory elements in the human genome can only be fully realized when integrated with gene-based functional annotations, the ENCODE Portal will extend beyond the queries currently available to enable discovery based on gene names, gene symbols, unique IDs from the GENCODE project, NCBI, Ensembl, HGNC, UniProt, and controlled vocabulary terms that describe a gene's function or biological role, such as Gene Ontology or OMIM (Figure 7) [3, 37, 48-52]. This extension increases the accessibility of ENCODE data to researchers who are interested in whether a list of genes shares a common set of regulatory elements or the physician who queries the data using an OMIM ID because a patient may have a rare SNP in the regulatory region of a disease associated gene. Query results will summarize and integrate ENCODE results with annotations from Ensembl Regulation [53] as well as functional annotations from external sources, such as the GOA project or dbSNP [51, 52].

The massive amount and variety of data provided by the ENCODE project requires extremely fast indexed searching; "facets" to limit and filter search results, and an intuitive and functional user interface. Commercial shopping sites have had great success using this strategy to enhance access to huge product catalogs. The software team at the DCC will create all of the above features in a robust design process. In addition to the web interface itself, web services will be provided to programmatically access both data files and metadata describing those files. Each data file in the ENCODE repository will be accessible by a unique URL and made available via standard protocols such as HTTP and FTP. Metadata can be returned to a Web Service user in a variety of formats, including plain text, JSON, or XML.

Figure 7: Sample workflow of extensions to the ENCODE Portal. *Entry points into the Portal are supported using a variety of identifiers, such as Gene Ontology, OMIM, UniProt protein accessions, HGNC gene symbols, dbSNP reference cluster, PubMed, MOD genes and coordinates for model organism data, to name a few. In addition to the information shown the Portal will provide a central source for many reports and documents. This will include technical reports on data production methods, protocol documentation provided by the production labs, data standards developed by the ENCODE project, bibliography of ENCODE-related publications including those external labs using the project's data.*


Gene list
OMIM ID
Genomic region
Gene Ontology term
Help
ENCODE
Novel entry points into ENCODE Portal
Help
Gene 1
Gene 2
Gene 3
Search results will present a summarized view of regulatory regions relative to a gene
Ensembl
dbSNP and other variation database
UCSC Genome Browser
Model Organism Databases
UniProt and other databases with functional annotations
Search results will provide links to external databases describing biological functions and in-depth details of the summaries

A powerful new code feature will be provided that will simplify the management of sets of data files. The Portal will work directly with lists of file IDs as input from users or defined by search results. The ability to retrieve any defined set of files enhances how sets of files can be defined, shared and searched. The list can include specific file versions or the specifying latest version. The retrieval simply requires the file name, equivalent to a unique ID or thing of it as an accession number, to be provided in a list. The requested files could be retrieved as a tar file and downloaded. This will provide the basic mode of retrieving a set of files whether from an input list or the request of a database query. The list can be saved to define a specific version of data used in an analysis. The list can be shared with colleagues to allow them easy and specific access to a set of results. Metadata associated with a list of files, datasets, will allow a new ability to dig deep into the extensive ENCODE data.

2. The ENCODE Portal will allow customization of user experiences

We divide the users of the Encode Portal into five roles: Casual (guest), Power, Provider, Admin (Wrangler) and Oversight. These roles are defined programmatically within the Portal login system that will allow or deny read/write access to various components of the data and metadata, depending on the specific role. Roles are controlled by standard username/password authentication methods to assure those viewing the data have the appropriate credentials.

The least specific are the casual users, who are usually looking for one or more essentially random pieces of information. These users will have open access to all public data and tools and thus do not need to provide a login unless they wish to save search preferences. A casual user may only use the site once in a year to look something up or to create a figure for a presentation. There is little guideline or feedback from typical casual users on what they want or what they are looking for, thus the Portal will be designed with a "Google" like search interface to facilitate browsing of broad features and topics that are commonly desired as output. Casual users would only have access to fully released results (although there is no restriction on who can access preliminary results in effect becoming Power users). "Power" users on the other hand come to the Portal for very specific purposes. They might, for example, need all results, final and provisional, for histone modification marks. Power users have their own login account (using OpenID or a similar system) and can save preferences for their searches and set up alerts on types of data. They log in frequently (possibly once a month or more) and we expect that they will complain loudly if interfaces are slow, unnecessarily complicated, or opaque. Many Power users will eschew a web-browser interface and instead access the ENCODE data programmatically via the Web Services we provide nonetheless they an important source of feedback on the basic web interface. The use of Web Services is particularly true of bioinformatics analysts. The AWG are of the Power user class, although if necessary can be granted additional access.

A "Provider" user is an informatics contact from a production lab. They will have Power user access to all data, and special access to their own lab's data. Generally this will include viewing all versions (including "deleted" or "deprecated" data files; no data will actually be deleted once it is released as a preliminary result). Wrangler or Admin users have the privileges to update, modify, or deprecate data from any data provider.

Contributors and managers of specific ENCODE projects require real-time reporting on project status. This information will also be available from the Portal. The status of all data submitted to the DCC will be tracked by the DCC relational database and software will be developed that can access this information to provide live tracking of specific submitted files or applying any of the filters presented earlier in this aim. Monthly reports will automatically generate reports that detail the progress of the DCC. The real-time nature of these reports and the use of web services will allow groups with access to develop automatic reports as they desire.

NHGRI or others as defined by program staff will have access to summary reports for all data, which defines the final "Oversight" role. These reports will include exportable data tables and charts, and the Oversight role will have read access to the entire data pipeline for an experiment or group of experiments. We will automate much of the monthly report generation process using these tables and charts.

3. The ENCODE Portal will provide extensive documentation

The Portal will also include extensive documentation and tutorials for the use of the Portal, the ENCODE components of the UCSC Genome Browser, European Bioinformatics Institute (EBI) and National Center for Biotechnology Information (NCBI) tools. The most effective tutorials are short videos of about 1-2 minutes in length. We will also provide presentations than can be used for teaching and self-study. Additional details are described in Specific Aim 8.

In addition to providing data access to the greater scientific community, the ENCODE portal will provide documentation on the progress, mission, publications and discoveries associated with the project.

B.2.d **Aim 4: The DCC will provide enhanced access to data and computing environments via the cloud and Galaxy.**

Distribution of all levels of data maintained by the DCC will be provided via standard Internet protocols from a state of the art computing facility. This will include HTTP, FTP, and cloud protocols such as Simple Storage Service (S3, http://aws.amazon.com/s3/). The DCC will maintain a Cluster of Computers (DCCCC or DC4) in the same location as the Big Data Hub for use by the AWG, so that AWG researchers need not download the immense amount of data required for some of their tasks. The DC4 will be configured for flexibility as a cloud, providing researchers with virtual machines conforming to their own computer configurations. The DCC will set up an instance of Galaxy with the ENCODE data built-in, and that will be able to use the DC4 as a computational backend. The DCC will make arrangements with various large cloud computing facilities, both commercial and academic, to provide dependable and efficient access to the data throughout the world.

1. Background about the cloud

"The cloud" is a deliberately nebulous term covering several increasingly useful information technologies. The cloud was developed originally in conjunction with web sites, particularly Amazon, Yahoo, and Google. The unifying theme of the cloud is to make the resources from a large data center available on demand. This leads to efficiencies of scale because large data centers are cheaper to build and maintain than multiple small ones with the same capacity. It also leads to a peak-leveling effect where, since the peak usage of all users in the cloud rarely if ever coincides exactly, the peak demand for machines in the cloud as a whole is smaller than the sum of the peak demands of all of the individual cloud users. Thus the large data center need not be as big as the sum of all of the smaller data centers it is replacing to serve peak demand. The combination of efficiencies of scale and peak usage averaging has allowed the various cloud providers to make a profit while saving their customers money.

The key technologies behind the cloud are virtual machines for computation, and URL based storage for data. There are several variations on these technologies and how they are packaged together. Amazon was the first to widely market their cloud services as EC2 (Elastic Compute Cloud, http://aws.amazon.com/ec2/) on the computing side and S3 on the data side. Google's cloud operations emerged somewhat later, and have not caught on to the same degree as Amazon's, in part because while a computer in the EC2 can be viewed for the most part as just a remote Linux box (or more recently a Window's box) configured according to your specifications, the computing environment in the Google cloud is particular to Google. Yahoo for the most part has followed the pattern of the Amazon cloud, and has been instrumental in developing open source versions of the Amazon systems, many of which are included in the OpenStack (http://openstack.org/) protocols we plan to support at the DCC.

Though the cloud was originally developed for web oriented computing, it has proven useful in many other areas. In particular, it shows great promise as a way of enabling smaller computational labs to take advantage of resources hosted in larger computational groups. However cloud technologies do not, at least as yet, solve all of the problems one encounters in building flexible high throughput systems for bioinformatics. In particular, computer programs that have both intensive I/O and intensive CPU requirements remain a challenge, better addressed by Google's software than Amazon's or Yahoo's. As a result the DCC plans to take a fairly nuanced approach to cloud computing, to support indirectly new research in I/O intensive cloud computing, and to offer limited access to the DC4 via more direct non-cloud methods.

2. Use of the cloud for data storage

The master copy of the ENCODE data will be kept at the San Diego Supercomputer Center (SDSC) on a storage cluster running IBM's General Parallel File System (GPFS, [54]). This storage system, in UCSC's experience, is able to keep up with the demands of 100s of I/O intensive programs running simultaneously. It is possible to add capacity to it without down time. UCSC has used it for five years under demanding conditions without any data loss. GPFS is a POSIX-compliant file system, meaning that, unlike S3, it can be used directly by existing bioinformatics programs without modification. The ENCODE files will be available to Linux machines inside the SDSC including web servers and the DC4 simply as normal files. The web servers will provide outside access to the same files via HTTP, HTTPS and FTP protocols. We will also provide external S3 access via OpenStack software for cloud access.

While GPFS is capable of handling the requirements of 100s of I/O intensive programs, performance degrades unacceptably when faced with the demands of 1000s of programs in this class such as short read mappers and peak callers. The approach that comes closest to working for these programs on the scale of 1000s and beyond is the Map/Reduce approach (http://hadoop.apache.org/mapreduce/) pioneered commercially by Google, and developed in open source by the Hadoop project (http://hadoop.apache.org). The Hadoop system is designed to work with large numbers of modestly powered computers including both disk and CPU in the same box. The Hadoop File System replicates data in three places, and the Hadoop scheduler tries to execute the program using a CPU in the same box as the data, or failing that at least on the same local network as a computer with the data. Hadoop is good at what it does, but has several significant limitations. Programs using Hadoop must use Hadoop APIs to access data rather than regular POSIX-compliant APIs, and so, as with S3, existing programs can't be used without modification. The replication requirements of Hadoop also require significantly more disk space to be purchased than for other file systems. For very commonly used data sets, even three replicates are not enough to ensure that each time the data is needed a piece of that data will be near. The calculation of sequence alignments typically requires two or more significant pieces of data and finding those two pieces of data together is rare, obviating the data locality that is Hadoop's main advantage. The ability to analyze genomic data is a significant challenge for the Hadoop style of computing.

3. Use of the cloud for data backups

We will mirror the data at the Open Science Data Cloud (OSDC) Chicago Data Center. The data will be available there via HTTP, FTP, and S3 mechanisms just as it is from the SDSC (Figure 8). The data will also be available internally at OSDC via Bionimbus protocols. Bionimbus (www.bionimbus.org) is a project organized by Robert Grossman, University of Chicago, which promises to address many of the limitations of Hadoop (see the letter from Dr. Grossman). Though as a research project many aspects of this problem are beyond the scope of this grant proposal, the ENCODE data set is a perfect test case for Bionimbus. Having the data at OSDC serves the dual purpose of supporting the next generation of cloud computing and being an off-site backup for the ENCODE data, a backup that gets enough use that any data corruption will likely be noticed immediately. Sadly this is often not the case with traditional tape-based backup systems. The network connectivity between SDSC and OSDC is exceedingly good using StarLight connections (www.startap.net), and is currently capable of transferring 100 terabytes/day, and is estimated to be capable of transferring 1000 terabytes/day by the end of this project. This transfer rate will enable full recovery from OSDC in a day.

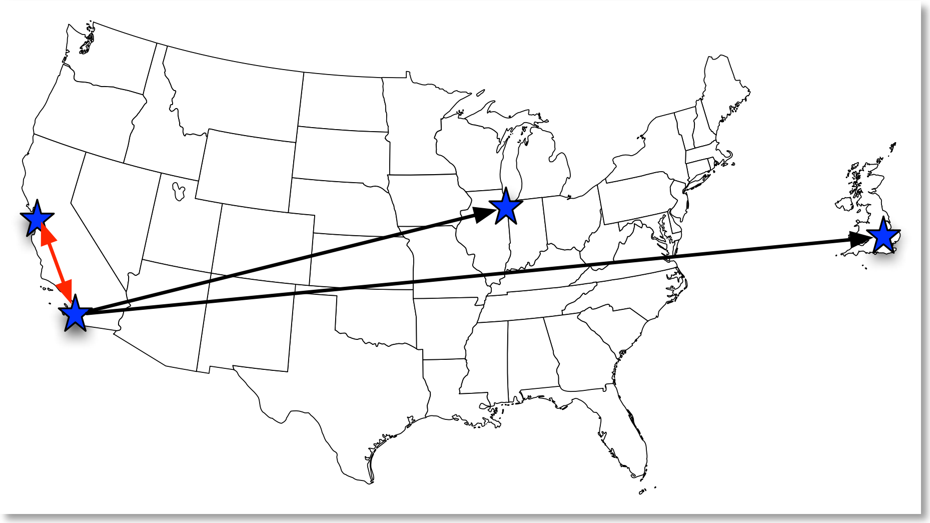


Figure 8. Data is distributed geographically providing more access to the data*. ENCODE data will be located at the San Diego Supercomputer Center (SDSC) and distributed from there to other sites. The black lines represent data transfer from the SDSC to the Open Science Data Cloud (OSDC) in Chicago and data transfer to the EBI in Cambridge UK that will provide fast access to our European colleagues. The red double-headed arrow shows the fast connection between the two primary sites of the DCC (Stanford and UCSC) and SDSC.*

We will also mirror the data at the EBI cluster in Cambridge UK under the management of Ewan Birney. An EBI mirror provides off-continent backup, and providing much better HTTP and FTP bandwidth to Europe. A limited number of accounts will be available to members of the AWG on the EBI's very large computer cluster, involving thousands of computer nodes. The existing ENCODE AWG has found the EBI cluster to be useful for remapping reads with improved algorithms, peak calling, and other tasks that require access to the large, low level, ENCODE data sets, as opposed to the small, more processed data sets that suffice for most purposes. Unfortunately that current cluster environment involves systems that predate the emergence of cloud technologies, and require restrictions to protect the cluster from overuse that could result in unavailability of the systems for EBI as a whole. Therefore only a small number of AWG members, each of whom will require careful training, will be provided accounts at the EBI.

4. Computing with data in the cloud

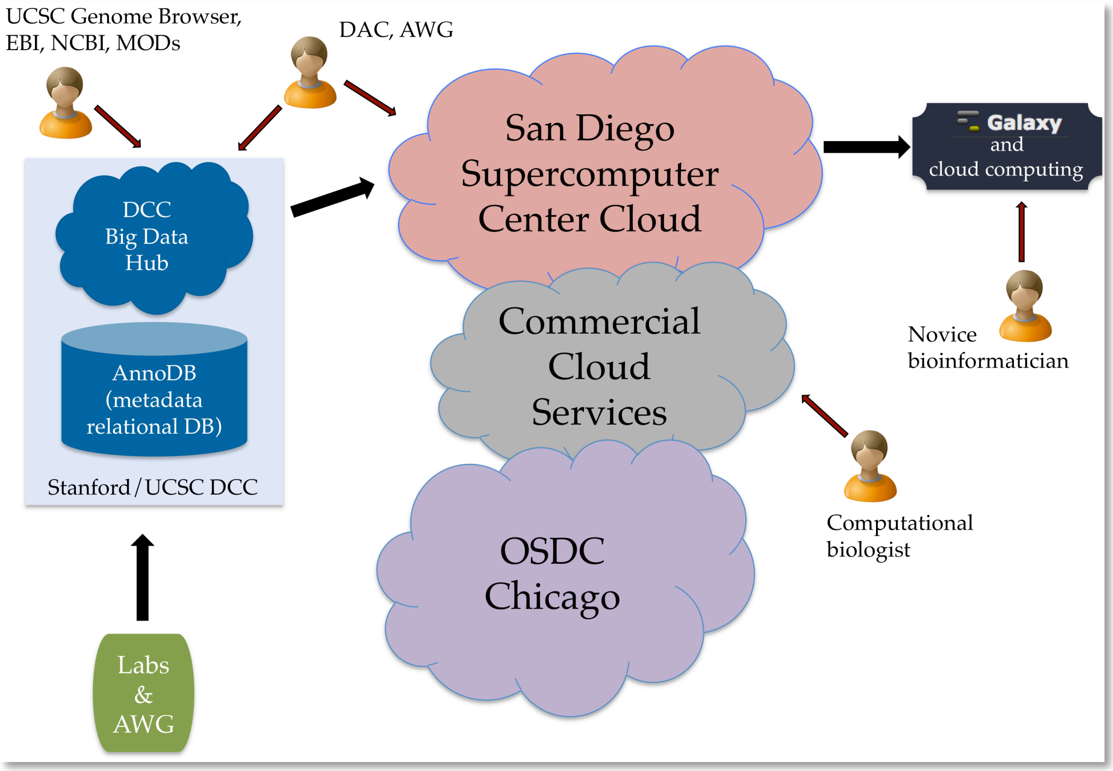
We will also provide access to the DCC specific DC4 cluster to a somewhat larger group of AWG members via EC2 compatible methods as implemented in OpenStack. The EC2 approach enables researchers to start virtual machines conforming to their favorite computer operating systems and configurations, and avoids the need to download the ENCODE data sets to local computers. EC2 by itself only specifies how to bring up a single virtual machine, although there are tools that build on top of this to allow users to build "virtual clusters". Because the DC4 cluster is relatively small, just 128 nodes, and because it is used for Galaxy (see below) and the DCC's internal needs, we are likely to limit the use of the DC4 to no more than 32 nodes per user without prior arrangement.

Cloud resources provided as virtual machines offer enormous flexibility, but still require informatics expertise to use. To make DCC data available to the majority of biomedical researchers who lack this expertise, we will provide data and analysis services through Galaxy. Galaxy [55] makes existing computational tools available in an entirely web based environment where users can analyze large datasets, construct analysis workflows, and easily share results. Galaxy CloudMan [56, 57] couples this accessible interface with infrastructure for scalable computing on cloud resources. The Galaxy instance provided by the DCC will be hosted on the OpenStack cloud described above, with high-performance access to all DCC datasets. We will extend Galaxy to interact with the DCC metadata infrastructure, enabling easy discovery and analysis of all DCC datasets within Galaxy. This will enable users to process DCC data with hundred of analysis tools and workflows already enabled in Galaxy, and to integrate across dozens of other biological databases with which Galaxy integrates. Additionally, Galaxy is designed for rapid integration of new analysis tools; we will work to assist in the integration of key analysis tools identified by the AWG for use by the larger community.

To further enable accessible analysis of ENCODE data by the public, we intend to make DCC data available through commercial clouds, specifically Amazon Web Services. We will request that Amazon host the data as a public resource at no cost as they currently do for Ensembl, 1000 Genomes, modENCODE and other data.

Amazon benefits from such an arrangement because they charge for transferring data from the public data sets into user's local cloud space. Amazon also benefits by promoting cloud computing and from the goodwill hosting public data sets generates. Even if Amazon does not agree, users in the Amazon Cloud will be able to access the data from SDSC, OSDC, or in the future EBI via S3 protocols. In addition, for even more accessibility, we will provide a virtual machine image pre-configured for accessing this data. The CloudBioLinux (cloudbiolinux.org) project currently provides virtual machines with hundreds of common bioinformatics tools preinstalled, including an automated process for building derived virtual machines. We will extend the standard CloudBioLinux image to automate access to the DCC data mirrored within the commercial cloud. CloudBioLinux also integrates CloudMan and Galaxy by default, and we will extend the virtual image to automatically load DCC data and metadata into the Galaxy environment, allowing users to immediately analyze this data through this powerful web interface.

Figure 9: ENCODE data will be moved from the production labs to the Big Data Hub and then injected into the cloud in San Diego. *From SDSC other academic clouds such as at OSDC and commercial data clouds such as Amazon S3 have access to the data. The Galaxy software environment provides many features that simplify use of ENCODE data. These cloud data and computing services enhance the availability of ENCODE results to the great research community.*



B.2.e **Aim 5: The DCC will deposit data into the appropriate public archival repositories, model organism databases, and international bioinformatics resources.**

Each of the four levels of data will be deposited in the appropriate public archival resource. The DCC will prepare the data files in acceptable formats for submission to these resources. The processing pipeline will be enhanced as needed to maintain the full data verification and format specifications required by these public sites. Specifically, the DCC will provide data to NCBI (GEO [52] and SRA [58]), UCSC Genome Browser [22], EBI (Ensembl [51] and ArrayExpress [59]), and the appropriate model organism databases and other resources as specified by the AWG and NIH. The Export pipeline as diagramed in Figure 5 illustrates the conversion of data into formats required for integration into these remote databases. These sites are all archival in nature thus it is essential for the DCC to build close relationships with these resources to insure the high quality and rich metadata are accurately provided via these resources. These facilities will promote the use of ENCODE data by the greater research communities. Several already integrate ENCODE into their user interfaces, enhancing the discovery process by integrating the ENCODE results with other types of data. In some cases these sites provide easy to use forms that have become common analysis tools.

1. Export to NCBI. Experimental results and metadata will be submitted to GEO and the SRA at NCBI. Currently, all ENCODE data in BED, bigWig, BAM, and FASTQ files are submitted to GEO, which will submit the FASTQ files to the SRA. Metadata is automatically mapped by DCC to the SOFT file format used by GEO and other resources [60]. We will continue to enhance and optimize this process. While much of this data transfer and metadata conversion is automatic, there are several continuing needs that require a full-time Data Wrangler to interface with the NCBI, including defining mappings between ontologies and controlled vocabularies used by the ENCODE project and controlled vocabularies required for the SOFT file format. We understand there is a similar commitment at NCBI for the receipt and integration of ENCODE data into their databases.

2. Export to EBI. Experimental results and metadata will be submitted to ArrayExpress and Ensembl at the EBI. Quality control checks will be run at EBI on the read level data (Level 1) to confirm that the reads are derived from the cell line (using SNP-based characterizations) and examine the reproducibility of assays (using the IDR method, [61]). These data will then be integrated into Ensembl to provide a gene-centric view of the ENCODE results. Incorporation at EBI will require coordination and pre-registration of the cell line sample information into the EBI BioSample database[62], which is regularly exchanged with the NCBI BioSample database. Inclusion in the EBI BioSample database requires curation of the metadata in the Experimental Factor Ontology (EFO,[63]). The integration and metadata curation work will be handled by the EBI group via a subcontract with Ewan Birney's group. Birney's group will have direct access to the DCC Big Data Hub. The procedures for effective data transfer and subsequent data conversion will be defined and developed by the EBI group as is appropriate.

3. Export to the UCSC Genome Browser. Data integration into the UCSC Browser will be conducted by this project in two ways. All ENCODE data will be presented in the Big Data Hub allowing selected data to be imported into the Browser by interested users. A subset of the data, in particular those created by the DAC and DCC, will be included as integrated views and will be loaded into the Browser's standard built-in tracks. The metadata for these integrated views of the data will be converted to the text format used by the UCSC Genome Browser.

4. Export to Model Organism Databases (MODs). There is no single web resource available to explore the human genome. Currently the user must query and integrate data presented by Ensembl, UCSC browser, and NCBI to retrieve what is known about a region of the genome, or its genes and their regulation. Thus it is essential that the ENCODE Portal combine information from a variety human resources including UCSC (co-Investigator: Kent) and Ensembl (co-Investigator: Birney). However, well-developed databases already exist for *Mus*, *Drosophila*, and *Caenorhabditis* information and the new ENCODE results should thus be integrated in their respective Model Organism Database (MODs). All new produced *Mus*, *Drosophila* and *Caenorhabditis* data will be provided to the appropriate community database, MGI [64], FlyBase [27], and WormBase [65] respectively.

The DCC will use the same certification, validation, and verification pipeline used for human data on all submitted *Mus*, *Drosophila* and *Caenorhabditis* data generated by the ENCODE production labs. The DCC will format all the data in the same standard formats and provide consistent metadata as specified for the human experiments with extensions needed for that model system. However, each of the three MODs is built from different software architectures, search interfaces, and visualization tools. Due to these differences it is difficult for the DCC to directly provide assistance for the ENCODE data at the MODs, however we will work with each MOD to ensure the data we provide can be incorporated into their systems appropriately.

Therefore, we are proposing that one FTE Data Wrangler will manage MOD-specific data issues. This is not only transferring data to the MOD but also includes providing expertise in organism specific nomenclature and biosample details that will be captured by AnnoDB. The DCC will act as a liaison between the EDCAC and the MODs in order to ensure the best possible integration of the experimental data and the rich metadata at that MOD. The range of activities will include, but are not limited to, any of the following: providing assistance to the MOD staff on all aspects of the samples metadata, work with the MOD and DCC software engineers to create pipelines for automatic download of information, obtain the controlled vocabularies and ontologies that are specific to the organism and integrate them into the metadata the DCC collects, suggesting methods to take advantage of the richness of the dataset and promote their use by the MOD communities, summarizing the experimental data and metadata to be consistent with the MOD's representation of chromosomal features, ensuring the experimental data are mapped to the current version of the reference genome, and exporting the results in standard file formats to the MODs.

Interactions with FlyBase and WormBase will be better defined after we understand the number of fly and worm projects that will be part of ENCODE. It is too early to estimate the volume of model organism information that will be generated in the next round of ENCODE. The PIs of the MODs are eager to work with us on new data for their resources, see the letters of support from Janan Eppig (MGI), Bill Gelbart (FlyBase) and Paul Sternberg (WormBase) indicating their dedication to this effort. To enhance the integration of ENCODE datasets by the MODs, we suggest the NHGRI consider supplements to each of the MOD projects. The DCC will aid the MOD with data formats and mappings; however, we cannot realistically help them in modifying their displays or with the integration of the data into their databases. Some MODs have already begun hosting results similar to that created by ENCODE and some have not yet reached that point. We feel it is necessary for the ENCODE DCC to make every effort to assist the MODs in obtaining the data in forms that can quickly integrate into their resource. The quicker these data are provided to the model organism communities, the greater the success of ENCODE. The DCC will integrate the *Mus*, *Drosophila* and *Caenorhabditis* data with the AnnoDB and Big Data Hub, and thus all resources downstream of there. We will help the MODs in using the Web Services, AnnoDB queries and URL file access to enhance their resource.

B.2.f **Aim 6: The DCC will incorporate and maintain previously produced data from ENCODE, modENCODE and other genomic projects.**

The new DCC will incorporate existing data from other projects, striving to unify the metadata across multiple projects with a common set of controlled vocabulary terms and version control. The use of a single set of metadata standards will increase the power of searching for similar datasets across multiple organisms. Integrating these other data into the new DCC environment and into the ENCODE Portal will allow all these data to be searched and browsed from the same site.

This include human and mouse data maintained by the current ENCODE DCC project at UCSC and the modENCODE DCC at Ontario Institute for Cancer Research, Lawrence Berkeley National Laboratory and University of Cambridge [66]. Human and mouse metadata will be converted into a newly defined set of controlled vocabulary terms as described in Aim 1. The data files will be migrated into the new file system as outlined in Aim 3. In addition, the metadata for *Drosophila* and *Caenorhabditis* datasets will also be converted to the newly defined controlled vocabulary if needed and added to the AnnoDB database. As described in Aim 3 we will provide a file storage structure that will enable us to provide several virtual views of the data organization and file naming via Unix soft links. Thus using our new system we will appear to maintain the modENCODE DCC download directory structure but it will actually be in our new structure at the Big Data Hub. Thus, there should be minimal impact on the current modENCODE users.

We anticipate being able to incorporate the majority of the current data provided by the modENCODE DCC before February 2013, the end date recently proposed by Lincoln Stein, the modENCODE DCC PI. If this timeline is approved, there will be an eight-month overlap between the new EDCAC and the modENCODE DCC, providing enough overlap for a smooth and faithful transition. During this time of overlap, migration of the modENCODE data will be coordinated with the modENCODE Data Wrangler to ensure accurate metadata mapping into AnnoDB and the data transfer into our new file system. Translating the modENCODE metadata will require different term mappings as compared to the human and mouse ENCODE data because the modENCODE metadata standards use slightly different vocabularies and, of course, are for very different organisms and anatomies. However, the assays and file formats are identical to the ENCODE project. The majority of the conversion will be automated, however as with all such tasks, manual mappings between different controlled vocabulary terms and ontologies may be required plus errors will need to be manually reviewed before we can process integrating the dataset into the new DCC.

The funding for the modENCODE production labs will end in March 2012. In the months following March 2012, the modENCODE AWG will continue working with the data. All of the integrative analyses will likely not be completed by February 2013. In which case, the modENCODE AWG results, as well as updates to current files, will become the responsibility of the new ENCODE DCC. In addition to newly submitted results from the modENCODE project we must assume that there will be *Drosophila* and *Caenorhabditis* projects funded as part of the new ENCODE U54 RFA to allow appropriate planning. Key components of the existing modENCODE data submission pipeline can be moved to Stanford to allow newly funded U54 projects access to the pipeline they have previously used. However, all funded U54 projects will need to be trained to use new software interfaces and to learn the newly defined metadata requirements. The distribution of effort between incorporating current modENCODE metadata and data files, transferring the current modENCODE data submission pipeline, and training the new U54 projects will be decided once we learn the number of U54 labs that will be generating model organisms data.

We will also incorporate epigenomics data such as from the NIH Roadmap Epigenomics project (see letter from the Epigenomics EDACC director Aleksander Milosavljevic, Baylor College of Medicine, Houston) [67]. Much of these data are already available from the UCSC Genome Browser and can be easily incorporated using the translation scripts to be created for the ENCODE data. In addition to the Epigenomics project's EDACC these data are available from several other sources including the Washington University, EBI and NCBI. Our efforts will focus on mapping the metadata into the new AnnoDB. We have had initial discussions with Ting Wang at Washington University in St. Louis and Aleks Milosavljevic at Baylor (see letters of support) to collaborate on effective procedures for integration of the Epigenomics data into the DCC. Wang has extensive experience with the data and provides additional insight for the presentation of these data. The epigenomics metadata standards are being enhanced and we look forward to working with the international consortium producing these data. If the ENCODE DCC is asked to maintain the Epigenomics project's data, we will be prepared. The Epigenomics EDACC with the International Human Epigenome Consortium (IHEC) standards for the their data have been proposed [68]. We will integrate these standards as we define the new metadata standards for ENCODE data.

The incorporation of these previously produced (ENCODE, modENCODE and Roadmap Epigenomics projects) data will require extensive efforts from Data Wranglers, Software Engineers and QA Engineers. Data will require shepherding though the newly defined validation pipeline and verification. Because these data were not submitted to our enhanced pipeline we anticipate there will be issues that need to be addressed. We will not remap the data to the current reference genome except for specific datasets that are needed for integrative analysis. Remapping is not a good idea because remapped data must be analyzed. The new DCC will conduct QA procedures to guarantee the highest quality presentation of these data.

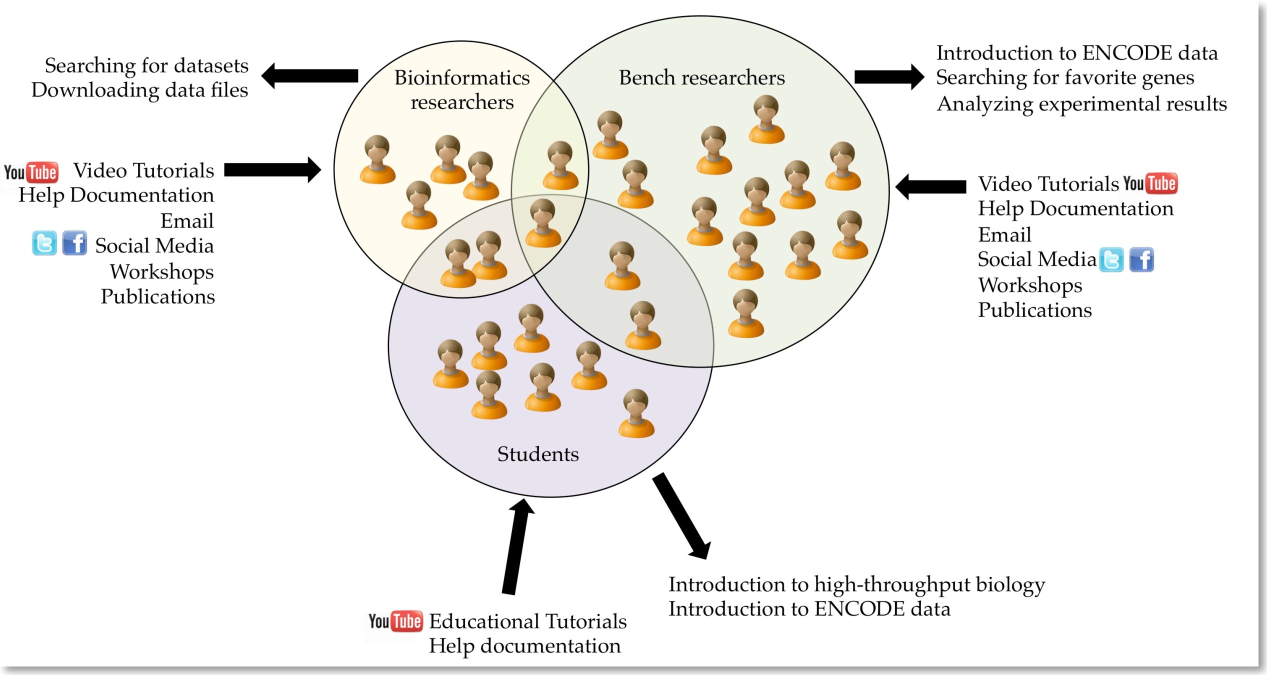
B.2.g **Aim 7: The DCC will be integrated with the DAC to form a unified ENCODE Data Coordination and Analysis Center (EDCAC).**

The DCC PI and co-Investigators will work with the DAC PI and its co-Investigators to specify and implement the most productive EDCAC possible by facilitating access to the data, ensuring reproducibility of analysis, and supporting the bioinformatics needs of the DAC. This will include integration of the ENCODE Analysis Working Group. We are confident that our technical skills and resource management experience will allow us to work with our DAC partners to build a robust and dynamic EDCAC for the betterment of all participating labs. Our software engineering and data wrangling will complement the integrative and computational analyses provided by the DAC. We anticipate the implementation of the interface between the DCC and DAC to be ongoing as new analyses, data types and data formats, and visualization technologies become available. The integrated EDCAC will involve many experienced personnel to successfully complete our challenging tasks described in Aims 1, 3, and 4 at the DCC. However, this Specific Aim is an essential element of the DCC and will have the full attention of the PI and co-Investigators.

The DCC will support the informatics needs of the AWG by providing enhanced access to the submitted data. We will facilitate the AWG's ability to query AnnoDB before these features have been integrated into the ENCODE Portal. We anticipate the DAC will be distributed at multiple sites and its member groups will use diverse algorithms in search of functional elements within the genome. Thus these sites may require access to different levels and amounts of data. We will ensure that the DAC co-Investigators will have access to all the data they require for their work. They will also provide us with use cases that are representative of the computational biologists and genomics power users. The new DCC will provide a file versioning facility that we will strongly encourage all AWG members to use. In previous analyses, the data provenance was lost because work was conducted on data files that were not versioned and thus the exact when, what, and from where of the data and the analysis conducted was not recorded. Proper versioning of intermediate steps during analysis will provide a proper description of the final results. The DCC AnnoDB and file versioning will allow each step of the analysis to be deposited and members of the AWG will explicitly know the provenance of the data. This AnnoDB driven system eliminates the need for 'soft' freezes requested by the AWG and those that were defined outside of NHGRI milestones. A list of file IDs can be shared, the specified files represent an explicit set of data and their metadata can be similarly retrieved.

The DCC will manage the software specification and implementation used by the labs for data validation. The DCC is responsible for the verification and shepherding of data through the pipeline and ensuring the quality of the metadata. We anticipate the AWG will define changes to the methods used to enhance the verification and analysis pipelines as the project progresses and will work with the AWG to apply their new specifications into the pipelines. The components of the EDCAC and AWG may share the implementation of some components of the pipeline. In such cases, the DCC and DAC PIs will determine the necessary requirements and staffing to allow an effective work relationship. To facilitate tracking and reproducibility of analytical and computational results, the DCC will specify metadata that applies to computational protocols. Members of the AWG or any other computational analysts will be encouraged to submit the results of their calculations back to the DCC, along with references to source code and other data preparation and normalization protocols such that all can benefit from their efforts, while still tracking data provenance. As described in Aim 4, the DCC will provide computational resources for a small number of AWG members as defined by the DAC. We can provide this subgroup of the AWG access to the DC4 resources at the SDSC. This is more appropriate for analyses that require fast I/O and access to lower levels of the datasets. Many other types of analysis work on Level 3 or 4 data and typically do not have similar I/O constraints. While we are not able to provide computing resources for the full AWG, they will however, have full access to the information stored within the Big Data Hub. The full datasets will be available from several geographically dispersed locations, San Diego, Chicago, Cambridge UK and the Amazon S3 cloud (Aim 4; also Figure 9).

Figure 10: Different communities who will use ENCODE data, their needs, and how the DCC will provide outreach material to meet those needs.



The DCC will be the primary contact to the data production labs for the EDCAC. We will work closely with the labs to ensure, and assist as needed, that they are meeting their requirements as specified by the EDCAC and the greater ENCODE project. The DCC is responsible for facilitating the discussion and specification of data provided to the AWG. The results from AWG analyses will be incorporated into the ENCODE file store and AnnoDB for distribution to the public. As mentioned above the DCC strongly encourages the AWG to deposit their intermediate data files to the DCC to track and share their important work. DCC PI will also interact closely with the NIH to ensure the success of the ENCODE Project.

The DCC PI is committed to working effectively with the DAC PI to resolve any issues that may arise. It is of the upmost importance that the two PIs pay attention to what is happening within their groups. Minor differences in expectations, if not addressed, can grow into major problems. It is important for the PIs to meet with his staff in small groups to obtain the perspective of the team members. Concerns are often not mentioned within a larger meeting if they are seen as trivial or as too critical. The EDCAC PIs will take actions as needed to address any challenges that diminish the effectiveness of the projects.

B.2.h **Aim 8: The DCC will maintain service to and interactions with the research community.**

The ENCODE project creates highly detailed results using state of the art experimental data. The results of this project are unique and useful to all academic and commercial investigators who are exploring human biology. If the other Specific Aims of this proposal are successful, then this Specific Aim becomes the most important. The DCC will create a variety of training materials and communicate through a variety of media outlets to educate those seeking to better understand the biology. The scientific community will be instructed in how to access the rich data and results from the ENCODE project to accelerate the pace of their research. Each of the communities represented in Figure 10 provides a unique challenge for outreach and instruction. The ENCODE project is much more than a set of genomic assays. Through the integration with other resources it provides the primary source of human gene regulation, function and how these processes change during development and within differing tissues. In the future, combining these data with sequence variation, disease phenotypes, cancer genomic abnormalities and the interactions of its gene products will produce a powerful information source.

1. Documentation. As mentioned in Aim 3, the ENCODE Portal will provide the following types of documentation: descriptions of data standards, the data pipeline, and the metadata, help guides to aid in the search, retrieval, and visualization of the ENCODE data, and progress reports regarding different aspects of the project. These forms of documentation will provide guidelines on the appropriate use and application of the ENCODE data for the casual user as well as the power user, as defined in Aim 3.

The ENCODE Portal will update and expand on the ENCODE-specific tutorials, FAQs, and user guides that are currently available for bench biologists and computational scientists at the UCSC Genome Browser website. These forms of documentation include descriptions of the data that can be found through the ENCODE Portal, short answers to commonly asked questions regarding the data and search tools, and detailed workflows to query, view, and download the data of interest.

The ENCODE Portal will also produce video tutorials to highlight components of the ENCODE Portal. These video tutorials will be made available through an ENCODE channel on YouTube, similar to the GenomeTV YouTube channel [69]. Many short tutorials (1-2 minutes) will describe the full range of browser and search features as well as introductions to the data and results coming from the ENCODE project. These shorter tutorials can often substitute for longer text-based help guides because the workflow is demonstrated visually and not described. Longer educational tutorials (8-10 minutes) will also be produced to educate the audience in greater depth regarding the techniques and experimental methods used by different labs in the ENCODE project. Students who are new to functional genomics will benefit from the longer tutorials as an introduction to how different experimental methods can be analyzed using a variety of algorithms and software. The longer tutorials provide a mechanism to introduce the ENCODE project and the technologies used by the ENCODE project while providing an introduction to the world of functional genomics.

In addition to the online help documentation and video tutorials, the DCC staff will write blog posts to provide details about different aspects of the ENCODE project. These posts could include an in-depth description about a recently released dataset, a publication that uses ENCODE data for discovery, or a tip about a unique feature of the ENCODE Portal. These blog posts can complement the more structured help documentation by providing a more detailed look into a particular dataset or a particular workflow. Using thoughtful category tags, such as 'Search tips', will allow blog posts to become an integral part of the user guide by collecting all related posts under a single category.

2. Modes of communication. To provide customized assistance to the scientific community about the ENCODE data, the DCC staff will be active on community discussion lists, such as BioStar (http://biostar.stackexchange.com/), engage the scientific community and the public via social media outlets, and answer questions via a traditional Help Desk. Due to the diversity of scientific communities who will benefit from the ENCODE data, having personalized modes of communication allows a greater understanding of how ENCODE data will be used. This knowledge can be used to improve the documentation, such as the user guides or video tutorials.

3. Meetings and workshops. The DCC staff will attend, present posters and talks, and organize workshops and tutorials at domestic and international meetings. The audience at these meetings will range from the core users of the ENCODE data (such as the AWG) to researchers investigating different aspects of human biology (such as the annual meeting of the American Society of Human Genetics) to a diverse community of researchers studying a single biological process (such as meetings focused on transcription regulation). At

these meetings, we will distribute printed materials that provide overview of the ENCODE project and present short workflows to access the data as well as organize tutorials and workshops to present these information in greater detail.

4. Publications. Publications from the DCC staff will include overviews of the data available from the ENCODE Portal and other data repositories and how these data can be accessed. In addition, publications will describe the processes required to ensure data accuracy and integrity in the Big Data Hub and AnnoDB. We will also publish software and to enhance the data wrangling tasks to allow other projects to learn from our experience.

5. Surveys. To ensure that the goals and resource planning for the DCC is consistent with the functionality required by the scientific community, we will conduct frequent user surveys to assess usability of the ENCODE portal and resources and inquire about new features. For example, the scientific community can be queried to see if there is a need and a demand for smart phone app to view and search ENCODE data. The design of the survey is important to allow both feedback on data and tools currently provided but also to be able to draw out ideas and needed from the users on what where benefit there research.

B.3 ACCESS AND DISSEMINATION

All released data will be provided freely via the Big Data Hub (Aim 2 and 4), the ENCODE Portal (Aim 3), and any of several other sites that will mirror the data (Aim 5). Appropriate restrictions will be placed on aspects of the data that is not ready for release. The data will be backed up via the mirror at the OSDC Chicago Data Center and the EBI.

B.4 TRAINING / USER SUPPORT

As described in Specific Aim 8 we will provide extensive training and user support of the ENCODE Portal. This includes the focused presentation for the different levels of use, production labs, AWG and the public.

B.5 ADMINISTRATION AND MANAGEMENT

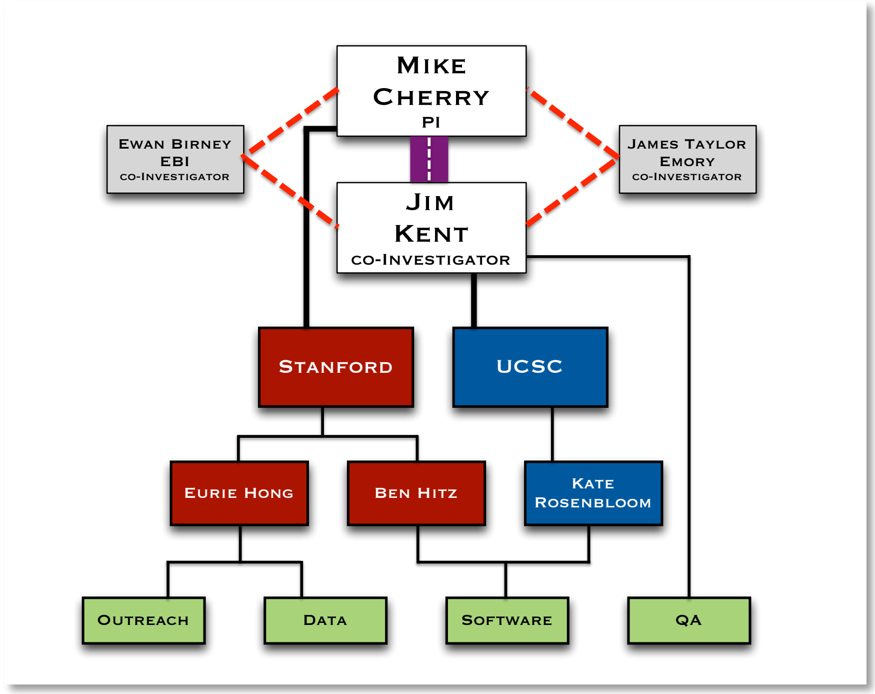
In the first year we anticipate 100 TB of data will be generated. The cost of disk space is currently approximately $1K for each 1 TB of high quality RAID-mounted disk. Instead of tape backups, will be have the data duplicated off site and these additional disks will cost the same price. In the initial year there will be ~3500 current human ENCODE datasets, ~300 mouse ENCODE datasets, ~1500 modENCODE tracks, and ~3000 Roadmap Epigenomics datasets. These ~8300 datasets will require a substantial use of staff to accurately translate into the new metadata standard to allow complete access via the enhanced Portal. We anticipate this task alone will require three Wranglers, two QA engineers and two software engineers. However, this task should only require a year to complete. The previous experience of the ENCODE and modENCODE DCC groups has found that each data track (typically a dataset) requires ~1 FTE per day from submission to final release. The current peak submission rate for ENCODE is ~200/month. With 1 FTE*day per dataset, 200 FTE days of Wrangler staff is required to effectively and efficiently handle the data. That is equal to 6.5 FTE Wranglers just for the regularly submitted data if it is equal to the current peak rate. We anticipate this will be an underestimation once the new higher throughput production centers reach capacity. Informal estimates suggest the new rate could be 2-3X more than the current submission rate. The task is slightly more difficult for each additional organism added to the mix. Human and mouse data are very similar in structure and nomenclature. Fly and worm metadata and community nomenclature are considerably different and will require added expertise for effective integration. For this estimate assuming one Data Wrangler can process 500 tracks per year (higher than possible for the first two years), the cost per track for data wrangling alone is ($150K Total cost per Wrangler / 500 tracks per Wrangler) $300 per track total cost ($200 Direct cost). Salaries will increase slightly over the course of the project, and we must assume the efficiency of our work will also improve, and the software components of the verification system will become more robust and thus, the cost per track for data wrangling should decrease by 10-15% over the course of the project. The actual cost per track will be at least 75% higher with software engineering, equipment, distribution to NCBI and EBI, and QA are factored into the equation. That amount would not include the cost of software engineering for the Portal and data distribution, Portal design, user support and outreach, project management. The cost per terabyte of data is not an interesting a number for three reasons; First the amount of data associated per assay is different; second, the amount of data per sequencing run is increasing; and third, the value of a discovery is not measured by the amount of information used to make the discovery.

B.5.a ORGANIZATIONAL STRUCTURE AND STAFF RESPONSIBILITIES

The overall project management will be the responsibility of the PI. Dr. Cherry will be responsible for the operations of all aspects of the DCC provided resources and operations. The availability of Dr. Cherry to the co-Investigators will be foremost in his priorities to ensure all project sites are functioning together well. Stanford and UCSC sites will be integrated at many levels and this will require Dr. Cherry to be paying attention to the workings of both groups. Integration will start with Drs. Cherry and Kent always being in close contact as represented by the highway between the two in Figure 11. The expertise provided to this project by Drs. Cherry and Kent is very complementary. These two senior scientists together have over 25 years experience creating and maintaining world-class bioinformatics resources. Professional interactions between these two have been occurring for many years. Dr. Cherry has visited Dr. Kent many times over the years learning of his methods in managing and motivating his engineering team working on the outstanding UCSC Genome Browser. Over this same period Dr. Cherry has shared his experiences integrating the SGD project into the entire budding yeast research community. The management of biocurators and his direction of the Gene Ontology Consortium have also provided Dr. Cherry with valuable management abilities for this DCC. The engineering design and quality assurance components will be lead by Dr. Kent. The curation of metadata and interactions with production laboratory staff, Data Wrangling, the Portal interface design and user outreach will be lead by Dr. Cherry.

Concerns are often not mentioned within a larger meeting if they are seen as trivial or if seen as out of place. Drs. Cherry and Kent will have regular face-to-face and electronic discussions to assess the progress of each task and the interactions between the DAC, and the labs. The Stanford and Santa Cruz groups will regularly make the 45-mile drive to meet at each other's site. The working relationships between the two groups have already been rewarding as we learn from our individual experiences and successes.

Figure 11: Management Structure of the ENCODE Data Coordinating Component of the EDCAC.



Drs. Cherry and Kent have an excellent project staff and they will interface together, (black lines in Figure 11). Dr. Rosenbloom is the Project Manager at UCSC. She will work jointly with Drs. Kent, Cherry and Hitz to set the specific engineering tasks that must be completed. It is critical that the UCSC and Stanford engineers/programmers act as a team. Daily management of the development team will be the responsibility of Drs. Rosenbloom and Hitz. As with any development group managers will assign tasks based on the expertise of team members. Dr. Hong will be responsible for the overall Data Wrangler management and will coordinate with Ms. Sloan, lead of the UCSC Wranglers, to maintain a unified team. During the initial months of this project Ms. Sloan will provide continuity for the current labs by maintaining the interactions and data submission pipeline that is currently in place. With development of the pipeline and hiring of Wranglers at Stanford the daily management and task prioritization for the Wranglers project will move to Stanford. Dr. Hong is very accomplished at managing a large staff of biocurators for the SGD project. The tasks facing the Data Wranglers are comparable in their requirement of technical scientific knowledge data and research methodology. The Wranglers have expert knowledge of the controlled vocabularies and the patience to work through and help the production lab's staff.

Dr. Kent is the primary manager for the Quality Assurance Engineers. Dr. Kent defines the QA tasks in coordination with Drs. Cherry and Hong to ensure a clear understanding of how issues will be resolved and the tracking system that will be used to report any identified concerns. Drs. Cherry and Hong will be the primaries for defining and executing the outreach program. Drs. Cherry and Hong have many years of experience in preparing documentation, instructional tutorials and teaching presentations and working with scientific communities to understand what they need to do better science. This experience will be supplemented with assistance from outside experts to provide the best service possible interfaces and tutorials to scientific researchers and educators.

Regular conference calls between the Stanford, UCSC, Emory and EBI staff will be a requirement of participating sites. The data wrangler team will meet weekly at Stanford. Components of the Software Engineering team will interact frequently via email and Skype, with a face-to-face meeting at Stanford or UCSC at least once a month. The Data Wranglers working from UCSC will work one day a week at Stanford to allow a tight integration of that group which is so critical for the project's success. A general staff meeting between UCSC and Stanford will occur weekly via teleconference and WebEx, with all subcomponents of the project meeting in person on a regular basis. Meeting agenda and minutes will be tracked with a combination of Redmine (or alternate calendar software) and the ENCODE wiki.

Drs. Cherry and Kent will share responsibility for setting priorities with Drs. Taylor (Emory) and Birney (EBI) subcontracts (red lines in Figure 11). As with all the management tasks the ultimate responsibility for the project is the responsibility of Dr. Cherry. The tasks to be conducted by Dr. Taylor at Emory University will be integrated into Aim 4 to provide access to the data via the Galaxy environment. The Galaxy environment will be part of Aim 8 to be presented to our users. Dr. Birney's group at the EBI will interact closely with the Stanford staff in Aim 5 for the transfer and integration of all levels of data to the EBI, as well as Aim 7 in support of the AWG. This interaction will also involve the engineering staff as part of Aim 2 for the development of the verification pipeline. Project management for this large and geographically distributed group is a challenge. The Redmine tracking system will be used as an issue and project time lining tracker. The UCSC site has used the software for some time and Stanford has used a similar system for years (bugzilla) and has now migrated to Redmine. The EBI and Emory sites will be required to use the Redmine system maintained by Stanford.

The PI will participate in all the required ENCODE telephone conferences and meetings to maintain the integration of all DCC operations and products with the production labs and the AWG. It is of particular importance that the production lab PIs are involved with defining the procedures and standards designed to accurately represent their data. The lab's staff will work directly with the Data Wranglers. Dr. Cherry will be actively engaged with the lab PIs so there is an appropriate agreement of the amount of effort required for each task.

Dr. Cherry will attend to the resolution of all issues or disagreements between the DCC and the PIs and co-Investigators of the greater EDCAC, production labs, AWG and computational projects. Dr. Cherry will pay close attention to communications between the ENCODE project members to act quickly to avoid any misunderstandings and the stress that could occur.

B.5.b SCIENTIFIC ADVISORY BOARD

The EDCAC will serve a diverse constituency including data production laboratories, bioinformatics experts and the greater biomedical research community. As part of the ENCODE project and the organizational structure defined by the NHGRI there will be appropriate and sufficient exchange between all PIs of the project including production labs, computational development groups, EDCAC and NHGRI program staff. Thus, additional advice is needed to help the DCC define our development of the Portal. We propose an advisory board including experts in online education, designers of user interfaces that provide access to complex information, professional managers of outreach sites, as well as representation by academic bioinformatics scholars from small and diverse institutions. The scholars from diverse institutions provide us with excellent advice in reaching out to small colleges and universities that have not yet connected with the ENCODE resources. This advisory board will help define best practices in creating the interface and tools that will provide the most intuitive and approachable Portal. The goal of the Portal is to open up the world of human genomics and the encyclopedic knowledgebase represented by the ENCODE Portal. An advisory of diverse professionals is often challenging to schedule. Having two meetings a year for the first year would greatly assist our effort in building the Portal it is likely that in reality we will not be able to organize two meeting for all members of the board and will thus need to accept a targeted group for one of the meetings. Hence, we suggest an advisory committee for the design and goal setting for the optimal ENCODE Portal and Community Outreach would be formed.

B.5.c MILESTONES

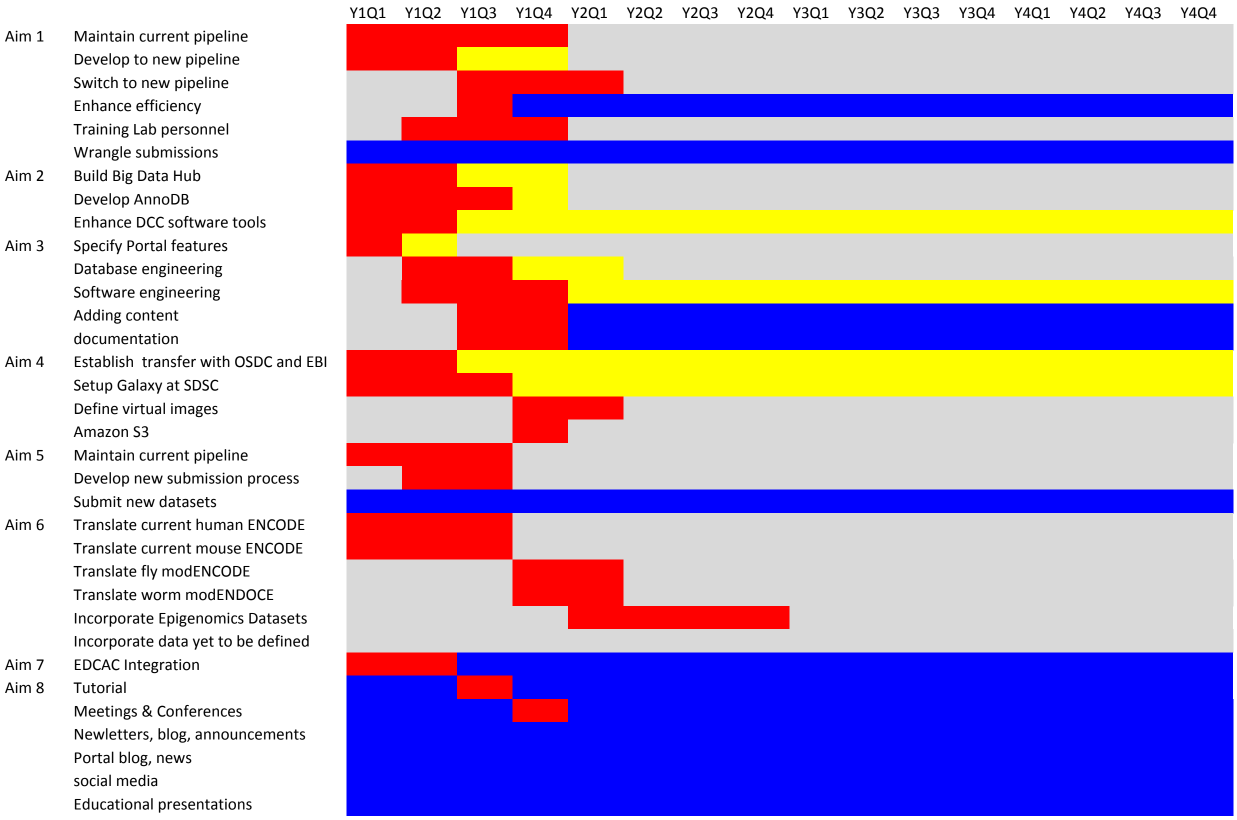


Figure 12: An estimate of the timeline Specific Aims are indicated per quarter*.* *Red indicates high priority during that quarter. Blue indicates an ongoing long-term Wrangler task and yellow indicates ongoing long-term Engineer task.*

A good estimate for the speed of completing these tasks will depend on the data submission rate, the type of assay and the organism. There is a lot of work to be done in the first year, building a new environment, loading old data and supporting the old pipeline. Initial development of the new DCC environment: submission pipelines, database, Portal and Big Data Hub will require much of the first year to design and implement while loading with the previously collected data. A phased process of migrating some parts of the pipeline will begin in Y1Q3 with training the lab personnel. The scripts for translating old ENCODE to new ENCODE metadata standard will begin in Y1Q2. Once the Big Data Hub and AnnoDB are ready the translation can begin in Y1Q3. Translation of human and mouse data will take until Y2Q2. The development of the Portal will begin Y1Q3. Components will be released as completed throughout the term of the project. Maintenance and enhancement of the pipelines, validator, verification software, interfaces to AnnoDB will continue for the life of the project. User support occurs for the life of the project. Wrangling of submitted datasets occurs for the life of the project. We hope to reach 500 datasets processed per FTE Wrangler per year. Many of the tasks will have a cyclic demand on staff time. For example data providers are known to submit the most datasets just before a data freeze this means nothing else can happen until the data is verified. The Portal search interface, Web Services, web design and documentation are ongoing tasks. Project management is a continual task. The DCC management will be dedicated to maintaining an integrated EDCAC for the course of the project.

B.6 SUMMARY OF RESEARCH APPROACH

The partnership between Stanford, UCSC, Emory and EBI will begin a team devoted to providing service to the data providers. However, the purpose of the DCC as is ENCODE is to aid biomedical discovery. Through our understanding of the experimental methods, resulting data, and integrative analyses we will contribute to the consortium in a unique manner focused on public use of the data and building a resource where ENCODE is the hub of all human genomic information.

C. Bibliography and Resources Cited

1. Lander, E.S., L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J.C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R.H. Waterston, R.K. Wilson, L.W. Hillier, J.D. McPherson, M.A. Marra, E.R. Mardis, L.A. Fulton, A.T. Chinwalla, K.H. Pepin, W.R. Gish, S.L. Chissoe, M.C. Wendl, K.D. Delehaunty, T.L. Miner, A. Delehaunty, J.B. Kramer, L.L. Cook, R.S. Fulton, D.L. Johnson, P.J. Minx, S.W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R.A. Gibbs, D.M. Muzny, S.E. Scherer, J.B. Bouck, E.J. Sodergren, K.C. Worley, C.M. Rives, J.H. Gorrell, M.L. Metzker, S.L. Naylor, R.S. Kucherlapati, D.L. Nelson, G.M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D.R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H.M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R.W. Davis, N.A. Federspiel, A.P. Abola, M.J. Proctor, R.M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D.R. Cox, M.V. Olson, R. Kaul, N. Shimizu, K. Kawasaki, S. Minoshima, G.A. Evans, M. Athanasiou, R. Schultz, B.A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W.R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J.A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D.G. Brown, C.B. Burge, L. Cerutti, H.C. Chen, D. Church, M. Clamp, R.R. Copley, T. Doerks, S.R. Eddy, E.E. Eichler, T.S. Furey, J. Galagan, J.G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L.S. Johnson, T.A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W.J. Kent, P. Kitts, E.V. Koonin, I. Korf, D. Kulp, D. Lancet, T.M. Lowe, A. McLysaght, T. Mikkelsen, J.V. Moran, N. Mulder, V.J. Pollara, C.P. Ponting, G. Schuler, J. Schultz, G. Slater, A.F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y.I. Wolf, K.H. Wolfe, S.P. Yang, R.F. Yeh, F. Collins, M.S. Guyer, J. Peterson, A. Felsenfeld, K.A. Wetterstrand, A. Patrinos, M.J. Morgan, P. de Jong, J.J. Catanese, K. Osoegawa, H. Shizuya, S. Choi and Y.J. Chen, *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

2. International Human Genome Sequencing Consortium, *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-45.

3. Harrow, J., F. Denoeud, A. Frankish, A. Reymond, C.K. Chen, J. Chrast, J. Lagarde, J.G. Gilbert, R. Storey, D. Swarbreck, C. Rossier, C. Ucla, T. Hubbard, S.E. Antonarakis, and R. Guigo, *GENCODE: producing a reference annotation for ENCODE.* Genome Biol, 2006. **7 Suppl 1**: p. S4 1-9. PMC1810553.

4. Takahashi, H., S. Kato, M. Murata, and P. Carninci, *CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks.* Methods in molecular biology, 2012. **786**: p. 181-200.

5. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11. PMC2672628.

6. Nammo, T., S.A. Rodriguez-Segui, and J. Ferrer, *Mapping open chromatin with formaldehyde-assisted isolation of regulatory elements.* Methods in molecular biology, 2011. **791**: p. 287-96.

7. Li, G., M.J. Fullwood, H. Xu, F.H. Mulawadi, S. Velkov, V. Vega, P.N. Ariyaratne, Y.B. Mohamed, H.S. Ooi, C. Tennakoon, C.L. Wei, Y. Ruan, and W.K. Sung, *ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing.* Genome biology, 2010. **11**(2): p. R22. PMC2872882.

8. van Berkum, N.L. and J. Dekker, *Determining spatial chromatin organization of large genomic regions using 5C technology.* Methods in molecular biology, 2009. **567**: p. 189-213.

9. Zhao, J., T.K. Ohsumi, J.T. Kung, Y. Ogawa, D.J. Grau, K. Sarma, J.J. Song, R.E. Kingston, M. Borowsky, and J.T. Lee, *Genome-wide identification of polycomb-associated RNAs by RIP-seq.* Molecular cell, 2010. **40**(6): p. 939-53. PMC3021903.

10. Cheng, C., K.K. Yan, K.Y. Yip, J. Rozowsky, R. Alexander, C. Shou, and M. Gerstein, *A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets.* Genome Biol, 2011. **12**(2): p. R15. PMC3188797.

11. Gerstein, M.B., Z.J. Lu, E.L. Van Nostrand, C. Cheng, B.I. Arshinoff, T. Liu, K.Y. Yip, R. Robilotto, A. Rechtsteiner, K. Ikegami, P. Alves, A. Chateigner, M. Perry, M. Morris, R.K. Auerbach, X. Feng, J. Leng, A. Vielle, W. Niu, K. Rhrissorrakrai, A. Agarwal, R.P. Alexander, G. Barber, C.M. Brdlik, J. Brennan, J.J. Brouillet, A. Carr, M.S. Cheung, H. Clawson, S. Contrino, L.O. Dannenberg, A.F. Dernburg, A. Desai, L. Dick, A.C. Dose, J. Du, T. Egelhofer, S. Ercan, G. Euskirchen, B. Ewing, E.A. Feingold, R. Gassmann, P.J. Good, P. Green, F. Gullier, M. Gutwein, M.S. Guyer, L. Habegger, T. Han, J.G. Henikoff, S.R. Henz, A. Hinrichs, H. Holster, T. Hyman, A.L. Iniguez, J. Janette, M. Jensen, M. Kato, W.J. Kent, E. Kephart, V. Khivansara, E. Khurana, J.K. Kim, P. Kolasinska-Zwierz, E.C. Lai, I. Latorre, A. Leahey, S. Lewis, P. Lloyd, L. Lochovsky, R.F. Lowdon, Y. Lubling, R. Lyne, M. MacCoss, S.D. Mackowiak, M. Mangone, S. McKay, D. Mecenas, G. Merrihew, D.M. Miller, 3rd, A. Muroyama, J.I. Murray, S.L. Ooi, H. Pham, T. Phippen, E.A. Preston, N. Rajewsky, G. Ratsch, H. Rosenbaum, J. Rozowsky, K. Rutherford, P. Ruzanov, M. Sarov, R. Sasidharan, A. Sboner, P. Scheid, E. Segal, H. Shin, C. Shou, F.J. Slack, C. Slightam, R. Smith, W.C. Spencer, E.O. Stinson, S. Taing, T. Takasaki, D. Vafeados, K. Voronina, G. Wang, N.L. Washington, C.M. Whittle, B. Wu, K.K. Yan, G. Zeller, Z. Zha, M. Zhong, X. Zhou, J. Ahringer, S. Strome, K.C. Gunsalus, G. Micklem, X.S. Liu, V. Reinke, S.K. Kim, L.W. Hillier, S. Henikoff, F. Piano, M. Snyder, L. Stein, J.D. Lieb and R.H. Waterston, *Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project.* Science, 2010. **330**(6012): p. 1775-87. PMC3142569.

12. Roy, S., J. Ernst, P.V. Kharchenko, P. Kheradpour, N. Negre, M.L. Eaton, J.M. Landolin, C.A. Bristow, L. Ma, M.F. Lin, S. Washietl, B.I. Arshinoff, F. Ay, P.E. Meyer, N. Robine, N.L. Washington, L. Di Stefano, E. Berezikov, C.D. Brown, R. Candeias, J.W. Carlson, A. Carr, I. Jungreis, D. Marbach, R. Sealfon, M.Y. Tolstorukov, S. Will, A.A. Alekseyenko, C. Artieri, B.W. Booth, A.N. Brooks, Q. Dai, C.A. Davis, M.O. Duff, X. Feng, A.A. Gorchakov, T. Gu, J.G. Henikoff, P. Kapranov, R. Li, H.K. MacAlpine, J. Malone, A. Minoda, J. Nordman, K. Okamura, M. Perry, S.K. Powell, N.C. Riddle, A. Sakai, A. Samsonova, J.E. Sandler, Y.B. Schwartz, N. Sher, R. Spokony, D. Sturgill, M. van Baren, K.H. Wan, L. Yang, C. Yu, E. Feingold, P. Good, M. Guyer, R. Lowdon, K. Ahmad, J. Andrews, B. Berger, S.E. Brenner, M.R. Brent, L. Cherbas, S.C. Elgin, T.R. Gingeras, R. Grossman, R.A. Hoskins, T.C. Kaufman, W. Kent, M.I. Kuroda, T. Orr-Weaver, N. Perrimon, V. Pirrotta, J.W. Posakony, B. Ren, S. Russell, P. Cherbas, B.R. Graveley, S. Lewis, G. Micklem, B. Oliver, P.J. Park, S.E. Celniker, S. Henikoff, G.H. Karpen, E.C. Lai, D.M. MacAlpine, L.D. Stein, K.P. White, and M. Kellis, *Identification of functional elements and regulatory circuits by Drosophila modENCODE.* Science, 2010. **330**(6012): p. 1787-97. PMC3192495.

13. Zhang, Z.D., A. Paccanaro, Y. Fu, S. Weissman, Z. Weng, J. Chang, M. Snyder, and M.B. Gerstein, *Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions.* Genome Res, 2007. **17**(6): p. 787-97. PMC1891338.

14. Meissner, A., A. Gnirke, G.W. Bell, B. Ramsahoye, E.S. Lander, and R. Jaenisch, *Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis.* Nucleic Acids Research, 2005. **33**(18): p. 5868-77. PMC1258174.

15. Brunner, A.L., D.S. Johnson, S.W. Kim, A. Valouev, T.E. Reddy, N.F. Neff, E. Anton, C. Medina, L. Nguyen, E. Chiao, C.B. Oyolu, G.P. Schroth, D.M. Absher, J.C. Baker, and R.M. Myers, *Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver.* Genome Res, 2009. **19**(6): p. 1044-56. PMC2694474.

16. Song, L. and G.E. Crawford, *DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells.* Cold Spring Harb Protoc, 2010. **2010**(2): p. pdb prot5384.

17. Pellegrini, M. and R. Ferrari, *Epigenetic Analysis: ChIP-chip and ChIP-seq.* Methods in molecular biology, 2012. **802**: p. 377-87.

18. Birney, E., J.A. Stamatoyannopoulos, A. Dutta, R. Guigo, T.R. Gingeras, E.H. Margulies, Z. Weng, M. Snyder, E.T. Dermitzakis, R.E. Thurman, M.S. Kuehn, C.M. Taylor, S. Neph, C.M. Koch, S. Asthana, A. Malhotra, I. Adzhubei, J.A. Greenbaum, R.M. Andrews, P. Flicek, P.J. Boyle, H. Cao, N.P. Carter, G.K. Clelland, S. Davis, N. Day, P. Dhami, S.C. Dillon, M.O. Dorschner, H. Fiegler, P.G. Giresi, J. Goldy, M. Hawrylycz, A. Haydock, R. Humbert, K.D. James, B.E. Johnson, E.M. Johnson, T.T. Frum, E.R. Rosenzweig, N. Karnani, K. Lee, G.C. Lefebvre, P.A. Navas, F. Neri, S.C. Parker, P.J. Sabo, R. Sandstrom, A. Shafer, D. Vetrie, M. Weaver, S. Wilcox, M. Yu, F.S. Collins, J. Dekker, J.D. Lieb, T.D. Tullius, G.E. Crawford, S. Sunyaev, W.S. Noble, I. Dunham, F. Denoeud, A. Reymond, P. Kapranov, J. Rozowsky, D. Zheng, R. Castelo, A. Frankish, J. Harrow, S. Ghosh, A. Sandelin, I.L. Hofacker, R. Baertsch, D. Keefe, S. Dike, J. Cheng, H.A. Hirsch, E.A. Sekinger, J. Lagarde, J.F. Abril, A. Shahab, C. Flamm, C. Fried, J. Hackermuller, J. Hertel, M. Lindemeyer, K. Missal, A. Tanzer, S. Washietl, J. Korbel, O. Emanuelsson, J.S. Pedersen, N. Holroyd, R. Taylor, D. Swarbreck, N. Matthews, M.C. Dickson, D.J. Thomas, M.T. Weirauch, J. Gilbert, J. Drenkow, I. Bell, X. Zhao, K.G. Srinivasan, W.K. Sung, H.S. Ooi, K.P. Chiu, S. Foissac, T. Alioto, M. Brent, L. Pachter, M.L. Tress, A. Valencia, S.W. Choo, C.Y. Choo, C. Ucla, C. Manzano, C. Wyss, E. Cheung, T.G. Clark, J.B. Brown, M. Ganesh, S. Patel, H. Tammana, J. Chrast, C.N. Henrichsen, C. Kai, J. Kawai, U. Nagalakshmi, J. Wu, Z. Lian, J. Lian, P. Newburger, X. Zhang, P. Bickel, J.S. Mattick, P. Carninci, Y. Hayashizaki, S. Weissman, T. Hubbard, R.M. Myers, J. Rogers, P.F. Stadler, T.M. Lowe, C.L. Wei, Y. Ruan, K. Struhl, M. Gerstein, S.E. Antonarakis, Y. Fu, E.D. Green, U. Karaoz, A. Siepel, J. Taylor, L.A. Liefer, K.A. Wetterstrand, P.J. Good, E.A. Feingold, M.S. Guyer, G.M. Cooper, G. Asimenos, C.N. Dewey, M. Hou, S. Nikolaev, J.I. Montoya-Burgos, A. Loytynoja, S. Whelan, F. Pardi, T. Massingham, H. Huang, N.R. Zhang, I. Holmes, J.C. Mullikin, A. Ureta-Vidal, B. Paten, M. Seringhaus, D. Church, K. Rosenbloom, W.J. Kent, E.A. Stone, S. Batzoglou, N. Goldman, R.C. Hardison, D. Haussler, W. Miller, A. Sidow, N.D. Trinklein, Z.D. Zhang, L. Barrera, R. Stuart, D.C. King, A. Ameur, S. Enroth, M.C. Bieda, J. Kim, A.A. Bhinge, N. Jiang, J. Liu, F. Yao, V.B. Vega, C.W. Lee, P. Ng, A. Yang, Z. Moqtaderi, Z. Zhu, X. Xu, S. Squazzo, M.J. Oberley, D. Inman, M.A. Singer, T.A. Richmond, K.J. Munn, A. Rada-Iglesias, O. Wallerman, J. Komorowski, J.C. Fowler, P. Couttet, A.W. Bruce, O.M. Dovey, P.D. Ellis, C.F. Langford, D.A. Nix, G. Euskirchen, S. Hartman, A.E. Urban, P. Kraus, S. Van Calcar, N. Heintzman, T.H. Kim, K. Wang, C. Qu, G. Hon, R. Luna, C.K. Glass, M.G. Rosenfeld, S.F. Aldred, S.J. Cooper, A. Halees, J.M. Lin, H.P. Shulha, M. Xu, J.N. Haidar, Y. Yu, V.R. Iyer, R.D. Green, C. Wadelius, P.J. Farnham, B. Ren, R.A. Harte, A.S. Hinrichs, H. Trumbower, H. Clawson, J. Hillman-Jackson, A.S. Zweig, K. Smith, A. Thakkapallayil, G. Barber, R.M. Kuhn, D. Karolchik, L. Armengol, C.P. Bird, P.I. de Bakker, A.D. Kern, N. Lopez-Bigas, J.D. Martin, B.E. Stranger, A. Woodroffe, E. Davydov, A. Dimas, E. Eyras, I.B. Hallgrimsdottir, J. Huppert, M.C. Zody, G.R. Abecasis, X. Estivill, G.G. Bouffard, X. Guan, N.F. Hansen, J.R. Idol, V.V. Maduro, B. Maskeri, J.C. McDowell, M. Park, P.J. Thomas, A.C. Young, R.W. Blakesley, D.M. Muzny, E. Sodergren, D.A. Wheeler, K.C. Worley, H. Jiang, G.M. Weinstock, R.A. Gibbs, T. Graves, R. Fulton, E.R. Mardis, R.K. Wilson, M. Clamp, J. Cuff, S. Gnerre, D.B. Jaffe, J.L. Chang, K. Lindblad-Toh, E.S. Lander, M. Koriabine, M. Nefedov, K. Osoegawa, Y. Yoshinaga, B. Zhu and P.J. de Jong, *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816. PMC2212820.

19. ENCODE Project Consortium, *The ENCODE (ENCyclopedia Of DNA Elements) Project.* Science, 2004. **306**(5696): p. 636-40.

20. Rosenbloom, K.R., T.R. Dreszer, J.C. Long, V.S. Malladi, C.A. Sloan, B.J. Raney, M.S. Cline, D. Karolchik, G.P. Barber, H. Clawson, M. Diekhans, P.A. Fujita, M. Goldman, R.C. Gravell, R.A. Harte, A.S. Hinrichs, V.M. Kirkup, R.M. Kuhn, K. Learned, M. Maddren, L.R. Meyer, A. Pohl, B. Rhead, M.C. Wong, A.S. Zweig, D. Haussler, and W.J. Kent, *ENCODE whole-genome data in the UCSC Genome Browser: update 2012.* Nucleic Acids Research, 2011. PMC Journal - In Process.

21. Muers, M., *Functional genomics: the modENCODE guide to the genome.* Nat Rev Genet, 2011. **12**(2): p. 80.

22. Dreszer, T.R., D. Karolchik, A.S. Zweig, A.S. Hinrichs, B.J. Raney, R.M. Kuhn, L.R. Meyer, M. Wong, C.A. Sloan, K.R. Rosenbloom, G. Roe, B. Rhead, A. Pohl, V.S. Malladi, C.H. Li, K. Learned, V. Kirkup, F. Hsu, R.A. Harte, L. Guruvadoo, M. Goldman, B.M. Giardine, P.A. Fujita, M. Diekhans, M.S. Cline, H. Clawson, G.P. Barber, D. Haussler, and W. James Kent, *The UCSC Genome Browser database: extensions and updates 2011.* Nucleic Acids Research, 2011. PMC Journal - In Process.

23. ENCODE data at GEO: http://www.ncbi.nlm.nih.gov/geo/info/ENCODE.html

24. Rosenbloom, K.R., T.R. Dreszer, M. Pheasant, G.P. Barber, L.R. Meyer, A. Pohl, B.J. Raney, T. Wang, A.S. Hinrichs, A.S. Zweig, P.A. Fujita, K. Learned, B. Rhead, K.E. Smith, R.M. Kuhn, D. Karolchik, D. Haussler, and W.J. Kent, *ENCODE whole-genome data in the UCSC Genome Browser.* Nucleic Acids Research, 2010. **38**(Database issue): p. D620-5. PMC2808953.

25. Raney, B.J., M.S. Cline, K.R. Rosenbloom, T.R. Dreszer, K. Learned, G.P. Barber, L.R. Meyer, C.A. Sloan, V.S. Malladi, K.M. Roskin, B.B. Suh, A.S. Hinrichs, H. Clawson, A.S. Zweig, V. Kirkup, P.A. Fujita, B. Rhead, K.E. Smith, A. Pohl, R.M. Kuhn, D. Karolchik, D. Haussler, and W.J. Kent, *ENCODE whole-genome data in the UCSC genome browser (2011 update).* Nucleic Acids Research, 2011. **39**(Database issue): p. D871-5. PMC3013645.

26. Barrett, T., D.B. Troup, S.E. Wilhite, P. Ledoux, C. Evangelista, I.F. Kim, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, R.N. Muertter, M. Holko, O. Ayanbule, A. Yefanov, and A. Soboleva, *NCBI GEO: archive for functional genomics data sets--10 years on.* Nucleic Acids Research, 2011. **39**(Database issue): p. D1005-10. PMC3013736.

27. McQuilton, P., S.E. St Pierre, and J. Thurmond, *FlyBase 101 - the basics of navigating FlyBase.* Nucleic Acids Research, 2011. PMC Journal - In Process.

28. Yook, K., T.W. Harris, T. Bieri, A. Cabunoc, J. Chan, W.J. Chen, P. Davis, N. de la Cruz, A. Duong, R. Fang, U. Ganesan, C. Grove, K. Howe, S. Kadam, R. Kishore, R. Lee, Y. Li, H.M. Muller, C. Nakamura, B. Nash, P. Ozersky, M. Paulini, D. Raciti, A. Rangarajan, G. Schindelman, X. Shi, E.M. Schwarz, M. Ann Tuli, K. Van Auken, D. Wang, X. Wang, G. Williams, J. Hodgkin, M. Berriman, R. Durbin, P. Kersey, J. Spieth, L. Stein, and P.W. Sternberg, *WormBase 2012: more genomes, more data, new website.* Nucleic Acids Research, 2011. PMC Journal - In Process.

29. GBrowse for worm modENCODE data: http://modencode.oicr.on.ca/fgb2/gbrowse/worm/

30. GBrowse for fly modENCODE data: http://modencode.oicr.on.ca/fgb2/gbrowse/fly/

31. Donlin, M.J., *Using the Generic Genome Browser (GBrowse).* Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.], 2009. **Chapter 9**: p. Unit 9 9.

32. Contrino, S., R.N. Smith, D. Butano, A. Carr, F. Hu, R. Lyne, K. Rutherford, A. Kalderimis, J. Sullivan, S. Carbon, E.T. Kephart, P. Lloyd, E.O. Stinson, N.L. Washington, M.D. Perry, P. Ruzanov, Z. Zha, S.E. Lewis, L.D. Stein, and G. Micklem, *modMine: flexible access to modENCODE data.* Nucleic Acids Research, 2011. PMC Journal - In Process.

33. modMine: http://intermine.modencode.org/

34. QuEST: http://mendel.stanford.edu/sidowlab/downloads/quest/

35. Langmead, B., C. Trapnell, M. Pop, and S.L. Salzberg, *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome biology, 2009. **10**(3): p. R25. PMC2690996.

36. Langmead, B., *Aligning short sequencing reads with Bowtie.* Curr Protoc Bioinformatics, 2010. **Chapter 11**: p. Unit 11 7. PMC3010897.

37. The Gene Ontology Consortium, *The Gene Ontology: enhancements for 2011.* Nucleic Acids Research, 2011. PMC Journal - In Process.

38. Eilbeck, K., S.E. Lewis, C.J. Mungall, M. Yandell, L. Stein, R. Durbin, and M. Ashburner, *The Sequence Ontology: a tool for the unification of genome annotations.* Genome Biol, 2005. **6**(5): p. R44. PMC1175956.

39. Brinkman, R.R., M. Courtot, D. Derom, J.M. Fostel, Y. He, P. Lord, J. Malone, H. Parkinson, B. Peters, P. Rocca-Serra, A. Ruttenberg, S.A. Sansone, L.N. Soldatova, C.J. Stoeckert, Jr., J.A. Turner, and J. Zheng, *Modeling biomedical experimental processes with OBI.* J Biomed Semantics, 2010. **1 Suppl 1**: p. S7. PMC2903726.

40. Evidence Code Ontology: http://www.obofoundry.org/cgi-bin/detail.cgi?id=evidence_code

41. de Matos, P., N. Adams, J. Hastings, P. Moreno, and C. Steinbeck, *A Database for Chemical Proteomics: ChEBI.* Methods Mol Biol, 2012. **803**: p. 273-96.

42. Meehan, T.F., A.M. Masci, A. Abdulla, L.G. Cowell, J.A. Blake, C.J. Mungall, and A.D. Diehl, *Logical development of the cell ontology.* BMC Bioinformatics, 2011. **12**: p. 6. PMC3024222.

43. Gremse, M., A. Chang, I. Schomburg, A. Grote, M. Scheer, C. Ebeling, and D. Schomburg, *The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources.* Nucleic Acids Research, 2011. **39**(Database issue): p. D507-13. PMC3013802.

44. Whetzel, P.L., H. Parkinson, H.C. Causton, L. Fan, J. Fostel, G. Fragoso, L. Game, M. Heiskanen, N. Morrison, P. Rocca-Serra, S.A. Sansone, C. Taylor, J. White, and C.J. Stoeckert, Jr., *The MGED Ontology: a resource for semantics-based description of microarray experiments.* Bioinformatics, 2006. **22**(7): p. 866-73.

45. Mungall, C.J. and D.B. Emmert, *A Chado case study: an ontology-based modular schema for representing genome-associated biological information.* Bioinformatics, 2007. **23**(13): p. i337-46.

46. Zhou, P., D. Emmert, and P. Zhang, *Using Chado to store genome annotation data.* Current protocols in bioinformatics, 2006. **Chapter 9**: p. Unit 9.6.

47. Hu, Z., J.H. Hung, Y. Wang, Y.C. Chang, C.L. Huang, M. Huyck, and C. DeLisi, *VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology.* Nucleic Acids Research, 2009. **37**(Web Server issue): p. W115-21. PMC2703932.

48. Seal, R.L., S.M. Gordon, M.J. Lush, M.W. Wright, and E.A. Bruford, *genenames.org: the HGNC resources in 2011.* Nucleic Acids Research, 2011. **39**(Database issue): p. D514-9. PMC3013772.

49. Dimmer, E.C., R.P. Huntley, Y. Alam-Faruque, T. Sawford, C. O'Donovan, M.J. Martin, B. Bely, P. Browne, W. Mun Chan, R. Eberhardt, M. Gardner, K. Laiho, D. Legge, M. Magrane, K. Pichler, D. Poggioli, H. Sehra, A. Auchincloss, K. Axelsen, M.C. Blatter, E. Boutet, S. Braconi-Quintaje, L. Breuza, A. Bridge, E. Coudert, A. Estreicher, L. Famiglietti, S. Ferro-Rojas, M. Feuermann, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, J. James, S. Jimenez, F. Jungo, G. Keller, P. Lemercier, D. Lieberherr, P. Masson, M. Moinat, I. Pedruzzi, S. Poux, C. Rivoire, B. Roechert, M. Schneider, A. Stutz, S. Sundaram, M. Tognolli, L. Bougueleret, G. Argoud-Puy, I. Cusin, P. Duek-Roggli, I. Xenarios, and R. Apweiler, *The UniProt-GO Annotation database in 2011.* Nucleic Acids Research, 2011. PMC Journal - In Process.

50. Magrane, M. and U. Consortium, *UniProt Knowledgebase: a hub of integrated protein data.* Database (Oxford), 2011. **2011**: p. bar009. PMC3070428.

51. Flicek, P., M.R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A.K. Kahari, D. Keefe, S. Keenan, R. Kinsella, M. Komorowska, G. Koscielny, E. Kulesha, P. Larsson, I. Longden, W. McLaren, M. Muffato, B. Overduin, M. Pignatelli, B. Pritchard, H.S. Riat, G.R. Ritchie, M. Ruffier, M. Schuster, D. Sobral, Y.A. Tang, K. Taylor, S. Trevanion, J. Vandrovcova, S. White, M. Wilson, S.P. Wilder, B.L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X.M. Fernandez-Suarez, J. Harrow, J. Herrero, T.J. Hubbard, A. Parker, G. Proctor, G. Spudich, J. Vogel, A. Yates, A. Zadissa, and S.M. Searle, *Ensembl 2012.* Nucleic Acids Research, 2011. PMC Journal - In Process.

52. Sayers, E.W., T. Barrett, D.A. Benson, E. Bolton, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. Dicuccio, S. Federhen, M. Feolo, I.M. Fingerman, L.Y. Geer, W. Helmberg, Y. Kapustin, S. Krasnov, D. Landsman, D.J. Lipman, Z. Lu, T.L. Madden, T. Madej, D.R. Maglott, A. Marchler-Bauer, V. Miller, I. Karsch-Mizrachi, J. Ostell, A. Panchenko, L. Phan, K.D. Pruitt, G.D. Schuler, E. Sequeira, S.T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T.A. Tatusova, L. Wagner, Y. Wang, W.J. Wilbur, E. Yaschenko, and J. Ye, *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Research, 2011. PMC Journal - In Process.

53. Ensembl Regulation: http://www.ensembl.org/info/docs/funcgen/

54. Jones, T., A. Koniges, and R.K. Yates *Performance of the IBM general parallel file system*. International Parallel and Distributed Processing Symposium, 2000. 673-68 DOI: doi: 10.1109/IPDPS.2000.846052.

55. Goecks, J., A. Nekrutenko, and J. Taylor, *Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.* Genome biology, 2010. **11**(8): p. R86. PMC2945788.

56. Afgan, E., D. Baker, N. Coraor, B. Chapman, A. Nekrutenko, and J. Taylor, *Galaxy CloudMan: delivering cloud compute clusters.* BMC Bioinformatics, 2010. **11 Suppl 12**: p. S4. PMC3040530.

57. Afgan, E., D. Baker, N. Coraor, H. Goto, I.M. Paul, K.D. Makova, A. Nekrutenko, and J. Taylor, *Harnessing cloud computing with Galaxy Cloud.* Nature biotechnology, 2011. **29**(11): p. 972-4.

58. Kodama, Y., M. Shumway, and R. Leinonen, *The sequence read archive: explosive growth of sequencing data.* Nucleic Acids Research, 2011. PMC Journal - In Process.

59. Parkinson, H., U. Sarkans, N. Kolesnikov, N. Abeygunawardena, T. Burdett, M. Dylag, I. Emam, A. Farne, E. Hastings, E. Holloway, N. Kurbatova, M. Lukk, J. Malone, R. Mani, E. Pilicheva, G. Rustici, A. Sharma, E. Williams, T. Adamusiak, M. Brandizi, N. Sklyar, and A. Brazma, *ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments.* Nucleic Acids Research, 2011. **39**(Database issue): p. D1002-4. PMC3013660.

60. SOFT2 Documentation: http://www.ncbi.nlm.nih.gov/geo/info/soft2.html

61. Li, Q., J.B. Brown, H. Huang, and P.J. Bickel, *Measuring reproducibility of high-throughput experiments.* Annals of Applied Statistics, 2011. **5**: p. 1752—1779.

62. Gostev, M., A. Faulconbridge, M. Brandizi, J. Fernandez-Banet, U. Sarkans, A. Brazma, and H. Parkinson, *The BioSample Database (BioSD) at the European Bioinformatics Institute.* Nucleic Acids Research, 2011. PMC Journal - In Process.

63. Malone, J., E. Holloway, T. Adamusiak, M. Kapushesky, J. Zheng, N. Kolesnikov, A. Zhukova, A. Brazma, and H. Parkinson, *Modeling sample variables with an Experimental Factor Ontology.* Bioinformatics, 2010. **26**(8): p. 1112-8. PMC2853691.

64. Eppig, J.T., J.A. Blake, C.J. Bult, J.A. Kadin, and J.E. Richardson, *The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse.* Nucleic Acids Research, 2011. PMC Journal - In Process.

65. Harris, T.W., I. Antoshechkin, T. Bieri, D. Blasiar, J. Chan, W.J. Chen, N. De La Cruz, P. Davis, M. Duesbury, R. Fang, J. Fernandes, M. Han, R. Kishore, R. Lee, H.M. Muller, C. Nakamura, P. Ozersky, A. Petcherski, A. Rangarajan, A. Rogers, G. Schindelman, E.M. Schwarz, M.A. Tuli, K. Van Auken, D. Wang, X. Wang, G. Williams, K. Yook, R. Durbin, L.D. Stein, J. Spieth, and P.W. Sternberg, *WormBase: a comprehensive resource for nematode research.* Nucleic Acids Research, 2010. **38**(Database issue): p. D463-7. PMC2808986.

66. Washington, N.L., E.O. Stinson, M.D. Perry, P. Ruzanov, S. Contrino, R. Smith, Z. Zha, R. Lyne, A. Carr, P. Lloyd, E. Kephart, S.J. McKay, G. Micklem, L.D. Stein, and S.E. Lewis, *The modENCODE Data Coordination Center: lessons in harvesting comprehensive experimental details.* Database (Oxford), 2011. **2011**: p. bar023. PMC3170170.

67. Bernstein, B.E., J.A. Stamatoyannopoulos, J.F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M.A. Marra, A.L. Beaudet, J.R. Ecker, P.J. Farnham, M. Hirst, E.S. Lander, T.S. Mikkelsen, and J.A. Thomson, *The NIH Roadmap Epigenomics Mapping Consortium.* Nat Biotechnol, 2010. **28**(10): p. 1045-8.

68. Epigenomics metadata proposal: http://www.ihec-epigenomes.org/DownloadDocs/NIH-Roadmap-Metadata-sent-10-23-2011.pdf

69. GenomeTV YouTube channel: http://www.youtube.com/user/GenomeTV

6. PROTECTION OF HUMAN SUBJECTS : Human subjects are not a component of this proposed work.

10. VERTEBRATE ANIMALS : Animal research is not part of this work.

13. CONSORTIUM/CONTRACTUAL ARRANGEMENTS

We have created a unique team from four institutions known to be unique for the work they will bring to this project. Each of the groups participating in this project brings a unique perspective to the development, maintenance and distribution of the ENCODE products, the collection of unique and essential data sets for the exploration of human biology. This project created by the four investigators links the largest genome browser and informatics resources, with the major analysis pipeline tool Galaxy, and the complete service solutions provided for a model for database resources.

Each of the co-Investigators is the leader of the work to be done. Dr. Kent at UCSC maintains the Genome Browser that is accessed more than 2,000,000,000 times each year. Dr. Kent has maintained the DCC for the ENCODE project for the past five years. There is no one else that can bring the resources, knowledge and creativity necessary for the presentation of the ENCODE data but Dr. Kent. Dr. Birney at EBI directs the Ensembl project with its genome browser and is connected to essential informatics resources such as the UniProt, ArrayExpress, InterPro, European Nucleotide Archive (ENA) and a large number of other important databases. Post-doctoral scholars in Dr. Birney's group will be defining the ENCODE integration of ENCODE into the EBI databases. The EBI will also mirror all of the ENCODE data at the EBI providing an alternative site for access to and backup of this priceless data. Dr. Taylor an original member of the Galaxy Project's development team is an expert with distributed cloud computing and will bring these services and the Galaxy software environment to the ENCODE Portal. Dr. Cherry is a leader in the development and maintenance of community databases. His expertise focuses on how to integrate complex biological data into a resource and then provide tools that allow scientist to access the data via an approachable interface.

The primary sites UCSC and Stanford will work collaboratively on the core components of the DCC. This collaboration is possible because of their close proximity, just an hour commute between UCSC and Stanford campuses. Specifically these groups will collaborate in the specification, development and implementation of software into pipelines and automatic tools for the DCC's operations, and providing access to the outside world. The UCSC has enhanced the Quality Assurance (QA) processes for genomic data and they will continue focusing this important responsibility, checking that all software developed either at Stanford or UCSC for this project works substantially according to specifications. The data submission and metadata curation processes, called here Data Wrangling, has been successfully provided from 2006-2011 by the UCSC team. The Stanford team has advanced skills in the areas of data curation and user support and will thus enhance the wrangling services. The Stanford team is very eager to work with the data production labs and to enhance the smooth operation of the project.

14. LETTERS OF SUPPORT

Co-Investigators of this Proposal

Ewan Birney, Ph.D.
Head of DNA data, EBI
Senior Scientist, EMBL
European Bioinformatics Institute
Hinxton, Cambridge, UK

James Taylor, Ph.D.
Assistant Professor
Department of Biology
Department of Mathematics and Computer Science
Emory University
Atlanta, Georgia

modENCODE PI and co-Investigators

Lincoln Stein, M.D./Ph.D.
Director, Informatics & Biocomputing
Senior Principal Investigator
Ontario Institute for Cancer Research
Toronto, Canada

Suzanna Lewis
Staff Scientist
Lawrence Berkeley National Laboratory
Berkeley Bioinformatics Open Sources Projects
Berkeley, California

Gos Micklem, Ph.D.
Director, Cambridge Computational Biology Institute
Department of Genetics
University of Cambridge
Cambridge, UK

PI of Mouse Genome Database, *Mus* Community Resource

Janan Eppig, Ph.D.
Professor
Mouse Genome Database
The Jackson Laboratory
Bar Harbor, Maine

PI of WormBase, *Caenorhabditis* Community Resource

Paul Sternberg, Ph.D.
Investigator
Thomas Hunt Morgan Professor of Biology
Division of Biology
California Institute of Technology
Pasadena, California

PI of FlyBase, *Drosophila* Community Resource

Bill Gelbart, Ph.D.
Professor
Department of Molecular and Cellular Biology
Harvard University
Cambridge, Massachusetts

Roadmap Epigenomics Data Analysis and Coordination Center (EDACC)

Aleksandar Milosavljevic, Ph.D.
Associate Professor
Department of Molecular and Human Genetics
Baylor School of Medicine
Houston, Texas

Roadmap Epigenomics Visualization Hub

Ting Wang, Ph.D.
Assistant Professor
Department of Genetics
Center for Genome Sciences
Washington University School of Medicine
St. Louis, Missouri

Current ENCODE Data Producer

Greg Crawford, Ph.D.
Assistant Professor
Institute for Genome Sciences & Policy
Department of Pediatrics
Duke University
Durham, North Carolina

Current ENCODE Data Producer Informatics Lead

Richard Sandstrom
Sr. Computational Scientist
Lead, Data Management and Dissemination
UW ENCODE Project/Northwestern Reference Epigenome Mapping Center
Department of Genome Sciences
University of Washington
Seattle, Washington

Cloud Computing Developer

Robert Grossman, Ph.D.
Core Faculty and Director of Informatics
Institute for Genomics and Systems Biology
University of Chicago
Chicago, Illinois

Laboratoire Européen de Biologie Moléculaire
European Molecular Biology Laboratory
Europäisches Laboratorium für Molekularbiologie

EMBL- EBI

EMBL Outstation Hinxton — The European Bioinformatics Institute

16 December 2011

Dear Mike and Jim,

This is a letter expanding on my Bio sketch and CV indicating why I am excited to be included in this DCC grant and how I will manage my increasingly complex role if we are successful in this grant.

Firstly I must stress how important the Data Coordination Center is to project such as ENCODE. Much of the work a Data Coordination Center is conceptually mundane – the business of tracking submissions, tracking meta-data and associating the meta data, data and tracking processes seem easy enough when written on paper. However, every single error in this process propagates minimally as a complex downstream error and worse still in some cases can lead to apparently valid but fundamentally incorrect results. The sheer data flow complexity coupled with the scale of the data – both in raw disk storage but more importantly in the dimensionality of cell type, samples and assay types means that there are many opportunities for well meaning but erroneous data entry by submitting groups. I think the combination of Mike's group – with a long experience of high dimensionality data in Yeast, and a long experience of complex meta data based management and Jim's group – with a proven track record in data flow at this scale and browsing is a great team, and I am very excited to be a proposed component of this group.

Secondly I would like to outline what my group at the EBI can provide. There are three main components – 1. access and smooth information flow to the EBI BioSampleDatabase (BioSD) group . This provides a worldwide coordination of the publically known samples. This database has been in prototype mode for about 1 year and is now taking on an increasingly central role at the EBI; it is a peer of the NCBI

BioSample database group and has a robust exchange mechanism set up. By coordinating up front the ENCODE meta-data information with this globally synchronised meta-data system (common across EBI and NCBI) we will both place the cell type and tissue in the context of existing ontologies and future proof the integration of ENCODE data with other datasets. 2. The provision of a "output" view of the data via the Ensembl web site. The UCSC genome browser is an excellent and exemplar browser system, particularly focused on serving "genomics users" and users in the US. Ensembl provides a series of complementary views with a stronger gene centric view point and a different geographic emphasis in Europe. To maximise the utility of the ENCODE data is worth seeing the information flow onto as many access platforms as possible, and the combination of the UCSC genome browser and the Ensembl system provides two of the largest portals of genomic information. 3 The development of routine QC and low level analysis routines. In the previous ENCODE scale up project there was a stop-start freeze based scheme for major analysis. I think it is clear that this is not optimal in particular in capturing quality events and errors – a number of datasets were only detected as erroneous late in the system. Instead a more proactive analysis mainly as excellent QC mechanisms should run continuously. Obviously in the ENCODE context it is likely that there will be a productive interplay – as in previous projects – with the Data Analysis Centre (DAC), but low level analysis, in particular as it informs the QC of datasets is better run inside the DCC.

Thirdly I'd like to outline my personal role. As you know, in April 2012 I will be undertaking a broader, more strategic role at the EBI. In many ways over 2011 I have been fulfilling this role and it was clear in 2011 that if I was to become Associate Director this was incompatible with me playing a leading analysis role as I did in the ENCODE Scale up project. That said I am confident that I can provide the leadership to contribute to the EBI portion of this. Firstly this role is more engineering based than innovative analysis based – no less hard to execute, but easier to describe and lock down in an engineering fashion. Secondly my experience in ENCODE allows me to precisely assess the requirements of this so I am not overstretching my own role. Finally in the EMBL system I will be able to hire a "Staff Scientist" – a senior postdoc – independent of this money who will help support my on going research interests. Experience is that this allows for a better splitting of my time with strategic roles and research roles. I will make a commitment to attend in person at least once a year at a meeting of your choice, and I am likely to attend more than that.

Many thanks for bringing me into this project and I look forward to working with you.

Ewan Birney

Head of DNA data, EBI
Senior Scientist, EMBL

EMORY
COLLEGE
OF ARTS AND
SCIENCES

Faculty of Arts and Sciences
Department of Biology
Department of Mathematics and Computer Science

James Taylor
Assistant Professor

December 13, 2011

J. Michael Cherry
Department of Genetics
Stanford University
Stanford, California 94305-5120

Dear Mike,

I'm writing to express my enthusiasm for the opportunity to work with you on the next phase of the ENCODE Data Coordination Center, particularly on the development of a data architecture and additional infrastructure to make future ENCODE data more easily used by a wider community.

As you know, I have been working for a number of years on making data intensive analysis more accessible to biomedical researchers, primarily through our continuing development of the Galaxy platform. In that work, I have collaborated extensively with Jim Kent and the browser team at UCSC on tight integration of between our data analysis framework, and their data storage, query, and visualization tools. Together these resources provide a complete solution for data mining, analysis, and visualization.

I'm excited that you will be taking the lead on the next phase of the ENCODE DCC. With the increasing depth and complexity of data the DCC will be managing, I think your expertise in data curation and ontologies is exactly what is needed to move the DCC forward, enabling researchers to better discover and understand ENCODE datasets. The ENCODE project is an amazing resource which I believe is not yet as widely used at it could be.

In addition to data discovery, enabling scientists to analyze ENCODE data is crucial to maximizing its utility. I will work with you to better enable data analysis in two ways. For researchers comfortable working with analysis tools directly, we will make the ENCODE DCC data available in infrastructure cloud environments, allowing efficient analysis of the data while providing a completely flexible and scalable computing environment. In addition, we will provide a Galaxy analysis instance and pre-configured virtual machines, allowing researchers without informatics expertise to work directly with ENCODE DCC data. I believe this combination of strategies is the best approach to make this data usable to the entire biomedical research community.

Looking forward to working with you,

Sincerely,

James Taylor

Ontario Institute for Cancer Research

science → discoveries → solutions

Suite 800, MaRS South Tower
101 College Street
Toronto, ON M5G 0A3
CANADA

December 11, 2011

Dr. J. Michael Cherry
Department of Genetics
Stanford University
Stanford, CA 94305

Dear Mike,

I am delighted that you and Jim Kent have joined forces on a proposal to manage the data coordinating center for the ENCODE project. As you know, I am the lead principal investigator for the modENCODE DCC; my group has had primary responsibility for the modENCODE genome browser and the modENCODE Amazon cloud images.

As expressed also in the letters from my co-PIs Suzi and Gos, we will be happy to work with you to transfer to the new DCC the modENCODE data, metadata and any of the infrastructural tools we have developed, including controlled vocabularies, file formats, and databases. My group will work closely with your software engineers and data curators in order to make the transition happen as smoothly and quickly as possible. As you know, we have requested an extension of partial funding for the modENCODE DCC for the 2012-2013 year for the purpose of ensuring an orderly transition of data into the new DCC as well as the model organism databases. With luck this request will be granted in whole or in part, so that we can work out the details of the transition without urgent time pressures, but in either event, I will be happy to discuss logistical details before your award decision is rendered.

I wish you the very best of luck with the proposal.

Lincoln D. Stein, MD/PhD
Director, Informatics & Biocomputing and
Senior Principal Investigator, OICR

Professor of Bioinformatics, Cold Spring Harbor Laboratory

Lawrence Berkeley National Laboratory
Berkeley Bioinformatics Open Source Projects
1 Cyclotron Road Mail Stop—64R0121
Berkeley, CA 94720

Prof. J. Michael Cherry
Department of Genetics
Stanford University
Stanford
California 94305-5120

December 11, 2011

Dear Mike,

I was very pleased to learn that you and Jim will be submitting an application to take over the running of the modENCODE Data Coordination Centre as part of the new ENCODE DCC. As you know my group has been responsible for modENCODE data collection and quality checking (aka data wrangling). We designed and developed the structured Wiki that data producers use to describe their experiments and protocols before submitting data to the DCC. We also developed the pipeline that runs a configurable set of checks on the data, customizes the browser interface, and loads the data into a Chado database for delivery of the integrated, published data to modMine in Cambridge.

We will be glad to work with you transfer our experience and software. This will be a natural extension of our existing collaborations and we are more than happy to meet with the new data wranglers. We have been collaborators for over a decade and have known one another for close to 20 years. One factor that has made our productive joint work possible is how close Stanford and Berkeley are, which will allow you and your staff to visit LBNL as needed.

Yours sincerely,

Suzanna Lewis

Suzanna Lewis
Staff Scientist

GATCGT Cambridge
TCCBIA Computational
TTACCG Biology
AGCCTA Institute
GAACGT

UNIVERSITY OF
CAMBRIDGE
Department of Genetics

Prof. J. Mike Cherry
Department of Genetics
Stanford University
Stanford
California 94305-5120

11th December 2011

Dear Mike,

I'm glad to hear that you will be submitting an application to take over the running of the modENCODE Data Coordination Centre as part of the new ENCODE DCC. As you know my group has been responsible for the modMine data warehouse, built using InterMine, which has acted as the primary integrated data resource for the project. We will be happy to transfer the running of modMine to you along with the associated software, data and our experience in building and maintaining it. This will be a natural extension of our existing collaboration, which supports the use of InterMine by SGD as well as almost all the other major model organism databases.

As you know we have funding to develop InterMine-based databases for at least the next three years and so we will be in a good position to support you if further fruit fly or nematode worm modENCODE projects are funded and modMine needs to be updated, or if you decide to extend the use of InterMine to further projects.

Best wishes,

Dr. Gos Micklem
Director, Cambridge Computational Biology Institute

The Jackson Laboratory
Leading the search for tomorrow's cures

J. Michael Cherry, PhD
Stanford University
Center for Genomics and Personalized Medicine
Department of Genetics
1501 S. California Ave, Rm 2419
Palo Alto, California 94304-5577

Jim Kent, PhD
University of California, Santa Cruz
Baskin School of Engineering
MS: CBSE-ITI
1156 High Street
Santa Cruz, CA 95064

5 December 2011

Dear Mike and Jim,

I was delighted to hear that you are applying for the new ENCODE Data Coordination Component (DCC) (RFA-HG-11-026) and I am very supportive of your application. Together you bring an outstanding set of skills and experience directly relevant to the execution of this program.

As the PI for the Mouse Genome Database (MGD), you contacted me to explore how MGD might incorporate some of the relevant ENCODE data in the future. The current ENCODE project includes a very small proportion of mouse datasets and we have not approached these data yet (even now, most mouse datasets are not yet released).

The new RFA for the next phase of ENCODE is anticipated to produce mostly human datasets, but with a significant portion of mouse datasets (depending, of course on what proposals are funded among those submitted). From the perspective of the MGD, the mouse ENCODE data will provide a rich source of new transcriptional, structural, and regulatory genome features that appropriately would be integrated into MGD and will provide new sources of biological insight into the functioning of the genome.

Your thoughts on developing a supplement to MGD to facilitate integration of relevant mouse ENCODE data seems like a great approach. The DCC would store the primary datasets. Relevant data could then be transferred to MGD once experiments have been processed and analyzed. You would facilitate this by providing resulting datasets to us in a form that MGD could import and integrate. In integrating these data, MGD would provide you the standardized nomenclature for the new genome features added, to allow us to easily associate them with existing genes and genome features in MGD. This will ensure that they are in the proper genome context and correctly linked to other biological information.

Given that it is an unknown as to how many or what kind of mouse projects will be awarded, we agree that such a supplement would need to be created later. I agree, and will be happy to work with you on this when the time is right.

In the meantime, best of luck on securing the ENCODE DCC award.

Yours,

Janan T. Eppig

Janan T. Eppig
Professor
Mouse Genome Database
The Jackson Laboratory
600 Main Street
Bar Harbor, Maine 04609

HHMI
HOWARD HUGHES MEDICAL INSTITUTE

December 8, 2011

J. Michael Cherry
Stanford Unversity

Dear Mike,

I am strongly in support of your application to run the ENCODE—Data Coordination Center (DCC).

From your perspective as a MOD leader, you will naturally ensure that data from the ENCODE project is accessible to MODs such as WormBase. I think your plan to arrange resources for a WormBase curator to handle any nematode ENCODE data produced is an excellent plan. We will be delighted to find and supervise such a curator.

WormBase is continuing to enjoy having modENCODE data, but look forward to a more streamlined incorporation process such that you envision.

Sincerely yours,

Paul W Sternberg

DEPARTMENT OF MOLECULAR AND
CELLULAR BIOLOGY
HARVARD UNIVERSITY

The Biological Laboratories
16 Divinity Avenue
Cambridge, Massachusetts 02138

William M. Gelbart
Phone: (617) 495-2906
Fax: (617) 496-1354
email: gelbart@morgan.harvard.edu

December 8, 2011

J. Michael Cherry
Center for Genomics and Personalized Medicine
Department of Genetics
1501 S. California Ave, Rm 2419
Palo Alto, California 94304-5577

Dear Mike,

I'm writing my letter to confirm FlyBase's interest in collaborating with the funded ENCODE DCC to help manage, synchronize and provide the scientific community with FlyBase-integrated access to any Drosophila data generated through grants funded as part of the upcoming ENCODE competition. From long experience of FlyBase and, more particularly, through our experience working with the modENCODE groups, I agree that anything that could be done to work together from the start of the project will be of great benefit to us all. You have suggested that the simplest mechanism for funding FlyBase's part of the data management would be through a supplement to FlyBase. While I certainly would entertain other options, I agree with you that the supplement would be the most straightforward.

Sincerely,

William M Gelbart

Bill Gelbart

BCM
Baylor College of Medicine

Aleksandar Milosavljevic, Ph.D.
Associate Professor
Department of Molecular and
Human Genetics
One Baylor Plaza, BCM 225
Jewish Building, Ste. 400D
Houston, Texas 77030-3498
TEL: (713) 798-8719
FAX: (713) 798-5556
e-mail: amilosav@bcm.edu

J. Michael Cherry
Department of Genetics
Stanford University
Stanford, California 94305

December 13th, 2011

Dear Mike,

I am writing to confirm our conversation about your “A Data Coordinating Center for ENCODE” proposal in response to RFA-HG-11-026. I look forward to working jointly to build metadata standards building on those already developed by our projects by adopting shared dictionaries / ontologies for cell and tissue types, assays, and other data elements in order to improve virtual integration of Human Epigenome Atlas releases and other data in the UCSC browser.

With shared standards, the issue of physical data integration is no longer a major issue. As you know the Epigenome Atlas data is already accessible from UCSC, EBI and NCBI and quarterly updates will be provided from us. As we discussed, we ask that our content be properly attributed. We release our updates as under the name Human Epigenome Atlas, which will be referenced in track descriptions at the ENCODE Portal in abbreviated form "EA Release <release number>". Human Epigenome Atlas Release 6 is due by the end of this year.

I am confident that our joint work will advance virtual data integration through the use of metadata standards for data coming form an increasing diversity of assays and projects.

Sincerely,

Aleksandar Milosavljevic

Washington University in St.Louis
SCHOOL OF MEDICINE

Ting Wang, Ph.D.
Department of Genetics

December 13, 2011

J. Michael Cherry, Ph.D.
Associate Professor, Department of Genetics
Stanford University School of Medicine

W. James Kent, Ph.D.
Director, UCSC Genome Browser Project
Research Scientist, Department of Biomolecular Engineering
UC Santa Cruz

Dear Michael and Jim,

I am writing enthusiastically in support of your proposal "A Data Coordinating Center for ENCODE" in response to RFA-HG-11-026. I want to use this opportunity to extend the same strong support from Dr. Joe Costello, PI of the currently funded UCSF-based Reference Epigenome Mapping Center, and PI of our proposed ENCODE U54 data production center to continue our work mapping transcriptomes and DNA methylomes. If our applications are successful we look forward to working with you in your important component of the ENCODE project. We have had a fruitful relationship with Jim and the UCSC group for several years and are confident that the new DCC with the addition of the Stanford group will create an even better resource.

At Washington University my lab has created a resource for NIH's Roadmap Epigenomics project. As you know, the Roadmap project has generated thousands of high quality, sequencing-based epigenomics datasets since the program started at the end of 2008. However, a mechanism that allows data immediately available to the community with tools that allow navigation through the data currently did not exist. It was quickly realized that these data would be useless if the scientific community cannot access and navigate them – they must be put on the UCSC Genome Browser – this is a single important request made by the external scientific advisory board at the Roadmap project steering committee meeting, reflecting the need of the community, and perhaps a dependency on the UCSC Genome Browser.

My lab's goal was to develop an advanced visualization platform to integrate, display and synthesize Roadmap Epigenomics data generated by all four REMCs. This "visualization hub (VizHub)" includes a UCSC Genome Browser mirror at Washington University, which hosts all Roadmap Epigenomics data; a remote data hub, which provides all data tracks of the Roadmap project to display remotely at UCSC's main browser site; and a new Epigenome Browser which has just been launched with a manuscript published at Nature Methods. The VizHub would facilitate data analysis and interpretation, and improve interactions among investigators. First and foremost, it would present the data to the general scientific community in a visually appealing and user friendly form, and in synergy with presentation of data from other consortiums, particularly ENCODE.

Washington University in St.Louis

SCHOOL OF MEDICINE

Ting Wang, Ph.D.
Department of Genetics

The new Epigenome Browser’s front end was developed using JavaScript and other advanced web technologies. I would be happy to share my expertise gained by developing this browser with you for creation of new displays for the ENCODE Portal. It is important that you provide tools that are easy to use and allow general users views of the data that are simple to navigate and display information at a level appropriate to your interest.

The expansion of the metadata pipeline and the use of Data Hubs will allow our center to work effectively with the DCC. The requirement for rich and precise metadata is critical for the ENCODE project. We are interested in working with you to enhance the metadata standards to be captured by the DCC. Your ideas to expand on the experiences of the current ENCODE DCC at UCSC are important. Retrieving our data from a Data Hub is an excellent application of this new tool. We also look forward to learning more about enhancements to the verification processes that will occur at Stanford. Of course these automatic processes are important to accelerate the release of our data. However, we are glad that you will continue providing Data Wranglers to assist us in defining metadata and those handling those difficult issues that always arise when new assays and software are put into place.

We have complete confidence in the success of your proposed work during this next phase of ENCODE, and look forward to further collaboration with you in this exciting venture.

Sincerely yours,

Ting

Ting Wang, PhD
Assistant Professor
Department of Genetics
Center for Genome Sciences
Washington University School of Medicine

DUKE INSTITUTE FOR
GENOME
SCIENCES & POLICY

Gregory E. Crawford, Ph.D.
Assistant Professor

Institute for Genome
Sciences & Policy
Department of Pediatrics
Division of Medical Genetics

Duke University
101 Science Drive
Box 3382
Durham, NC 27708

T 919-684-8196
F 919-668-0795

greg.crawford@duke.edu

December 14, 2011

To:
W. James Kent Ph.D.
University of California Santa Cruz

J. Michael Cherry Ph.D.
Stanford University

Dear Jim and Mike,

I'm writing in enthusiastic support of your proposal for an ENCODE Data Coordinating Center. As a PI of a production lab for the current ENCODE Project, and as a frequent user of the ENCODE databases, I know first hand how critical it is to have intuitive, experienced, and knowledgeable people running the DCC. My group has produced DNase, FAIRE, and ChIP experiments from a large number of diverse human cell types. I continue to notice more and more groups of researchers routinely using data from my lab and other ENCODE groups for their own research, which is great to see, and is a testament to your group.

Even though coordinating data from a diverse number of production groups is challenging, UCSC is doing an excellent job. The procedure for submitting data to the DCC is well thought out, and captures the important and relevant information from of our experiments without being unduly burdensome. The DCC staff is friendly, thoughtful, and gets back to me and my group immediately if we have any questions or concerns. I also have been extremely pleased with how the DCC does an excellent job of checking the data. There was one instance where a glitch on our end resulted in a duplicate dataset being submitted as an independent replicate experiment. Your group quickly caught this mistake and helped us sort out the problem before any of the data went live.

One of the advantages of working in the ENCODE consortium is being able to compare my group's data with that of many other labs working on the same cell lines. For example, we routinely compare DNase and FAIRE data to other ChIP datasets to learn more about the factors that bind to the regions that we identify, as well as identify transcription factor binding sites that do not bind to accessible chromatin. Having the various ENCODE data sets all in common formats, and viewable on a single browser has been a huge help in our research.

I'm glad to see that the UCSC team is planning to continue on working as part of the DCC with the next generation of the ENCODE project. As the project continues to grow in complexity, and is merging in with the ModENCODE project, it makes complete sense to team up with Mike Cherry's group. Mike's experience running SGD and collaborating with the other model organism databases should be very helpful in ensuring a smooth transition of the ModENCODE data into the next DCC. In addition, Mike's work with the Stanford Microarray Database and with the Gene Ontology Consortium will help organize the controlled vocabularies and other metadata, which are so important for the scientific community to understand and make the best use of the ENCODE data. Best of luck with your proposal!

Sincerely,

Gregory E. Crawford

Gregory E. Crawford

UNIVERSITY OF WASHINGTON SCHOOL OF MEDICINE

Dec 12, 2011

J. Michael Cherry Ph.D.
Associate Professor (Research)
Department of Genetics
Stanford University, Stanford CA 94305-5120

W. James Kent Ph.D.
Research Scientist
UCSC Genome Bioinformatics Group:
Center for Biomolecular Science & Engineering
Baskin School of Engineering, University of California, Santa Cruz
1156 High St., Room 373, Santa Cruz, CA 95064

Dear Drs. Cherry and Kent,

I am delighted to write this letter in support of your plan to develop the next iteration of the ENCODE Data Coordination Center (DCC).

As you know, I have worked closely with the DCC in the process of submitting over 500 UW datasets from multiple assay types to ENCODE and mouse ENCODE over the last 4 years. This has involved numerous interactions with DCC staff, from modeling new assay types to routine submissions through to final approval conference calls. Those interactions have given me an appreciation for the level of competence of the entire staff as well as a respect for the overall program design. Of particular note is the focus on quality control and ensuring that information presented to the public via the UCSC Genome Browser is of the highest integrity. I also work in a similar capacity managing the submission of UW data to the Roadmap Epigenome project.

The next phase of ENCODE promises to produce data at higher rate and of more complicated types. I feel that your group is in a great position to expand the capabilities to keep pace and have already proven to me that you can. For example:

- Moving from cell lines to more primary cell data (which will only increase in the next phase) has produced a need for the DCC to develop a donor tracking structure. This is of particular importance when multiple data production centers are producing data from the same donor. The DCC has been able to do this in two parts, an initial, interim solution that satisfied the requirement making data available to the scientific community quickly, while simultaneously developing a more robust, extensible system that will serve more complicated donor coordination requirements in the future.
- Another example of a more complicated type is perturbations, instances of which will help to produce a more dynamic view of genomic regulation. Some simple perturbation data has been produced recently and in that effort we, again working closely with receptive and helpful DCC staff, have developed a suitable meta-data scheme that both modeled the data accurately and allowed for the release of this same data to the public quickly.

It is apparent to me that the current culture of the DCC is such that it promotes scientific curiosity and engagement from its staff towards their interactions with the data production centers. I believe that this atmosphere is vital going forward and have confidence that it will be maintained. I look forward to an opportunity to work with you to make the next phase of ENCODE a success.

Sincerely,

Richard Sandstrom
Sr. Computational Scientist
Lead, Data Management and Dissemination
UW ENCODE Project/ Northwest Reference Epigenome Mapping Center
Dept. of Genome Sciences
University of Washington
2211 Elliott Avenue, 6th Floor
Seattle WA 98121
(206) 383-5162

Institute for
Genomics &
Systems Biology

ROBERT L. GROSSMAN, PH.D.
Core Faculty and Director of Informatics
Institute for Genomics and Systems Biology
University of Chicago
900 East 57th Street
Chicago, IL 60637

J. Michael Cherry
Department of Genetics
Stanford University
Stanford, CA 94305-5120

December 12, 2011

Jim Kent
Center for Biomedical Engineering
University of California Santa Cruz
1156 High St.
Santa Cruz, CA 95064

Dear Mike and Jim,

This is a letter of support for your proposal: A Data Coordinating Center (DCC) in response to the NHGRI solicitation RFA-HG-11-026 (Data Analysis and Coordination Center for the Encyclopedia of DNA Elements (ENCODE)).

Jim has estimated there will be 100 TB of ENCODE data after the first year and that the amount of data will double each year after that for the remaining three years of the grant.

As you know, I have been leading the development of an informatics platform called Bionimbus (www.bionimbus.org) for next gen sequencing, which uses large data clouds to manage and analyze the large datasets that next gen sequencing produces. Bionimbus is designed to support remote next-gen sequencing instruments and integrates technology for managing, archiving, analyzing and transporting large datasets. We have recently partially integrated Bionimbus with Galaxy. There is an open source version of Bionimbus available to those who wish to set up their own clouds.

The Bionimbus cloud peers with the StarLight Facility in Chicago and through StarLight with a variety of high performance research networks, such as the National Lambda Rail and Internet2. Bionimbus integrates a tool that we have developed called UDR that uses a version of rsync for data replication that relies on UDT (a high performance network transport) instead of TCP, and supports encryption.

As we discussed, you will be providing us with adequate hardware to store the data that the ENCODE Project produces. We have a 10G connection to StarLight that we expect to upgrade to 100G this coming year. Via StarLight, we can connect to the San Diego Supercomputing Center (SDSC) using multiple 10G paths.

Via this connection and using UDR, we will create a “hot backup” of the ENCODE data that you are managing at SDSC. In addition, we will contribute some of our own Bionimbus resources so that community can compute over the ENCODE data that we host.

We have used a similar approach and provided the community during the past year with a copy in Bionimbus of the modENCODE data that can be computed over.

I also look forward to working with you on this project to develop novel cloud-based computational methods to perform integrative analyses over the ENCODE data.

In summary, I strongly support your proposal, and if it is funded will provide: 1) high performance data replication services between SDSC and the Chicago Bionimbus Cloud using UDR; 2) a “hot backup” of the ENCODE data that you are managing in the SDSC using UDR; and 3) cloud based computing for the community over the entire ENCODE data set using Bionimbus.

Please let me know if I can provide any additional information.

Sincerely,

Robert Grossman
Chief Research Informatics Officer, Biological Sciences Division
Director of Informatics, Institute for Genomics and Systems Biology
Senior Fellow, Computation Institute
Professor, Department of Medicine, Section of Genetic Medicine
University of Chicago
robert.grossman@uchicago.edu
703 702 9765

B.5.b RESOURCES SHARING

Data Sharing Plan

Research tools and resources will be made available in full accordance with the NIH Grants Policy Statement and the Principles and Guidelines for Recipients of NIH Research Grants and Contracts. This in particular is a commitment to the principle of rapid data release. The policy at the time of this proposal was prepared to allow for a nine-month moratorium on presentations and publications of new data. This policy for the next phase of the ENCODE Project will be revisited by the PIs of the new projects. The DCC PI is in favor of immediate release of data as it fits the overall mission of the ENCODE Portal. Whatever the ENCODE Steering Committee implements the DCC will provide all the necessary notice and documentation to allow users to understand the policy.

The ENCODE Portal created by the DCC will be a comprehensive community bioinformatics resource for sharing the results and hosting the analyses from the ENCODE project. We do not generate animals or reagents. All resources from the DCC project are freely and openly available to all via publicly accessible web pages and download files, as well as being mirrored at the Open Science Data Cloud (OSDC) Chicago Data Center in Illinois and the European Bioinformatics Institute from Hinxton, Cambridge, UK. We endeavor to add value for the community by submitting directly, assisting in the submission and strongly encouraging that all data are submitted to, and thus available from, the appropriate public data archive. We will integrate data and information publicly available from international repositories as appropriate to enhance the information provided by the ENCODE Portal. Details of documentation, versioning, API, database downloads, and tools are discussed in the Research Strategy section above. Data created or assembled by the DCC will be available in standard formats. User communities will be notified of updates to the Portal web site, major data releases, and the availability of new tools via announcements using a variety of media including, web site, newsletters, email, social media, and presentations at meetings.

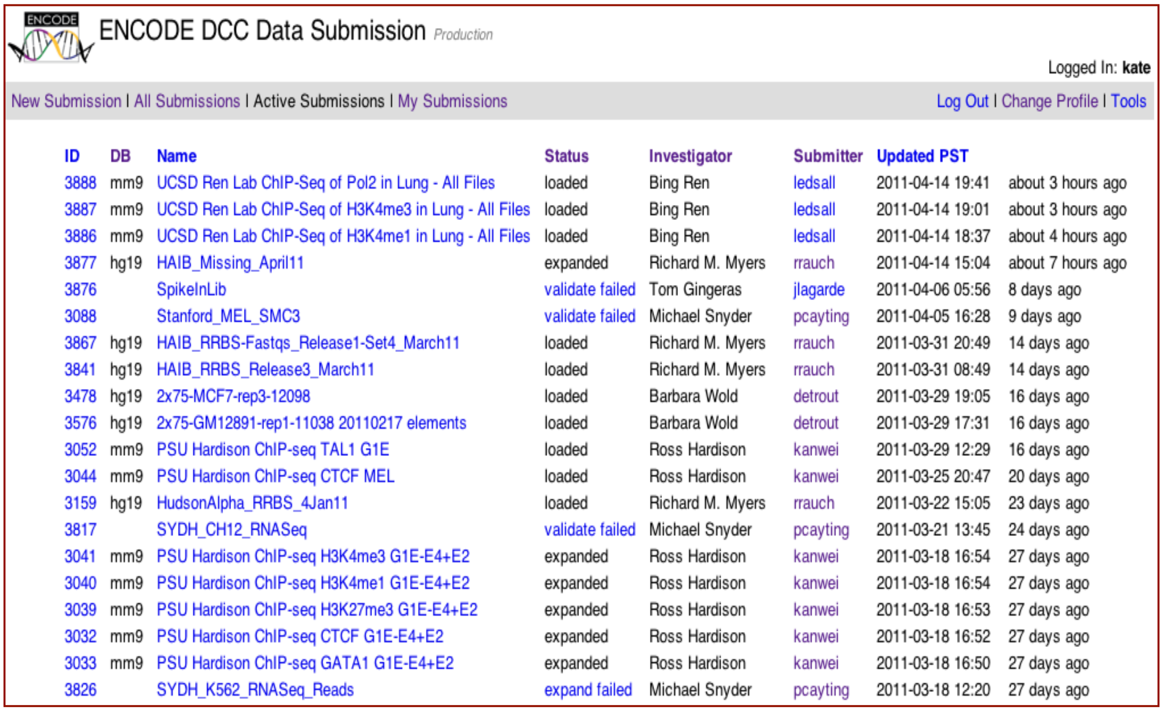
As described above in Aim 2, a software repository will be maintained to distribute all software created by the EDCAC. This repository will use the Github (https://github.com) environment and be freely available to academic and industrial institutions.

Those interested in the commercial resale of the information compiled by the DCC will be referred the National Human Genome Research Institute.

Resource Sharing Plan

As described in the above Aims the DCC will provide a rich resources to share all information released by the ENCODE project. Our plans to promote the ENCODE data for use by all biological researchers are presented in the ENCODE Portal (Aim 3) and the projects outreach (Aim 8). Access to a software repository will be provided via the modern Git software (Aim 2). As needed with will add other tools to enhance the resources we provide (Aim 2).

APPENDICES: TABLE OF CONTENTS



ENCODE DCC Data Submission *Production*

Logged In: **kate**

New Submission | All Submissions | Active Submissions | My Submissions

Log Out | Change Profile | Tools

ID	DB	Name	Status	Investigator	Submitter	Updated PST	
3888	mm9	UCSD Ren Lab ChIP-Seq of Pol2 in Lung - All Files	loaded	Bing Ren	ledsall	2011-04-14 19:41	about 3 hours ago
3887	mm9	UCSD Ren Lab ChIP-Seq of H3K4me3 in Lung - All Files	loaded	Bing Ren	ledsall	2011-04-14 19:01	about 3 hours ago
3886	mm9	UCSD Ren Lab ChIP-Seq of H3K4me1 in Lung - All Files	loaded	Bing Ren	ledsall	2011-04-14 18:37	about 4 hours ago
3877	hg19	HAIB_Missing_April11	expanded	Richard M. Myers	rrauch	2011-04-14 15:04	about 7 hours ago
3876		SpikeInLib	validate failed	Tom Gingeras	jlagarde	2011-04-06 05:56	8 days ago
3088		Stanford_MEL_SMC3	validate failed	Michael Snyder	pcayting	2011-04-05 16:28	9 days ago
3867	hg19	HAIB_RRBS-Fastqs_Release1-Set4_March11	loaded	Richard M. Myers	rrauch	2011-03-31 20:49	14 days ago
3841	hg19	HAIB_RRBS_Release3_March11	loaded	Richard M. Myers	rrauch	2011-03-31 08:49	14 days ago
3478	hg19	2x75-MCF7-rep3-12098	loaded	Barbara Wold	detrout	2011-03-29 19:05	16 days ago
3576	hg19	2x75-GM12891-rep1-11038 20110217 elements	loaded	Barbara Wold	detrout	2011-03-29 17:31	16 days ago
3052	mm9	PSU Hardison ChIP-seq TAL1 G1E	loaded	Ross Hardison	kanwei	2011-03-29 12:29	16 days ago
3044	mm9	PSU Hardison ChIP-seq CTCF MEL	loaded	Ross Hardison	kanwei	2011-03-25 20:47	20 days ago
3159	hg19	HudsonAlpha_RRBS_4Jan11	loaded	Richard M. Myers	rrauch	2011-03-22 15:05	23 days ago
3817		SYDH_CH12_RNASeq	validate failed	Michael Snyder	pcayting	2011-03-21 13:45	24 days ago
3041	mm9	PSU Hardison ChIP-seq H3K4me3 G1E-E4+E2	expanded	Ross Hardison	kanwei	2011-03-18 16:54	27 days ago
3040	mm9	PSU Hardison ChIP-seq H3K4me1 G1E-E4+E2	expanded	Ross Hardison	kanwei	2011-03-18 16:54	27 days ago
3039	mm9	PSU Hardison ChIP-seq H3K27me3 G1E-E4+E2	expanded	Ross Hardison	kanwei	2011-03-18 16:53	27 days ago
3032	mm9	PSU Hardison ChIP-seq CTCF G1E-E4+E2	expanded	Ross Hardison	kanwei	2011-03-18 16:52	27 days ago
3033	mm9	PSU Hardison ChIP-seq GATA1 G1E-E4+E2	expanded	Ross Hardison	kanwei	2011-03-18 16:50	27 days ago
3826		SYDH_K562_RNASeq_Reads	expand failed	Michael Snyder	pcayting	2011-03-18 12:20	27 days ago

Appendix 1.1: Screen shot of ENCODE DCC data submission interface summarizing all submissions. The interface lists the experiments unique identifier for each experiment, the reference genome and organism for that experiment, a human readable name for the submission, the status of the dataset in the pipeline, the PI of the project, the user who submitted the data, and a time stamp for the submission.

Assay	Number submitted
5C	15
CAGE	45
ChIA-pet	8
ChIP-seq	1080
Combined	26
DNA-PET	6
DNase-DGF	44
DNase-seq	145
Exon-array	120
FAIRE-seq	26
Gencode	2
Genotype	64
Mapability	4
Methyl Array	124
Methyl RRBS	93
Methyl-Seq	20
Nucleosome	2
ORChID	1
Proteogenomics	7
RIP-chip Gene ST	48
RIP-chip Tiling Array	8
RIP-seq	8
RNA-chip	26
RNA-PET	19
RNA-seq	192
SwitchGear	2
Grand Total	2135

Appendix 1.2: Number of ENCODE experiments in humans by assay type submitted to the ENCODE DCC as of November 2011.

Assay	Number submitted
ChIP-seq	177
DNase-seq	27
RNA-seq	27
Grand Total	231

Appendix 1.3: Number of ENCODE experiments in mouse by assay type submitted to the ENCODE DCC as of November 2011.

Metadata Field	Description	Controlled vocabulary at ENCODE?
Cell type category	Cell type category, such as T for tissue, L for cell line, P for primary cells	yes
Cell, tissue or DNA sample	Cell line or tissue used as the source of experimental material.	yes
Control or Input for ChIPseq	The type of control (or 'input') used in ChIP-seq experiments to remove background noise before peak calling.	yes
Experiment (Assay) type	The types of experiments such as ChIP-seq, DNAse-seq and RNA-seq.	yes
Cellular compartment	The cellular compartment from which RNA is extracted. Primarily used by the Transcriptome Project.	yes
Mapping algorithm	Algorithm used in high-throughput sequencing experiments to map sequenced tags to a particular location in the reference genome.	yes
Library Protocol	Lab specific protocol that may cover a number of steps in an experiment. Most typically this identifies methods for building a DNA or RNA library.	yes
Tissue Source Type	Source of tissue from either an indiviual organism or pooled set of organisms	yes
Age of donor organism	The age of the organism used to produce tissue or cell line.	yes
Antibody or target protein	The antibody to a specific protein. Used in immuno-precipitation to target certain fractions of biological interest.	yes
Is this auxillary data	This indicates the status of a file in the attic or not. It is an internal piece of metaData.	yes
Length of GIS DNA PET fragments	length of GIS DNA PET fragments, which has different values than fragLength	yes
Principal Investigator on grant	Principle investigator holding the grant by which a set of experiments are financed. Several labs led by other PI's may be under one grant.	yes
Insertion length	The length of the insertion for paired reads for RNA-seq experiments.	yes
Lab producing data	The name of the lab producing the data. Often many labs are working together under one grant or one project.	yes
Cell phase	The phase in a cell cycle. Some experiments attempt to isolate DNA from a single phase.	yes
Paired/Single reads lengths	Specific information about cDNA sequence reads including length, directionality and single versus paired read.	yes
Genomic region(s)	Genomic region(s) targeted by an experiment that is not whole-genome	yes
Restriction Enzyme used	The restriction enzyme used in an experiment, typically for DNA library preparation for a high-throughput sequencing experiment.	yes
RNA Extract	Fraction of total cellular RNA selected for by an experiment. This includes size fractionation (long versus short) and feature frationation (PolyA-, PolyA+, rRNA-).	yes
Sequencing Platform	Sequencing platform used in high-throughput sequencing experiment.	yes

Sex of donor organism	The sex of a cell line or tissue sample affects the genome target of an experiment.	yes
Strain of organism	The strain of the donor organism used in an experiment.	yes
Treatment	Treatment used as an experimental variable in a series of experiments.	yes
View - Peaks or Signals	Different track formats often allow different views of the data of a single experiment. These views sometimes represent different stages of processing, such as experimental 'signal' resulting directly from high-throughput sequencing and called 'peaks' which result from further analysis.	yes
UCSC replicate number	This replicate number is based on submission to the UCSC browser, it may not reflect the replicate number assigned by the contributing lab.	no
Obtained by	The production group that grew the cells and isolated genomic DNA.	no
GEO accession number provided by lab	A generic GEO accession number provided by the producing lab.	no
GENCODE annotation.	GENCODE specifies if an annotation is done manually or automatically.	no
Cross Lab Bio-Replicate ID	Cross lab sample ID number used to track bio-replicates PROVISIONAL.	no
UCSC Composite Track	Related tracks in the UCSC Genome Browser are often grouped into a named composite track.	no
ENCODE Data Freeze	The ENCODE project declares specific data freezes for data to be used in papers or analysis.	no
Date resubmitted to UCSC	Submitted data that was remapped to a new assembly, found to have errors or otherwise needed to be updated will have a date of resubmission.	no
Date submitted to UCSC	Date that a particular file was originally submitted to the UCSC Genome Browser.	no
Date restrictions end	ENCODE data is made publicly available but with restrictions on use for the first nine months since date submitted. After this date, the data is unrestricted.	no
UCSC Accession	The accession number provided by the UCSC Genome Browser.	no
Internal DCC Notes	Notes about tracks that are internal to the DCC.	no
Experiment ID	The ENCODE DCC experiment identification number.	no
Experimental variables	Experiment defining variables for a given ENCODE composite set of tracks.	no
File Name for downloading	The name of a downloadable file associated with a particular track in the browser.	no
Mean Length of DNA fragments	DNA libraries built for ChIP-seq and similar experiments often involve fragmenting the DNA into lengths close to this size.	no
GEO sample accession	GEO sample accession number applied to a single data set in a series of related data sets.	no
GEO series accession	GEO series accession number applied to all data sets in a related series.	no
Lab provided identifier	An ID provided by a lab that uniquely identifies an experiment (dataset) to that lab.	no
Lab specific protocol ID	Some labs produce experiments under multiple defined protocols which may or may not be experiment altering, but which should be recorded all the same.	no

Lab specific details	Free text field for labs to record miscellaneous details of their experiments and submitted data sets.	no
Gencode level	GENCODE level of annotation. Level 1: validated. Level 2: manually annotated. Level 3: automated annotation.	no
md5sum	The md5sum for the file associated with the object.	no
Assembly originally mapped to	Some experiments are originally performed and mapped to an earlier assembly of the reference genome and then are remapped or lifted over to a more recent assembly.	no
Privacy	Privacy may be applied to specific datasets where sequence and other identifying information is not provided	no
Project funded by	Project that funded this and a related set of experiments.	no
Rank of replicate	Experiments with multiple replicates may declare a preferred replicate for visualization.	no
Replicate number	The biological replicate of a particular experiment.	no
Experiment or Input	ChIP-seq experiments often require 'input' or control results to be subtracted from the experimental results. This term clearly identifies which datasets are being used as the experiment and which are the input.	no
Mapability windowing size	The windowing size used, typically in a whole genome computational analysis and characterization dataset.	no
Lab specific informatics	A free-text field for a lab to declare the software and computational methods used in an experiment.	no
Source Objects	Objects this object was built with.	no
Spike-In Pool	The spike-in pool that was used.	no
Submission ID	The submission ID assigned by the UCSC Genome Browser at time of actual data submission.	no
Version of data if resubmitted	When submitted data has been publicly released in the Genome Browser, but subsequently corrected and re-released, this data version will advance from the implicit V1.	no
Name of SQL table at UCSC	Downloadable files containing data displayed in tracks in the UCSC Genome Browser are often stored in SQL tables. The particular table can be located by this name.	no
File type	The type of the primary file associated with the object	no
Mapability Distinctions	Degree of region uniqueness or mapability disctinctions	no

Appendix 1.4: Full list of metadata fields currently captured by ENCODE.

Metadata Field	Example entry	Ontology or controlled vocabulary currently used by modENCODE
Protocols	chromatin_immunoprecipitation	MGED, OBI, internal CV
Sample inputs of protocol	chromatin	MGED, OBI, SO, GO, internal CV
Type of result from protocol	density	MGED, OBI, SO, GO, internal CV
Antibodies	Bre1 (FBgn0086694)	Wormbase or Flybase gene ID of antibody target
RNAi reagents	pncr013:4	Wormbase or Flybase gene ID of RNAi target
Developmental stage	pupae	Wormbase or Flybase developmental stages
Strains	32567	Wormbase or Flybase strain IDs
Tissue	dorsal mesothoracic disc	Wormbase or Flybase anatomy
Cell lines	ML-DmD17-c3	CL, DGRC cell line, anatomy

Appendix 1.5: Sample list of metadata fields captured by modENCODE. The modENCODE DCC uses a combination of DCC-managed controlled vocabulary lists, ontologies, and accepted nomenclature conventions to capture a wide range of metadata. Ontologies used by modENCODE include the Gene Ontology (GO), Sequence Ontology (SO), Ontology of Biomedical Investigations (OBI), Cell Type Ontology (CL), and the MGED Ontology (MO). The modENCODE DCC also uses nomenclature conventions, anatomy terms, and development stages defined and maintained by the model organism databases Flybase and Wormbase and the *Drosophila* stock centers.

	Description of check	Currently done at ENCODE?	Planned for future?
Metadata	All terms are in the controlled vocabulary or ontology	●	●
	Cell lines or tissues have approved protocols	●	●
	Antibodies have validation documents	●	●
	Submitted metadata values are consistent with each other		●
	New cell lines submitted for registration have growth protocols and ordering information		●
	Numerical metadata fields is an integer or floating point as required and in is in the right range.		●
File format	FASTQ files checks sequence name begins with “@” quality line begins with “+” length of sequence matches length of quality line there are 4 lines per sequence	●	●
	FASTA file checks header line begins with “>” sequence contains ACGTNacgtn0-3 and is nonzero length	●	●
	bigwig file checks valid bbi file chromosome names and sizes match the reference genome of organism	●	●
	BedGraph checks each line has 4 fields field 1 is a valid chrom name listed in chromInfo fields 2 and 3 are positive numbers (chrom position) field 3 is greater than field 2 (chrom start < chrom end) field 4 must be a float point number	●	●
	Bed file checks each line has more than 3 fields field 1 is a valid chrom name listed in chromInfo fields 2 and 3 are positive numbers (chrom position) field 3 is greater than field 2 (chrom start < chrom end) if the type is bed5float, the 6th field position exists if the type is bed5float the 6th field position is a floating point number	●	●
	broadPeak file checks check the chrom name check the chrom size check for score between 1 and 1000 check for floats in signalValue, pValue and qValue check strand for +,-	●	●
	narrowPeak all broadPeak checks if peak is >= 0 and <= (end-start)	●	●
	BAM file checks file structure is consistent with format accepted by GEO		●
	Chromosome and extra sequence haplotypes are in the agreed reference genome assembly		●
	All coordinates for a chromosome are in range for that chromosome		●
	Items referenced in two files, such as reads in both FASTQ and BAM files are consistent.		●
	All of the files supporting an experiment are present in all replicates		●

	Description of check	Currently done at ENCODE?	Planned for future?
Data accuracy	Two replicates correlate significantly above chance		•
	There is substantial overlap between RNA experiments and known gene sets		•
	There is substantial overlap between DNAse and other chromatin accessibility experiments and known promoter sets.		•
	ChIP-seq experiments aren't dominated by mappings to centromere fragments or mitochondrial pseudogenes.		•

Appendix 2.1: Description of the validation checks that are currently done during the submission pipeline at ENCODE on the metadata, the file format, and the accuracy of the data. In addition to current checks, new checks are proposed for these three categories of information. These checks will be run at the Certification, Validation, and Verification steps in the new DCC pipeline.

1. **Pre-QA**
   Focus is on reviewing the Human Genome Browser content. The goal is to get track descriptions more polished before full QA.

2. **Run auto-checks & review output**
   Run the encodeQaInit script, which (among other things) runs the following checks, and review output:
   countPerChrom
   check for entry in tableDescriptions table
   check that shortLabel does not exceed 17 characters
   check that longLabel does not exceed 80 characters
   check that there are no underscores in the table names
   check for indices on the tables
   check that positional tables are sorted
   checkTableCoords (checks for any illegal coordinates)

3. **Review the notes file**
   Get familiar with the release and what it consists of. This file contains list of all files and tables and details what is changing for this release compared to last one.

4. **Stage on track beta & check**
   Stage on beta and then make sure everything is staged properly (compare to notes file).

5. **Check the Human Genome Browser for Functionality & Content**
   This is usually where we identify large-scale issues with the data or release. While in checking the controls work and carefully reading the track description, we start exploring and understanding the data of the track.

   5.1 **Functionality** (track controls)
   Check that these all work and defaults are set properly: display modes, configuration settings of views, matrix (including headers), subtrack list, metaData, links.

   5.2 **Content** (.html description page)
   Data is described accurately, labels are in the right convention and are consistent, standard sections are present, and edit all for readability, grammar, spelling, style and format.

6. **Check the details pages of each subtrack type (view) in the combined track.**
   Make sure the details that are displayed correspond with the record in the table and the values seem correct. Make sure that the details displayed seem useful (you understand what is being displayed and why it is useful to the user; any internal or non-functioning fields are not displayed) and that all useful information available in the table is displayed. Make sure the details are presented/labeled clearly in a layout that is user friendly. Make sure and any links work and are labeled clearly.

7. **Check the Human Genome Browser display & performance**
   Check the display and performance of the track in the Human Genome Browser.

   7.1 **Display**
   Check the display of all Views in all display modes when zoomed in to the base pair level & zoomed out to 1 million bp. Check that an items' coordinates and other display features (exons, etc) display as expected/correctly based on table. Are items searchable; should they be? Search for a subtrack in track search. For human, Tier 1 and Tier 2 cell lines should be displayed in a unique color (other than black). Should this composite track be on by default? Check the default subtracks (tier1 & 2 and most important should be on by default, but not too many).

7.2 **Performance**

Check the time it takes to load the first signal subtrack on chrom1 & check time of loading all views for one experiment (e.g. Pol2 in GM12878 cells). Neither should take longer than 1 minute. Look at track in a gene-size region with track's default subtracks on and the default browser tracks to see loading time and how much data is displayed.

8. **Does the data makes sense?**

This is kept in mind through all previous steps, however, this step specifically asks this question in a broader sense. Compare subtracks within views: do all the subtracks within a view somewhat correlate; should they? Compare subtracks of related views, does what you observe make sense? Do the data make sense biologically? Compare with other tracks (e.g. a gene track or compare to subtracks of similar tracks).

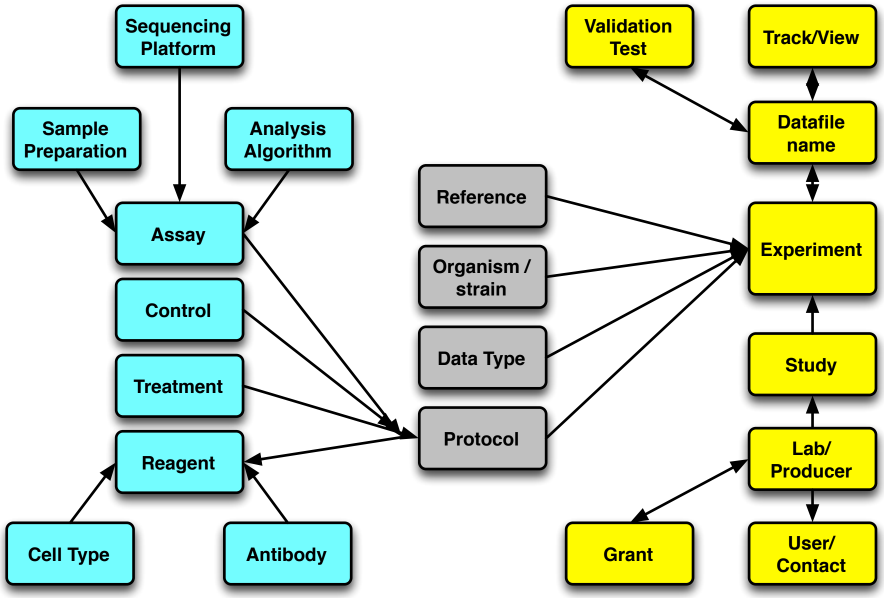
9. **Check files**

Make sure all the files are present. Check the functionality of hgFileUi works (buttons, sorting, fitlering) and that the columns are useful & have useful, concise titles. Check that links to download server works and readme is present/looks good. Check for any internal files & have wrangler remove them.

10. **Push to the production machines for release & check on production machines**

After pushing everything to the production machines, check to make sure the track has all the necessary subtracks, metadata, and files and turn default subtracks on in the Human Genome Browser to verify not only is it there, but that it loads.

Appendix 2.2: A checklist of Quality Assurance checks that are performed on the ENCODE data tracks by the QA group.



Appendix 3.1: A detailed representation of the schema for AnnoDB. There are three major areas of the metadata database: (1) details about reagents and specific assays (indicated by the turquoise shade) (2) grouping of reagents and assays into protocols (indicated by the grey) that are used by the labs, and (3) grouping of protocols and references that comprise an experiment that is done by a lab (indicated by the yellow).