APPLICATION FOR FEDERAL ASSISTANCE

SF 424 (R&R)

2. DATE SUBMITTED 04/26/2009	**Applicant Identifier** 36322
3. DATE RECEIVED BY STATE	**State Application Identifier**

1. * TYPE OF SUBMISSION

❍ Pre-application ● Application ❍ Changed/Corrected Application

4. Federal Identifier

5. APPLICANT INFORMATION *** Organizational DUNS:** 0527809180000

* Legal Name: UNIVERSITY OF MIAMI SCHOOL OF MEDICINE

Department: HUMAN GENETICS-31073 Division:

* Street1: 1507 Levante Avenue Street2:

* City: Coral Gables County: Miami-Dade * State: FL: Florida

Province: * Country: USA: UNITED STATES * ZIP / Postal Code: 33124

Person to be contacted on matters involving this application

Prefix:	* First Name:	Middle Name:	* Last Name:	Suffix:
Mr.	Tom		Gill	

* Phone Number: 305-243-6232 Fax Number: 305-243-4611 Email: grantsgov@med.miami.edu

6. * EMPLOYER IDENTIFICATION NUMBER *(EIN) or (TIN)*:
1590624458A1

7. * TYPE OF APPLICANT
O: Private Institution of Higher Education

Other (Specify):

Small Business Organization Type

❍ Women Owned ❍ Socially and Economically Disadvantaged

8. * TYPE OF APPLICATION: ● New

❍ Resubmission ❍ Renewal ❍ Continuation ❍ Revision

If Revision, mark appropriate box(es).

❍ A. Increase Award ❍ B. Decrease Award ❍ C. Increase Duration

❍ D. Decrease Duration ❍ E. Other *(specify)*:

9. * NAME OF FEDERAL AGENCY:
National Institutes of Health/DHHS

10. CATALOG OF FEDERAL DOMESTIC ASSISTANCE NUMBER:
93.701
TITLE:

* Is this application being submitted to other agencies? ❍ Yes ● No
What other Agencies?

11. * DESCRIPTIVE TITLE OF APPLICANT'S PROJECT:
Integrated NextGen workflow tool and genomic viewer

12. * AREAS AFFECTED BY PROJECT *(cities, counties, states, etc.)*
N/A.

13. PROPOSED PROJECT:

* Start Date	* Ending Date
09/30/2009	09/29/2011

14. CONGRESSIONAL DISTRICTS OF:

a. * Applicant	b. * Project
FL-18	FL-17

15. PROJECT DIRECTOR/PRINCIPAL INVESTIGATOR CONTACT INFORMATION

Prefix:	* First Name:	Middle Name:	* Last Name:	Suffix:
	SAWSAN		KHURI	

Position/Title: Research Assistant Professor * Organization Name: UNIVERSITY OF MIAMI SCHOOL OF MEDICINE

Department: HUMAN GENETICS-31073 Division:

* Street1: 1120 NW 14th ST Street2: 926 Clinical Research Building

* City: Miami County: * State: FL: Florida

Province: * Country: USA: UNITED STATES * ZIP / Postal Code: 33136

* Phone Number: 305-243-6069 Fax Number: 305-243-9304 * Email: skhuri@med.miami.edu

16. ESTIMATED PROJECT FUNDING	
a. * Total Estimated Project Funding	$948,549.00
b. * Total Federal & Non-Federal Funds	$948,549.00
c. * Estimated Program Income	$0.00

17* IS APPLICATION SUBJECT TO REVIEW BY STATE EXECUTIVE ORDER 12372 PROCESS?

a. YES ❍ THIS PREAPPLICATION/APPLICATION WAS MADE AVAILABLE TO THE STATE EXECUTIVE ORDER 12372 PROCESS FOR REVIEW ON:

DATE:

b. NO ● PROGRAM IS NOT COVERED BY E.O. 12372; OR

❍ PROGRAM HAS NOT BEEN SELECTED BY STATE FOR REVIEW

18. By signing this application, I certify (1) to the statements contained in the list of certifications* and (2) that the statements herein are true, complete and accurate to the best of my knowledge. I also provide the required assurances * and agree to comply with any resulting terms if I accept an award. I am aware that any false, fictitious, or fraudulent statements or claims may subject me to criminal, civil, or administrative penalties. (U.S. Code, Title 18, Section 1001)

● * I agree

** The list of certifications and assurances, or an Internet site where you may obtain this list, is contained in the announcement or agency specific instructions.*

19. Authorized Representative

Prefix:	* First Name:	Middle Name:	* Last Name:	Suffix:
Mr.	Tom		Gill	

* Position/Title: Director

* Organization Name: UNIVERSITY OF MIAMI SCHOOL OF MEDICINE

Department: OFFICE OF RESEARCH - 30120

Division:

* Street1: 1400 NW 10th Avenue

Street2: Dominion Tower, 10th Floor, Suite 1007

* City: Miami

County: Miami-Dade

* State: FL: Florida

Province:

* Country: USA: UNITED STATES

* ZIP / Postal Code: 33136

* Phone Number: 305-243-6232

Fax Number: 305-243-4611

* Email: grantsgov@med.miami.edu

*** Signature of Authorized Representative**

Mr. Tom Gill

*** Date Signed**

04/26/2009

20. Pre-application File Name: Mime Type:

21. Attach an additional list of Project Congressional Districts if needed.

File Name: Mime Type:

RESEARCH & RELATED Project/Performance Site Location(s)

Project/Performance Site Primary Location

Organization Name: UNIVERSITY OF MIAMI SCHOOL OF MEDICINE

* Street1: 1120 NW 14th Street		Street2: CRB 926
* City: Miami	County: Miami-Dade	* State: FL: Florida
Province:	* Country: USA: UNITED STATES	* Zip / Postal Code: 33136

	File Name	Mime Type

Additional Location(s)

RESEARCH & RELATED Other Project Information

1. * Are Human Subjects Involved?	❍ Yes ● No	
1.a. If YES to Human Subjects		
Is the IRB review Pending?	❍ Yes ❍ No	
IRB Approval Date:		
Exemption Number:	_ 1 _ 2 _ 3 _ 4 _ 5 _ 6	
Human Subject Assurance Number		
2. * Are Vertebrate Animals Used?	❍ Yes ● No	
2.a. If YES to Vertebrate Animals		
Is the IACUC review Pending?	❍ Yes ❍ No	
IACUC Approval Date:		
Animal Welfare Assurance Number		
3. * Is proprietary/privileged information included in the application?	❍ Yes ● No	
4.a. * Does this project have an actual or potential impact on the environment?	❍ Yes ● No	
4.b. If yes, please explain:		
4.c. If this project has an actual or potential impact on the environment, has an exemption been authorized or an environmental assessment (EA) or environmental impact statement (EIS) been performed?	❍ Yes ❍ No	
4.d. If yes, please explain:		
5.a. * Does this project involve activities outside the U.S. or partnership with International Collaborators?	❍ Yes ● No	
5.b. If yes, identify countries:		
5.c. Optional Explanation:		
6. * Project Summary/Abstract	abstract.pdf	Mime Type: MIMETYPE
7. * Project Narrative	projplan.pdf	Mime Type: MIMETYPE
8. Bibliography & References Cited	ref.pdf	Mime Type: MIMETYPE
9. Facilities & Other Resources	Facilities_Upload.pdf	Mime Type: MIMETYPE
10. Equipment	Major_Equipment_Upload.pdf	Mime Type: MIMETYPE

PROJECT SUMMARY

This application addresses broad Challenge Area (06) Enabling Technologies and specific Challenge Topic 06-HG-101, “New computational and statistical methods for the analysis of large data sets from next-generation sequencing technologies”. NextGen sequencing has enormous potential to push forward genomics research by allowing researchers to ask new questions and is being rapidly adopted by genomics research groups and clinical diagnostics labs. Increased adoption must be accompanied by advances in high-throughput analysis, strategies for mitigating systemic bias, and new sequence assembly and downstream analysis pipelines. We propose to develop optimized bioinformatics workflows for NextGen sequence assembly, gene expression and genomic variation analysis, and implement them in an end-to-end NextGen workflow and genomic visualization tool. The proposed system, named Aqwa (Automated Query and Workflow Agent), will be an easy-to-use tool for routinely running carry out tasks that would be difficult or impossible without significant bioinformatics resources. No similar system exists and Aqwa will be freely distributed as an open source software package. Aqwa is designed for researchers in large sequencing centers, genomics research groups, and clinical diagnostics laboratories. In addition to saving the time of researchers, the system will ultimately improve the quality of analyses that may lead to significant advances in the understanding, treatment and cure of human diseases.

PROJECT NARRATIVE

Rapidly falling NextGen sequencing costs mean increasing demand for NextGen bioinformatics infrastructure and tools. Researchers using NextGen data to ask new questions need high-throughput methods to handle the huge data volumes and new analysis pipelines for assembly and other analyses. We propose to develop optimized bioinformatics workflows for NextGen sequence assembly, gene expression and genomic variation analysis and integrate them into a software system, named Aqwa (Automated Query and Workflow Agent), which will serve as an end-to-end NextGen workflow and genomic visualization tool.

FACILITIES & OTHER RESOURCES

FACILITIES: Specify the facilities to be used for the conduct of the proposed research. Indicate the performance sites and describe capacities, pertinent capabilities, relative proximity, and extent of availability to the project. If research involving Select Agent(s) will occur at any performance site(s), the biocontainment resources available at each site should be described. Under "Other," identify support services such as machine shop, electronics shop, and specify the extent to which they will be available to the project. Use continuation pages if necessary.

Laboratory:

N/A

Clinical:

N/A

Biocontainment Resources Available: Complete if research involving Select Agent(s) will occur at any performance site(s), otherwise indicate N/A.

N/A

Animal:

N/A

Computer:

Client workstations are suitable for running necessary software analysis tools.

Office:

Offices of the Center for Computational Science (CCS) where the work will be performed are located in the Clinical Research Building on the University of Miami's Medical campus.

Other:

High Performance Computing

The CCS High Performance Computing (HPC) environment features approximately 650 central processing units (CPUs) with over 1 terabyte of memory; over four teraflops of computational power split between two distinct architectural paradigms: symmetric multi-processing (SMP) and massively/ embarrassingly parallel processing (M/EPP); a 256 CPU IBM Power5+ SMP cluster with high-speed interconnect, 96 CPU IBM Power4 SMP cluster, and a 32 CPU Sun Scalable Processor Architecture (SPARCIIIi) server; and a Linux Xeon cluster that runs RedHat Enterprise Linux with 256 processor cores for M/EPP programs and algorithms and an aggregate of 334 gigabytes of random access memory (RAM).

Software

Center users have a complete software suite at their fingertips, including standard scientific libraries and numerous optimized libraries and algorithms tuned for the computing environment. All of our programs and algorithms are implemented in 64-bit mode in order to address large memory problems, and we also offer compatible 32-bit libraries and algorithms. In addition, we use the Moab grid scheduling process to maximize the efficiency of our computational resources. Increased efficiency translates into the faster execution of programs, which provides our researchers with more resources.

The scientific software engineering group provides expertise in the areas of requirements identification and definition, systems design and development, and systems implementation and integration. The group provides these services through three teams: Software engineering, project management and application support.
The software engineering team designs and develops software systems, principally in Java. In addition to software development, the software engineering team provides systems design and consulting services. Project management provides leadership for development projects by handling coordination between project team members, collaborators, and others project members. Projects may be software development projects, but also include system section, system implementation and integration projects. The application support team provides expert and system level support for end user application services hosted by CCS. These systems include both internally developed and commercially available systems.

Bioinformatics
CCS's Bioinformatics Program was established to conduct research and offer services and training in the management and analysis of biological and medical/health record data. Our mission is to spearhead bioinformatics capacity at the University of Miami for all biological and medical applications. This includes data management, data mining, and data analysis capacities. We aim to achieve this mission through infrastructure, education, and expertise. In particular, we are providing an online portal for bioinformatics databases and web tools, and offering a number of data analysis services. We are concomitantly leading educational and training initiatives in bioinformatics analysis, and nourishing these activities with high impact bioinformatics research.

iBIS – UM's online Bioinformatics Integrated Services portal
iBIS is a bioinformatics portal that includes links to genomic databases, protein structure databases, clinical genetics databases, as well as numerous software tools for the analysis of gene expression, gene regulation, signaling and metabolic pathways, genomics, proteomics, and systems biology. In addition, the portal allows access to a suite of locally available tools and databases that we maintain on CCS's HPC servers. Access to most components in the portal is freely available to anyone with a University of Miami login name and password. The portal also offers online tutorials for the major databases and web tools, and the CCS Bioinformatics team provides regular training workshops for new users.

Bioinformatics Data Analysis
We provide data analysis training and expertise at a three levels, consulting, preliminary data generation, and fully collaborative, based on the time and complexity of the service requested. The analyses are undertaken by skilled analysts, and overseen by experienced faculty. We have been working mostly with microarray data and next generation sequencing data, and our analytical services include, but are not limited to, the following:

- Gene expression analysis for transcriptome profiling and/or gene regulatory network building,
- prognostics and/or diagnostic biomarker discovery,
- microRNA target analysis,
- copy number variant analysis, in this context we are testing the few existing algorithms and developing new ones for accurate and unambiguous discovery of copy number variation in the human genome,
- genome or transcriptome assembly from next generation sequencing data, and its visualization,
- SNP functionality analysis,
- Other projects include merging or correlating data from various data types for a holistic view of a particular pathway or disease process.

EQUIPMENT RESOURCES

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

High Performance Computing (HPC) environment:
The CCS High Performance Computing (HPC) environment features approximately 650 central processing units (CPUs) with over 1 terabyte of memory; over four teraflops of computational power split between two distinct architectural paradigms: symmetric multi-processing (SMP) and massively/ embarrassingly parallel processing (M/EPP); a 256 CPU IBM Power5+ SMP cluster with high-speed interconnect, 96 CPU IBM Power4 SMP cluster, and a 32 CPU Sun Scalable Processor Architecture (SPARCIIIi) server; and a Linux Xeon cluster that runs RedHat Enterprise Linux with 256 processor cores for M/EPP programs and algorithms and an aggregate of 334 gigabytes of random access memory (RAM).

Data Storage:
HPC at the CCS incorporates an integrated storage environment for both structured (relational) and unstructured (flat file) data. Our three high-performance general parallel file systems (GPFS) supply approximately 150 terabytes to store unstructured data. These systems are specifically tuned for data type and application requirements for serial access or highly parallelized. Servers and clients access data through a high-performance 10 gigabyte switch over a network file system (NFS). Clients also access data over multiple trunked gigabit lines, utilizing Samba.

The CCS offers structured data services through the most common relational database formats, including: Oracle, MySQL, and PostgreSQL. Investigators and project teams can access their space through SOA, and utilize their resources with the support of our integrated backend infrastructure.

Software
Center users have a complete software suite at their fingertips, including standard scientific libraries and numerous optimized libraries and algorithms tuned for the computing environment. All of our programs and algorithms are implemented in 64-bit mode in order to address large memory problems, and we also offer compatible 32-bit libraries and algorithms. In addition, we use the Moab grid scheduling process to maximize the efficiency of our computational resources. Increased efficiency translates into the faster execution of programs, which provides our researchers with more resources.

RESEARCH & RELATED Senior/Key Person Profile (Expanded)

PROFILE - Project Director/Principal Investigator

Prefix	* First Name	Middle Name	* Last Name	Suffix
	SAWSAN		KHURI	

Position/Title: Research Assistant Professor

Department: HUMAN GENETICS-31073

Organization Name: UNIVERSITY OF MIAMI SCHOOL OF MEDICINE

Division:

* Street1: 1120 NW 14th ST

Street2: 926 Clinical Research Building

* City: Miami

County:

* State: FL: Florida

Province:

* Country: USA: UNITED STATES

* Zip / Postal Code: 33136

*Phone Number	Fax Number	* E-Mail
305-243-6069	305-243-9304	skhuri@med.miami.edu

Credential, e.g., agency login: sawsankhur

*** Project Role:** PD/PI

Other Project Role Category:

	File Name	Mime Type
***Attach Biographical Sketch**	Bio_SAWSAN_KHURI_0.pdf	MIMETYPE
Attach Current & Pending Support		

PROFILE - Senior/Key Person_

Prefix	* First Name	Middle Name	* Last Name	Suffix
	NICHOLAS	F.	TSINOREMAS	

Position/Title: Research Professor

Department: CENTER FOR COMPUTATIONAL SCIEN

Organization Name: UNIVERSITY OF MIAMI

Division:

* Street1: 1190 Clinical Research Bldg

Street2: 1120 NW 14th Street

* City: Miami

County:

* State: FL: Florida

Province:

* Country: USA: UNITED STATES

* Zip / Postal Code: 33136

*Phone Number	Fax Number	* E-Mail
(305)243-7990	3052439304	Ntsinoremas@miami.edu

Credential, e.g., agency login: tsinoremas

*** Project Role:** Other (Specify)

Other Project Role Category: Co-Investigator

	File Name	Mime Type
***Attach Biographical Sketch**	Bio_NICHOLAS_F._TSINOREMAS_1.pdf	MIMETYPE
Attach Current & Pending Support		

RESEARCH & RELATED Senior/Key Person Profile (Expanded)

Additional Senior/Key Person Form Attachments

When submitting senior/key persons in excess of 8 individuals, please attach additional senior/key person forms here. Each additional form attached here, will provide you with the ability to identify another 8 individuals, up to a maximum of 4 attachments (32 people).

The means to obtain a supplementary form is provided here on this form, by the button below. In order to extract, fill, and attach each additional form, simply follow these steps:

- Select the "Select to Extract the R&R Additional Senior/Key Person Form" button, which appears below.
- Save the file using a descriptive name, that will help you remember the content of the supplemental form that you are creating. When assigning a name to the file, please remember to give it the extension ".xfd" (for example, "My_Senior_Key.xfd"). If you do not name your file with the ".xfd" extension you will be unable to open it later, using your PureEdge viewer software.
- Using the "Open Form" tool on your PureEdge viewer, open the new form that you have just saved.
- Enter your additional Senior/Key Person information in this supplemental form. It is essentially the same as the Senior/Key person form that you see in the main body of your application.
- When you have completed entering information in the supplemental form, save it and close it.
- Return to this "Additional Senior/Key Person Form Attachments" page.
- Attach the saved supplemental form, that you just filled in, to one of the blocks provided on this "attachments" form.

Important: Please attach additional Senior/Key Person forms, using the blocks below. Please remember that the files you attach must be Senior/Key Person Pure Edge forms, which were previously extracted using the process outlined above. Attaching any other type of file may result in the inability to submit your application to Grants.gov.

1) Please attach Attachment 1	
2) Please attach Attachment 2	
3) Please attach Attachment 3	
4) Please attach Attachment 4	

ADDITIONAL SENIOR/KEY PERSON PROFILE(S)	**Filename** **MimeType**
Additional Biographical Sketch(es) (Senior/Key Person)	**Filename** **MimeType**
Additional Current and Pending Support(s)	**Filename** **MimeType**

BIOGRAPHICAL SKETCH

Provide the following information for the 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	POSITION TITLE
Tsinoremas, Nicholas F., Ph.D.	Founding Director, Center for Computational Science
eRA COMMONS USER NAME tsinoremas	

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

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	YEAR(s)	FIELD OF STUDY
University of Athens, Greece	BS	1988	Chemistry
University of Leeds, England	Ph.D.	1991	Biochemistry and Molecular Biology

A. POSITIONS AND HONORS

1989 - 1991 Research Fellow, University of Leeds, Department de Physiologie Microbienne, Institut Pasteur, Paris, France
1992 - 1994 Research Associate, Texas A&M University
1994 - 1997 Research Scientist, Department of Biology, Texas A&M University
1997 - 1998 Research Scientist III, Department of Molecular Biology and Biochemistry, Energy Biosystems Corporation
1998 Biocomputational Scientist, Department of Computational Biology, Progenitor, Inc.
1999 Bioinformatics Scientist, Department of Bioinformatics, Incyte Pharmaceuticals
1999 - 2001 Director of Research, DoubleTwist, Inc.
2001 - 2002 Vice President of Genomics and Data Mining Tools, DoubleTwist, Inc.
2002 - 2004 Director of Genomic Discovery and Computational Genomics, Informatics, MRL Seattle, Merck & Co.
2004 - 2007 Director and Head of Bioinformatics, Scripps Florida, Jupiter, Florida
2005 - 2007 Research Professor, Department of Chemistry and Biochemistry, Florida Atlantic University
2006 - 2007 Professor, Department of Biochemistry and Genetics, University of Florida
2007 - Founding Director, Center for Computational Science, University of Miami
2007 - Research Professor, Department of Medicine, Leonard M. Miller School of Medicine, University of Miami
2007 - Research Professor, Department of Computer Science, University of Miami

OTHER EXPERIENCE AND PROFESSIONAL MEMBERSHIPS

2005- Scientific Advisory Board, NextBio Inc.
2006 - College of Engineering and Computer Science Executive Advisory Council, Florida Atlantic University
2006 - Biotechnology Advisory Committee, Miami Dade College
2006 - Board Member, Palm Beach County Library System

AWARDS

SERC Fellowship, Departement de Physiologie Microbienne, Institut Pasteur, Paris, France 1989-1991

B. PUBLICATIONS

1) CR Anderson, NF Tsinoremas, J Shelton, NV Lebedeva, J Yarrow, H Min, SS Golden (2000) Application of bioluminescence to the study of circadian rhythms in cyanobacteria. Methods Enzymol. 305:527-542

2) Z Kan, J Castle, JM Johnson, NF Tsinoremas. Detection of novel splice forms in human and mouse using cross-species approach. Pacific Symposium Biocomput. 2004:42-53
3) CK Raymond, J Castle, P Garrett-Engele, CD Armour, Z Kan, NF Tsinoremas, JM Johnson (2004) Expression of alternatively spliced sodium channel alpha-subunit genes. Unique splicing patterns are observed in dorsal root ganglia. J Biol Chem. 2004 Oct 29; 279(44):46234-41
4) EE Schadt, SW Edwards, D GuhaThakurta, D Holder, L Ying, V Svetnik, A Leonardson, KW Hart, A Russell, G Li, G Cavet, J Castle, Z Kan, R Chen, A Kasarskis, M Margarint, M Caceres, J Johnson, CD Armour, PW Garrett-Engele, NF Tsinoremas, DD Shoemaker (2004) A comprehensive transcript index of the human genome using microarrays and computational approaches. Genome Biol. 5(10):R73. Epub 2004 Sep 23
5) AC Cervino, M Gosink, M Fallahi, B Pascal, C Mader, NF Tsinoremas. A comprehensive mouse IBD database for the efficient location of quantitative trait loci. Mammalian Genome. 2006 June 17
6) AC Cervino, A Darvasi, M Fallahi, CC Mader, NF Tsinoremas. An integrated insilico gene mapping strategy in inbred mice. Genetics. 2006 October 8
7) BD Pascal, MJ Chalmers, SA Busby, CC Mader, MR Southern, NF Tsinoremas, PR Griffin. The Deuterator: software for the determination of backbone amide deuterium levels from H/D exchange MS data. BMC Bioinformatics. 2007 May 16; **8:**156
8) LS Liebovitch, NF Tsinoremas, A Pandya. Analysis of Biological Networks by Artificial Neural Networks. Society for Chaos Theory in Psychology and Life Sciences. 2006 August 3 – 6
9) AC Cervino, NF Tsinoremas, RW Hoffman. A genome-wide study of lupus: preliminary analysis and data release. Annals of the New York Academy of Science. 2007 September; 1110:131-9
10) MM Gosink, HT Petrie, NF Tsinoremas. Electronically subtracting expression patterns from a mixed cell population. Oxford Journals. 2007 October 22

C. RESEARCH SUPPORT

Completed

NGA/MLSCN 1 U54 MH074404-01
"Scripps Research Institute Molecular Screening Center"
This study is focused on high throughput chemical approaches integrated with state-of-the art post-genome sequence, cell, molecular, and in vivo biology to provide a rapid and facile mechanism for enhancing the process of biomedical science and the discovery of proof-of-concept molecules.
Role: Co-Investigator

MLSCN 1 U54 HG003914-01
"Molecular Library Screening Centers Network Center at Columbia University"
This study is focused on high throughput screening using phenotypic assays at the cellular and subcellular levels to identify bioactive compounds.
Role: Investigator – Informatics/Chemoinformatics

Dyadic SFP 1640
"Annotation of the Chrysoporium lucknowense ("C1") genome"
This study is focused on obtaining and annotating the genomic DNA sequence of C1, a proprietary fungal organism.
Role: Investigator

Ongoing

None.

BIOGRAPHICAL SKETCH

Provide the following information for the 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	POSITION TITLE
Young, Stuart	Post Doctoral Associate
eRA COMMONS USER NAME (credential, e.g., agency login)	

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

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	YEAR(s)	FIELD OF STUDY
University of Leicester, England	BSc	1986-1989	Biological Sciences
University of Edinburgh, Scotland	MA	1991-1995	Arabic and French
Open University, England	MBA	1999-2003	Technology Management
Keck Graduate Institute, CA	MS	2002-2003	Applied Biosciences
Indiana University, Bloomington, IN	MS	2003-2004	Bioinformatics
North Carolina State University, NC	PhD	2005-2009	Bioinformatics

A. Positions and Honors.

Positions and Employment

1995-1996 Coordinator for International Relations, Shiojiri City Hall, Nagano Prefecture, Japan
1996-1997 News Reporter, Reuters Japan, Tokyo, Japan
1999-2000 High-Tech Beat Reporter, Taipei Times, Taipei, Taiwan
2000-2000 Freelance Journalist, Business Week Magazine, Boston, MA
2000-2001 Investment Analyst, Primasia Securities, Taipei, Taiwan
2001-2002 Freelance Journalist, Taipei, Taiwan
2002-2003 Freelance Science Journalist, Keck Graduate Institute, Claremont, CA
2003-2004 Research Assistant (Bioinformatics), Indiana University, Bloomington, IN
2005-2008 Research Assistant (Bioinformatics), North Carolina State University, Raleigh, NC
2008-Present Post Doctoral Associate, Center for Computational Science, University of Miami, Miami, FL

Other Experience and Professional Memberships

1998-1989 Cultural Secretary, International Students Society, University of Leicester
1987-1988 President, Scientists Society, University of Leicester
2003-2004 Indiana University Teaching Assistantship

Honors

1998-1999 Johns Hopkins Academic Scholarship (Hopkins-Nanjing Center)
2002-2003 Keck Scholarship (Keck Graduate Institute)
2002-2003 Pioneer Hybrid Scholarship
2004-2008 Ronald McNair Scholar
2005-2008 UNC Genomics Scholar

B. Selected peer-reviewed publications (in chronological order).

None.

C. Research Support

None.

BIOGRAPHICAL SKETCH

Provide the following information for the 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	POSITION TITLE
Khuri, Sawsan	Bioinformatics Lead Scientist & Assistant Research Professor
eRA COMMONS USER NAME	

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

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	YEAR(s)	FIELD OF STUDY
The University of Reading, UK	Post-Doc	1994-1997	Molecular Biology
The University of Reading, UK	Sr Post-Doc	1997-2000	Bioinformatics
The University of Reading, UK	BSc	1985-1988	Agricultural Botany
Imperial College at Wye, University of London,	PhD	1988-1992	Plant Sciences

A. Positions and Honors.

Positions and Employment

2000 - 2006 Honorary Research Fellow, The University of Reading, UK
2002 - Visiting Assistant Professor, Dept. of Biology, American University of Beirut, Lebanon
2003 - 2007 Assistant Research Professor (Voluntary/Per Diem), The Dr. John T. Macdonald Foundation, Center for Medical Genetics, University of Miami Miller School of Medicine, Miami
2007–present Bioinformatics Lead Scientist, Center for Computational Science, Miami
2007–present Assistant Research Professor, Dr. John T. Macdonald Foundation Department of Human Genetics, University of Miami Miller School of Medicine, Miami

Other Experience

1992 - present Consultant, American University of Beirut, Lebanon
2001 - 2002 Research Associate, National Tropical Botanical Garden, Hawaii
2003 - 2004 Bioinformatics Consultant, Fidelity Biosciences Group, Boston
2005 - present Director, Cedrus Scientific Consulting, Miami
2005 - present Reviewer for NSF proposals on protein evolution and on chromatin structure

Honors and Awards

Winner (1988), UK government competitive award for PhD studies at the University of London
Winner (1997), Youth Entrepreneurs Scheme, UK

B. Selected peer-reviewed publications (in chronological order).

1. **Khuri S** and Moorby J (1995) Investigations into the role of sucrose in potato cv. Estima microtuber production *in vitro*. *Annals of Botany* **75**:295-303.5.
2. **Khuri S** and Talhouk SN (1999) Cedar of Lebanon. In *Conifers: Status Survey and Action Plan*, A. Farjon and C.N. Page (compilers), SSC/IUCN, pp 108-111.
3. Dunwell JM, **Khuri S** and Gane PJ (2000) Microbial relatives of the seed storage proteins of higher plants:
conservation of structure and diversification of function during evolution of the cupin superfamily. *Microbiology and Molecular Biology Reviews* **64**: 153-179.1.
4. **Khuri S**, Shmoury MR, Baalbaki R, Maunder M and Talhouk SN (2000) Conservation of the *Cedrus libani* populations in Lebanon: history, current status and experimental application of somatic embryogenesis. *Biodiversity and Conservation* **9**: 1261-1273.

5. **Khuri S**, Bakker FT and Dunwell JM (2001) Phylogeny, function and evolution of the cupins, a structurally conserved, functionally diverse superfamily of proteins. *Molecular Biology and Evolution* **18**: 593-605. Talhouk SN, Zurayk R and **Khuri S** (2001) Conifer conservation in Lebanon: Past, present and future prospects. *Oryx* **35**: 206-215.
6. Gibbings JG, Cook BP, Dufault MR, Madden SL, **Khuri S**, Turnbull CJ, and Dunwell JM (2003) Global transcript analysis of rice leaf and seed using SAGE technology. *Plant Biotechnology Journal* **1**: 271-285.
7. Dunwell JM, Purvis A and **Khuri S** (2004) Cupins: The most functionally diverse protein superfamily? *Phytochemistry* **65**: 7-17
8. Milledge T, **Khuri S**, Wei X, Yang C, Zheng G and Narasimhan G (2005) Sequence-Structure Patterns: Discovery and Applications. *Proceedings of the 6th Atlantic Symposium on Computational Biology and Genome Informatics (CBGI),* p1282-1285, July 2005.
9. Mnayer L, **Khuri S**, Al-Ali HM, Meroni G, and Elsas LJ (2006) A structure-function study of MID1 mutations associated with a mild Opitz phenotype. *Molecular Genetics and Metabolism* 87:198-203.
10. Luykx P, Bajic I, and **Khuri S** (2006) NXSensor: web tool for evaluating DNA for nucleosome exclusion sequences and accessibility to binding factors. *Nucleic Acids Research* 34: W560-W565.
11. Radwan A, Younis A, Hernandez MA, Ho H, Popa L, Shivaji S, and **Khuri S** (2007) BioFederator: A Data Federation System for Bioinformatics on the Web. Paper accepted for the Sixth International Workshop on Information Integration on the Web (IIWeb 2007), July 23, 2007, Vancouver, Canada. *Please note that conference papers in engineering are equivalent to peer reviewed journal articles in the biological sciences.*
12. P Jayakar, K Wierenga, **S Khuri**, R Lopez. A novel mutation in PLP is associated with severe Pelizaeus-Merzbacher Disease (PMD). Accepted to the ACMG conference, 2007
13. **S Khuri**, A Radwan, P Luykx, A Younis (2007) Nucleosome Exclusion Regions across the Human Genome. Poster presented at the 57th American Society for Human Genetics annual meeting, San Diego Oct 23-27.
14. Radwan A, Younis A, Luykx P, **Khuri S** (2008) Prediction and analysis of nucleosome exclusion regions in the human genome. BMC Genomics 9:186 [Epub ahead of print]
15. Badia R, Dasgupta G, Ezenwoye O, Fong L, Ho H, **Khuri S**, Liu Y, Luis S, Praino A, Prost J-P, Radwan A, Sadjadi SM, Shivaji S, Viswanathan B, Welsh P, Younis A (2008) Innovative Grid Technologies Applied to Bioinformatics and Hurricane Mitigation. In L. Grandinetti (ed.) *High Performance Computing and Grids in Action*, Advances in Parallel Computing, Volume 16, pp.436-462.

C. Research Support.

None

RESEARCH & RELATED BUDGET - SECTION A & B, BUDGET PERIOD 1

* ORGANIZATIONAL DUNS: 0527809180000

*** Budget Type:** ● Project ○ Subaward/Consortium

Enter name of Organization: UNIVERSITY OF MIAMI SCHOOL OF MEDICINE

*** Start Date:** 09-30-2009 *** End Date:** 09-29-2010 **Budget Period: 1**

A. Senior/Key Person

	Prefix	* First Name	Middle Name	* Last Name	Suffix	* Project Role	Base Salary ($)	Cal. Months	Acad. Months	Sum. Months	* Requested Salary ($)	* Fringe Benefits ($)	* Funds Requested ($)
1.		SAWSAN		KHURI		PD/PI	119,577.00	2.4			23,915.00	5,501.00	29,416.00
2.		NICHOLAS	F.	TSINOREMAS		Co-Investigator	204,616.00	1.2			20,462.00	4,706.00	25,168.00

Total Funds Requested for all Senior Key Persons in the attached file

Additional Senior Key Persons: File Name: Mime Type: **Total Senior/Key Person** **54,584.00**

B. Other Personnel

* Number of Personnel	* Project Role	Cal. Months	Acad. Months	Sum. Months	* Requested Salary ($)	* Fringe Benefits	* Funds Requested ($)
1	Post Doctoral Associates	12			43,690.00	13,063.00	56,753.00
	Graduate Students						
	Undergraduate Students						
	Secretarial/Clerical						
1	Software Engineering Advisor	1.2			13,458.00	4,024.00	17,482.00
1	Sr. Software Engineer	6			45,771.00	13,685.00	59,456.00
1	Jr. Software Engineer	12			72,817.00	21,772.00	94,589.00
4	**Total Number Other Personnel**				**Total Other Personnel**		**228,280.00**
					Total Salary, Wages and Fringe Benefits (A+B)		**282,864.00**

RESEARCH & RELATED Budget {A-B} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTION C, D, & E, BUDGET PERIOD 1

* ORGANIZATIONAL DUNS: 0527809180000

*** Budget Type:** ● Project ❍ Subaward/Consortium

Enter name of Organization: UNIVERSITY OF MIAMI SCHOOL OF MEDICINE

*** Start Date:** 09-30-2009 *** End Date:** 09-29-2010 **Budget Period: 1**

C. Equipment Description

List items and dollar amount for each item exceeding $5,000

Equipment Item	* Funds Requested ($)
1. Hardware	60,000.00

Total funds requested for all equipment listed in the attached file

Total Equipment **60,000.00**

Additional Equipment: File Name: Mime Type:

D. Travel **Funds Requested ($)**

1. Domestic Travel Costs (Incl. Canada, Mexico, and U.S. Possessions)
2. Foreign Travel Costs

Total Travel Cost

E. Participant/Trainee Support Costs **Funds Requested ($)**

1. Tuition/Fees/Health Insurance
2. Stipends
3. Travel
4. Subsistence
5. Other:

Number of Participants/Trainees **Total Participant/Trainee Support Costs**

RESEARCH & RELATED Budget {C-E} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTIONS F-K, BUDGET PERIOD 1

* ORGANIZATIONAL DUNS: 0527809180000

*** Budget Type:** ● Project ❍ Subaward/Consortium

Enter name of Organization: UNIVERSITY OF MIAMI SCHOOL OF MEDICINE

*** Start Date:** 09-30-2009 *** End Date:** 09-29-2010 **Budget Period: 1**

F. Other Direct Costs	**Funds Requested ($)**
Total Other Direct Costs	

G. Direct Costs	**Funds Requested ($)**
Total Direct Costs (A thru F)	**342,864.00**

H. Indirect Costs

Indirect Cost Type	**Indirect Cost Rate (%)**	**Indirect Cost Base ($)**	*** Funds Requested ($)**
1. MEDICAL ON CAMPUS-RESEARCH	53	282,865.00	149,918.00
		Total Indirect Costs	**149,918.00**

Cognizant Federal Agency DHHS, Darryl Mayes, 202-401-0215
(Agency Name, POC Name, and POC Phone Number)

I. Total Direct and Indirect Costs	**Funds Requested ($)**
Total Direct and Indirect Institutional Costs (G + H)	**492,782.00**

J. Fee	**Funds Requested ($)**

K. * Budget Justification	File Name: budget_justification_per1.pdf (Only attach one file.)	Mime Type: MIMETYPE

RESEARCH & RELATED Budget {F-K} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTION A & B, BUDGET PERIOD 2

* ORGANIZATIONAL DUNS: 0527809180000

*** Budget Type:** ● Project ○ Subaward/Consortium

Enter name of Organization: UNIVERSITY OF MIAMI SCHOOL OF MEDICINE

*** Start Date:** 09-30-2010 *** End Date:** 09-29-2011 **Budget Period: 2**

A. Senior/Key Person

	Prefix	* First Name	Middle Name	* Last Name	Suffix	* Project Role	Base Salary ($)	Cal. Months	Acad. Months	Sum. Months	* Requested Salary ($)	* Fringe Benefits ($)	* Funds Requested ($)
1.		SAWSAN		KHURI		PD/PI	123,164.00	2.4			24,633.00	5,666.00	30,299.00
2.		NICHOLAS	F.	TSINOREMAS		Co-Investigator	210,754.00	1.2			21,075.00	4,847.00	25,922.00

Total Funds Requested for all Senior Key Persons in the attached file

Additional Senior Key Persons: File Name: Mime Type: **Total Senior/Key Person** **56,221.00**

B. Other Personnel

* Number of Personnel	* Project Role	Cal. Months	Acad. Months	Sum. Months	* Requested Salary ($)	* Fringe Benefits	* Funds Requested ($)
1	Post Doctoral Associates	12			45,001.00	13,455.00	58,456.00
	Graduate Students						
	Undergraduate Students						
	Secretarial/Clerical						
1	Software Engineering Advisor	1.2			13,862.00	4,145.00	18,007.00
1	Sr. Software Engineer	6			47,144.00	14,096.00	61,240.00
1	Jr. Software Engineer	12			75,002.00	22,425.00	97,427.00
4	**Total Number Other Personnel**				**Total Other Personnel**		**235,130.00**
					Total Salary, Wages and Fringe Benefits (A+B)		**291,351.00**

RESEARCH & RELATED Budget {A-B} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTION C, D, & E, BUDGET PERIOD 2

* ORGANIZATIONAL DUNS: 0527809180000

*** Budget Type:** ● Project ❍ Subaward/Consortium

Enter name of Organization: UNIVERSITY OF MIAMI SCHOOL OF MEDICINE

*** Start Date:** 09-30-2010 *** End Date:** 09-29-2011 **Budget Period: 2**

C. Equipment Description

List items and dollar amount for each item exceeding $5,000

Equipment Item	* Funds Requested ($)
1. Hardware	10,000.00

Total funds requested for all equipment listed in the attached file

Total Equipment **10,000.00**

Additional Equipment: File Name: Mime Type:

D. Travel	Funds Requested ($)
1. Domestic Travel Costs (Incl. Canada, Mexico, and U.S. Possessions)	
2. Foreign Travel Costs	
Total Travel Cost	

E. Participant/Trainee Support Costs	Funds Requested ($)
1. Tuition/Fees/Health Insurance	
2. Stipends	
3. Travel	
4. Subsistence	
5. Other:	
Number of Participants/Trainees	**Total Participant/Trainee Support Costs**

RESEARCH & RELATED Budget {C-E} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTIONS F-K, BUDGET PERIOD 2

* ORGANIZATIONAL DUNS: 0527809180000

*** Budget Type:** ● Project ❍ Subaward/Consortium

Enter name of Organization: UNIVERSITY OF MIAMI SCHOOL OF MEDICINE

*** Start Date:** 09-30-2010 *** End Date:** 09-29-2011 **Budget Period: 2**

F. Other Direct Costs	**Funds Requested ($)**
Total Other Direct Costs	

G. Direct Costs	**Funds Requested ($)**
Total Direct Costs (A thru F)	**301,351.00**

H. Indirect Costs

Indirect Cost Type	**Indirect Cost Rate (%)**	**Indirect Cost Base ($)**	*** Funds Requested ($)**
1. MEDICAL ON CAMPUS-RESEARCH	53	291,351.00	154,416.00
		Total Indirect Costs	**154,416.00**

Cognizant Federal Agency DHHS, Darryl Mayes, 202-401-0215
(Agency Name, POC Name, and POC Phone Number)

I. Total Direct and Indirect Costs	**Funds Requested ($)**
Total Direct and Indirect Institutional Costs (G + H)	**455,767.00**

J. Fee	**Funds Requested ($)**

K. * Budget Justification	File Name: budget_justification_per1.pdf (Only attach one file.)	Mime Type: MIMETYPE

RESEARCH & RELATED Budget {F-K} (Funds Requested)

RESEARCH & RELATED BUDGET - Cumulative Budget

	Totals ($)	
Section A, Senior/Key Person		**110,805.00**
Section B, Other Personnel		**463,410.00**
Total Number Other Personnel	8	
Total Salary, Wages and Fringe Benefits (A+B)		**574,215.00**
Section C, Equipment		**70,000.00**
Section D, Travel		
1. Domestic		
2. Foreign		
Section E, Participant/Trainee Support Costs		
1. Tuition/Fees/Health Insurance		
2. Stipends		
3. Travel		
4. Subsistence		
5. Other		
6. Number of Participants/Trainees		
Section F, Other Direct Costs		
1. Materials and Supplies		
2. Publication Costs		
3. Consultant Services		
4. ADP/Computer Services		
5. Subawards/Consortium/Contractual Costs		
6. Equipment or Facility Rental/User Fees		
7. Alterations and Renovations		
8. Other 1		
9. Other 2		
10. Other 3		
Section G, Direct Costs (A thru F)		**644,215.00**
Section H, Indirect Costs		**304,334.00**
Section I, Total Direct and Indirect Costs (G + H)		**948,549.00**
Section J, Fee		

BUDGET JUSTIFICATION

Sawsan Khuri, Ph.D. Principal Investigator (2.4 calendar months)
Dr. Khuri is the Bioinformatics Lead Scientist at the Center for Computational Science, and Assistant Research Professor, Dr. John T. Macdonald Foundation Department of Human Genetics, University of Miami Miller School of Medicine. Dr. Khuri has 10 years experience in bioinformatics, covering a wide spectrum of areas including the regulation of gene expression, predicting the functional relevance of rare or common variants, and more recently, next generation sequence (NGS) data assembly and visualization. Dr. Khuri will train and work directly with the Bioinformatics postdoc to develop the analysis strategies that will be required for all projects. In particular, she will develop strategies for gene expression analysis and rare variant characterization and provide bioinformatics support for NGS data and the other bioinformatics needs of the program as they arise. 20%FTE is requested for all years.

Nicholas Tsinoremas, Ph.D. (Co-Investigator) (1.2 calendar months)
Dr. Tsinoremas is the Director of the Center for Computational Science and a Research Professor with the Departments of Medicine and Computer Science. Dr. Tsinoremas will provide direction and supervision of the overall project. He will review and specify the scientific pipelines and goals of the project. He will provide advice and guidance with the design of the system and Graphical User Interface.

Stuart Young, Ph.D. Post Doctoral Associate (12 calendar months)
Dr. Young has 4 years experience in bioinformatics research, covering deleterious mutation prediction in proteins, the design and implementation of bioinformatics database interfaces, and sequence annotation using machine learning and other methods. In particular, he will develop strategies for optimizing short read assembly, expression analysis, gene network and variant characterization workflows, and develop the system software. Dr. Young will handle the day-to-day project management and the other bioinformatics needs of the program as they arise. 100% FTE is requested for all years.

Chris Mader, Software Advisor (1.2 calendar months)
Mr. Mader is the Director of Application Systems Development at the Center for Computational Science. Mr. Mader will supervise software engineering for the project and provide guidance and advice on systems design and code.

Deepthi Puram, Software engineer (6 calendar months)
Deepthi Puram is Sr. Project Manager with the Center for Computational Science. Ms. Puram will develop object-oriented, modular code for the project, along with test suites and developer/user documentation. She will also assist in the iterative modification of the software design and deployment of the system.

TBA, Software engineer (12 calendar months)
The software engineer will develop object-oriented, modular code for the project, along with test suites and developer/user documentation. He/she will also assist in the iterative modification of the software design and deployment of the system.

EQUIPMENT
The budget includes a budget of $70,000 over two years to cover the purchase of a 16-node cluster, multi-gigabit routing switch, production web server, production database server, development web server, and a development database server.

PHS 398 Cover Page Supplement

OMB Number: 0925-0001
Expiration Date: 9/30/2007

1. Project Director / Principal Investigator (PD/PI)

Prefix:
* First Name: SAWSAN
Middle Name:
* Last Name: KHURI
Suffix:

* New Investigator? ○No ●Yes

Degrees: PhD

2. Human Subjects

Clinical Trial? ●No ○Yes

* Agency-Defined Phase III Clinical Trial? ●No ○Yes

3. Applicant Organization Contact

Person to be contacted on matters involving this application

Prefix: Mr.
* First Name: Tom
Middle Name:
* Last Name: Gill
Suffix:

* Phone Number: 305-243-6232
Fax Number: 305-243-4611
Email: grantsgov@med.miami.edu

* Title: Director

* Street1: 1400 NW 10th Avenue
Street2: Dominion Tower, 10th Floor, Suite 1007
* City: Miami
County: Miami-Dade
* State: FL: Florida
Province:

* Country: USA: UNITED STATES
* Zip / Postal Code: 33136

Tracking Number:

PHS 398 Cover Page Supplement

OMB Number: 0925-0001
Expiration Date: 9/30/2007

4. Human Embryonic Stem Cells

* Does the proposed project involve 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/registry/index.asp . Or, if a specific stem cell line cannot be referenced at this time, please check the box indicating that one from the registry will be used:

Cell Line(s): Specific stem cell line cannot be referenced at this time. One from the registry will be used.

Tracking Number:

OMB Number: 0925-0001
Expiration Date: 9/30/2007

PHS 398 Research Plan

1. Application Type:

From SF 424 (R&R) Cover Page and PHS398 Checklist. The responses provided on these pages, regarding the type of application being submitted, are repeated for your reference, as you attach the appropriate sections of the research plan.

*Type of Application:

● New ○ Resubmission ○ Renewal ○ Continuation ○ Revision

2. Research Plan Attachments:

Please attach applicable sections of the research plan, below.

1. Introduction to Application (for RESUBMISSION or REVISION only)	
2. Specific Aims	rplan_nar.pdf
3. Background and Significance	
4. Preliminary Studies / Progress Report	
5. Research Design and Methods	rplan_rdm.pdf
6. Inclusion Enrollment Report	
7. Progress Report Publication List	

Human Subjects Sections

Attachments 8-11 apply only when you have answered "yes" to the question "are human subjects involved" on the R&R Other Project Information Form. In this case, attachments 8-11 may be required, and you are encouraged to consult the Application guide instructions and/or the specific Funding Opportunity Announcement to determine which sections must be submitted with this application.

8. Protection of Human Subjects	
9. Inclusion of Women and Minorities	
10. Targeted/Planned Enrollment Table	
11. Inclusion of Children	

Other Research Plan Sections

12. Vertebrate Animals	
13. Select Agent Research	
14. Multiple PI Leadership Plan	
15. Consortium/Contractual Arrangements	
16. Letters of Support	rplan_con.pdf
17. Resource Sharing Plan(s)	rplan_res.pdf

18. Appendix

Attachments

IntroductionToApplication_attDataGroup0

File Name	Mime Type

SpecificAims_attDataGroup0

File Name	Mime Type
rplan_nar.pdf	MIMETYPE

BackgroundSignificance_attDataGroup0

File Name	Mime Type

ProgressReport_attDataGroup0

File Name	Mime Type

ResearchDesignMethods_attDataGroup0

File Name	Mime Type
rplan_rdm.pdf	MIMETYPE

InclusionEnrollmentReport_attDataGroup0

File Name	Mime Type

ProgressReportPublicationList_attDataGroup0

File Name	Mime Type

ProtectionOfHumanSubjects_attDataGroup0

File Name	Mime Type

InclusionOfWomenAndMinorities_attDataGroup0

File Name	Mime Type

TargetedPlannedEnrollmentTable_attDataGroup0

File Name	Mime Type

InclusionOfChildren_attDataGroup0

File Name	Mime Type

VertebrateAnimals_attDataGroup0

File Name	Mime Type

SelectAgentResearch_attDataGroup0

File Name	Mime Type

MultiplePILeadershipPlan_attDataGroup0

File Name	Mime Type

ConsortiumContractualArrangements_attDataGroup0

File Name	Mime Type

LettersOfSupport_attDataGroup0

File Name	Mime Type
rplan_con.pdf	MIMETYPE

ResourceSharingPlans_attDataGroup0

File Name	Mime Type
rplan_res.pdf	MIMETYPE

Appendix

File Name	Mime Type

SPECIFIC AIMS

NextGen sequencing technologies are fast approaching the '$1,000 genome' target (1): a $5,000 genome will be available in May 2009 by Comparative Genomics while other NextGen industry players are rapidly reducing the cost per Mbase. A new paradigm is emerging of the correlated and rapid analysis of individual genomic variation, methylation, histone-binding, expression analysis and other genome-wide factors that may begin to unlock the secrets of the cell (2) and create new avenues for clinical diagnostics. Bioinformatics infrastructure – hardware, software and personnel – is the bottleneck in the development of this new paradigm (2, 3). Costly investments are required in high performance computing clusters to cope with the large data volumes and in skilled personnel to develop, evaluate and run bioinformatics tools, and to integrate diverse biological data sources. Most biomedical research and diagnostics labs are unable to provide even the minimum of these hardware and personnel requirements. With regard to software, workflow tools are essential to allow non-technical staff to automate and run well-defined but complex analysis processes. These tools must be web-enabled for ease of access and flexible enough to support exploratory analysis through interaction with the data using a wide range of different software applications and data processing steps. They should also provide visualization functionality capable of handling large volumes of NextGen data and integrating heterogeneous external genome feature data sets. Given the budget considerations mentioned above, the ideal workflow tool should also be open source and freely available to the academic community.

To help address these opportunities, we propose the rapid deployment of a software system and analysis tools for managing NextGen sequencing projects, from short read generation to bioinformatics analysis to data visualization. The system will meet the following challenges: 1) facilitating the analysis of large-scale sequencing studies, 2) enabling expression analyses, and 3) determining the relationship of sequence variation and phenotypes to disease. These challenges will be addressed through the following specific aims:

Specific Aim 1: Develop and implement an optimized NextGen assembly workflow

We first propose carrying out an objective and thorough evaluation of current NextGen assemblers/aligners. Based on this assessment, we will provide an optimized workflow for each of the three main NextGen sequence platforms (Illumina/Solexa, Roche/454 and ABI/SOLiD) to generate assemblies and their associated quality control information. These workflows will be customizable by the user to suit their particular desired quality metrics or tradeoffs.

Specific Aim 2: Develop and implement NextGen genomic variation and expression analysis workflows

We propose developing a genomic variation annotation pipeline with defined quality control/assurance algorithms for verifying and annotating SNPs (single nucleotide polymorphisms), CNV (copy number variation) and large-scale structural variation. The pipeline will be integrated with current expression analysis packages. We also propose to develop new expression analysis algorithms. To facilitate better reporting and visualization of results, data filters will be designed based on user requirements to extract result subsets and provide genome-level views of the results integrated with external genomic features and exportable to downstream analysis applications.

Specific Aim 3: Develop an integrated NextGen workflow tool and genome viewer

Based on the requirements in Aims 1 and 2, we propose the development and implementation of a novel tool providing end-to-end integrated NextGen data analysis workflows, reporting and real-time genomic visualization of huge data sets. The tool, named Aqwa (Automated Query and Workflow Agent), will provide pre-optimized workflows for assembly/alignment, genomic variation and expression analysis and will also allow users to create their own customized workflows using any Linux-platform bioinformatics tools. The software development process will implement a user-centric approach including extensive pre- and post-release user testing at each project milestone to ensure improved usability compared to currently available tools.

1. NHGRI. NHGRI Seeks DNA Sequencing Technologies Fit for Routine Laboratory and Medical Use. 2008 [updated 2008; cited]; Available from: http://www.genome.gov/27527585.
2. Mardis ER. Next-Generation DNA Sequencing Methods. Annual Review of Genomics and Human Genetics. 2008;9(1):387-402.
3. Schuster SC. Next-generation sequencing transforms today's biology. Nat Meth. 2008;5(1):16-8.

Integrated NextGen workflow tool and genomic viewer

Research Area

This application addresses broad Challenge Area **(**06) Enabling Technologies and Challenge Topic 06-HG-101: New computational and statistical methods for the analysis of large data sets from next-generation sequencing technologies.

The Challenge and Potential Impact

The principal stakeholders in our study are large sequencing centers, genomics research groups, funders of genomics research, clinical diagnostics laboratories and a significant number of consumers of health care services. Genomics plays a part in nine of the Ten Leading Causes of Death in the United States (http://www.cdc.gov/genomics/faq.htm). Genomics funding by government and nonprofit organizations averaged $2.9 billion a year from 2003-2006, of which the United States accounted for 35%, half of which was provided by the NIH (National Institutes of Health) (1). Rapidly falling NextGen sequencing costs mean that, even given a reduction in funding due to the current global economic downturn, the demand for NextGen analysis services will continue to rise and the use of NextGen technologies will continue to spread into other research communities. As these technologies develop, they will present greater data infrastructure demands and new bioinformatics challenges. **The increased adoption of NextGen sequencing among genomics research groups and clinical diagnostics labs must be accompanied by advances in high-throughput analysis, strategies for mitigating systemic bias, and new sequence assembly and downstream analysis pipelines.**

This section describes the particular challenges of NextGen technologies, applications and bioinformatics in more detail and discusses the anticipated impact of solutions provided by this study.

Specific Aim 1: Develop and implement an optimized NextGen assembly workflow

In order to appreciate the particular problems and challenges of NextGen sequence analysis, we must first grasp the particular strengths and weaknesses of the different sequencing technologies. The current mainstream NextGen platforms produce millions of short (50bp – 400bp) sequence reads. Each of the three main platforms, namely Illumina/Solexa, Roche/454 and ABI/SOLiD, have their own inherent problems, including significant sequencing error rates and systematic errors. Large sequencing organizations such as genome centers, academic core facilities and commercial contract-sequencing enterprises across the globe have already adopted this NextGen technology (Figure 1) and smaller labs and molecular diagnostics facilities participating in growing numbers. A common refrain among adaptors of this technology is that the downstream bioinformatics analyses are often poorly understood and underestimated.

Alongside the rollout of NextGen sequencing platforms, third generation sequencing technologies are being developed to sequence single DNA molecules faster and cheaper with streamlined sample preparation. Real-time sequencing by synthesis is being developed by VisiGen (http://www.visigenbio.com) and Pacific Biosciences (http://www.pacificbiosciences.com). Pacific Biosciences is due to launch commercially in 2010 and has a mean DNA synthesis rate of approximately 4 bases per second, with a maximum read length of 4,000 bp. Also in development is sequencing based on sensing the bases of DNA molecules passed through nanopores (~5 nm in diameter). Different methods are being tested to create nanopores, including inorganic membranes (solid-state nanopores), genetically engineered protein channels by Oxford Nanopore Technologies (http://www.nanoporetech.com), polymer-based nanofluidic channels, and a combination of nanopores with sequencing by hybridization by NABsys (http://www.nabsys.com).

The currently available bioinformatics tools fall into four categories: reference aligners, de novo assemblers, variant-discovery tools and alignment viewers. Among the reference aligners are Eland (GAPipeline v0.30, Illumina), Mira, Genomics Workbench (CLC Bio), Seqman NGen (DNAStar), NextGene (Soft Genetics), MAQ and Shrimp. De novo assemblers include Edina, EULER-SR, SHARCGS, SSAKE, Velvet, and SOAPdenovo. Some NextGen statistical data-analysis tools are also available, such as JMP Genomics. **Despite the growing number of NextGen assembly/alignment tools, obtaining an accurately assembled sequence contig is still a very challenging problem.** The current tools vary widely in terms of data volume capacity (e.g., bacterial versus human data sets), number of reads aligned/assembled, error rates and bias, all of which may lead to suboptimal assemblies. Moreover, little is known about the comparative performance of the available tools because the few available performance statistics are based mainly on results from different non-human

data sets (e.g., phage, bacteria, yeast). Therefore, achieving an unbiased comparison between different assemblers is difficult even before considering how they perform on human sequence data.

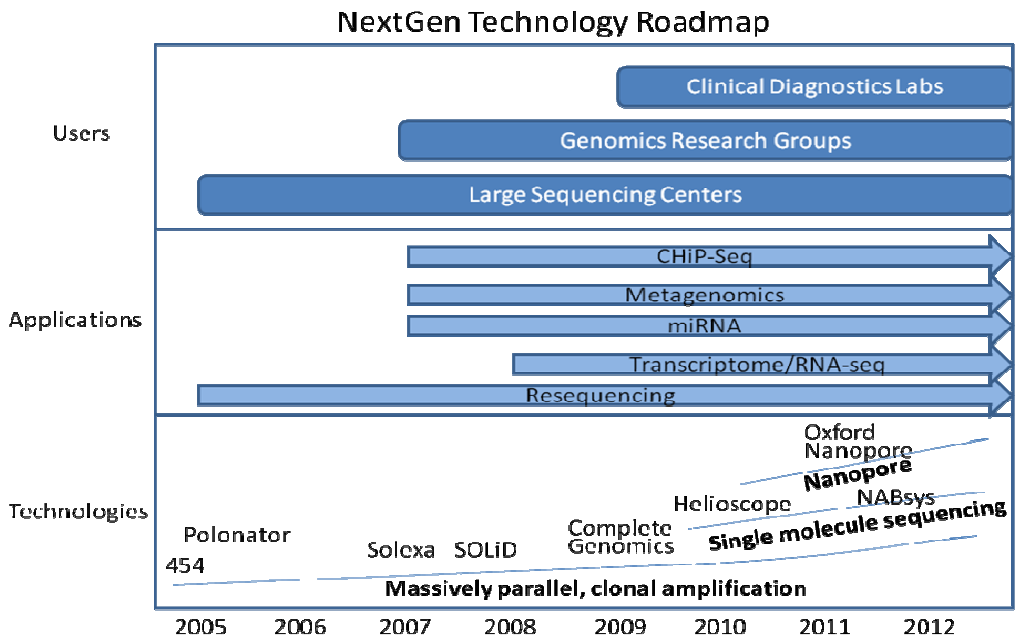

Figure 1. NextGen sequencing technology roadmap: the spread of uptake of NextGen sequencing and growth of applications parallels the development of new sequencing technologies. Data outputs are projected to increase rapidly as sequencing costs fall due to the rollout of new technologies.

Particularly in the early phases of the development of NextGen technology while many competing algorithms vie for supremacy, scientific publications will require comparisons of results using several different sequence analysis algorithms. **An objective, performance-based comparison of NextGen bioinformatics tools is an important step towards mitigating systemic bias in next generation data analysis. It will also lower the entry barrier allowing biomedical researchers to answer more penetrating questions and in less time.** An optimized NextGen assembly workflow will allow users to improve their work efficiency and the quality of their results. Customizable pipelines will also help meet the bioinformatics challenges faced by researchers at the cutting edge of life science exploration. Dynamic pipeline configuration coupled with high performance computing will enable researchers and other end users to rapidly develop and adapt different approaches to solving particular problems.

Another anticipated outcome will be the hastening of the transition to a mature technology, with fewer bioinformatics applications used for a wide range of applications. Extensive and objective comparisons between assembly/alignment tools will also serve to inform biomedical users which tools are more suitable for particular applications and data sets. A platform that is flexible enough to allow routine comparisons between the outputs of emerging algorithms for as-yet unknown NextGen bioinformatics challenges will be a valuable resource for research groups using genomics technologies and diagnostics labs alike.

Specific Aim 2: Develop and implement NextGen genomic variation and expression analysis workflows

Sequencing only the cDNA or transcribed portion of the genome focuses the analysis by reducing the size of the sequencing target space and also reduces costs. Expression profiling (a.k.a. RNA-seq or transcriptome analysis) is an increasingly popular application of NextGen sequencing (2) that has been shown to be robust and sensitive in comparison to five microarray platforms. NextGen sequencing also identified antisense transcription, which microarrays cannot detect, in 51% of all genes. In yeast, NextGen expression analysis has demonstrated a larger, more complex transcriptome than had been expected (3). An estimated 74.5% of the non-repetitive yeast genome was shown to be transcribed, as were many overlapping genes, alternative initiation codons and upstream open reading frames of yeast genes were demonstrated using short reads to generate a high-resolution map of the genome. Transcriptomes for mouse brain, liver and skeletal muscle were mapped by NextGen deep sequencing (4), providing a digital measure of the presence and prevalence of transcripts from known and previously unknown genes. RNA standards were used to quantify transcript prevalence and to test the linear range of transcript detection, which spanned five orders of magnitude.

Alongside the profound impact of NextGen applications in basic research, NextGen sequencing is also being adopted by clinical diagnostics laboratories for applications requiring deep sequence coverage and high-sensitivity such as rare HIV drug resistant variant detection (5). As the focus in human genetics research has shifted to genome wide association studies for the identification of genes involved with complex, multi-gene diseases, there is an increasing need for comprehensive diagnostic evaluations of SNPs as common or rare variants in multiple genes**.** At the base-pair level, NextGen sequencing has been shown to be highly suitable for high-throughput SNP acquisition using novel algorithms such as PolyBayes (6).

At the level of large-scale genomic variation, somatically acquired genomic rearrangements have been implicated in the development of various cancer phenotypes. The feasibility of using NextGen sequencing for the systematic, genome-wide characterization of rearrangements in human cancer genomes has already been demonstrated. The first high-resolution map of human genome structural variation revealed complex and large-scale structural variation in the form of insertions, deletions and inversions from a few thousand to millions of base pairs in length. NextGen sequencing has been used to characterize 306 germline structural variants and 103 somatic rearrangements to the base-pair level of resolution (7). **Improvements in the discovery of large scale genomic variations using NextGen sequencing will likely have a deep impact on the study of their involvement in cancer and other complex diseases.**

Copy number variations are another form of important large-scale variation: CNVs of 100 kilobases and greater contribute substantially to genomic variation between normal humans (8, 9) however their contribution to human phenotype remains unclear, and they are difficult to measure. Microarray-based approaches for detecting CNVs depend on microarray signal intensity differences to predict regions of variation and cannot detect inversions. Before the advent of CNV prediction based on NextGen sequencing, only a small fraction of CNV base pairs had been determined at the sequence level. NextGen CNV mapping allows the discovery of cancer-causing genes in genomic regions that show recurrent copy-number alterations (gains and losses) in tumor genomes (10). **Advances in CNV prediction and characterization can be expected to have a far-reaching impact in the study of their involvement not only in the characterization of different disease phenotypes, but also in the context of human phenotypic variation.**

Developing high-throughput approaches for the analysis of the mechanisms whereby genomic variation can cause disease is a major genomics research challenge. To the extent that gene expression is a proxy of disease phenotype, NextGen sequencing can be used provide evidence of the relationship of sequence variation and phenotypes to disease. One approach is to use transcriptome QTL (quantitative trait locus) mapping to identify chromosomal regions containing sequence variants that cause variations in downstream expression. However, one issue with quantitative trait mapping is that of determining an appropriate threshold value for declaring significant QTL effects. Another approach is to target SNPs of interest in particular individuals early in the design stage of a study or to target specific pathways when analyzing genome-wide data but this has the drawback of selection bias. SNPs and CNVs have been associated with gene expression, but in an uncorrelated way. In a study of individuals in the International HapMap project, SNPs captured 83.6% of the total detected variation in expression levels of 14,925 transcripts and CNVs captured 17.7%, although the signals from the two types of variation had little overlap (11). Non-parametric machine learning techniques have been applied to narrow down groups of SNPs that best capture phenotypic variation using information metrics to select SNPs (12) and greedy screening (13). In yet another approach, allelic imbalance in gene expression has been found in 20–50% of genes tested in human brain, liver and kidney samples with no obvious reason evident from the SNP itself (14) and a pipeline for screening allele-specific expression has been proposed (15). These and other new approaches and tools for characterizing the relationship between genomic variation and phenotype or disease will improve our understanding of diseases and may lead to ways to treat and prevent them. **Bioinformatics fills the space between NextGen sequencing technology and the interpretation of the data by the medical community to ultimately make an impact on human health.**

		Genome viewers and annotators						Workflow tools					Aqwa		
		UCSC Genome Browser	Ensemble Genome Browser	Generic Genome Browser	Integrative Genome Viewer	Apollo	Jbrowse	Pise (Gpipe)	Galaxy	Taverna	Wildfire	CBJ	Aqwa Alpha	Aqwa Beta	Aqwa Gamma
Genome Viewer Functions	Nucleotide level view	√	√	√	X	√	√	X	X	X	X	X	√	√	√
	Abundant genomic features	√	√	X	X	X	X	X	X	X	X	X	√	√	√
	Interactive display	√	√	√	√	√	√	X	X	X	X	X	√	√	√
	Rich context menu	X	X	X	X	√	X	X	X	X	X	X	√	√	√
	Extensible genomic features	√	√	√	X	√	√	X	X	X	X	X	√	√	√
	Multiple feature views	√	√	√	√	√	√	X	X	X	X	X	√	√	√
	Aggregate high level view	X	X	X	X	X	√	X	X	X	X	X	√	√	√
	Fast view update	X	X	X	√	X	√	√	X	X	X	X	√	√	√
	Filter view based on data	X	X	X	√	X	X	X	X	X	X	X	√	√	√
Workflow Tool Functions	Maintain state (action history)	X	X	X	X	X	X	X	√	√	√	√	√	√	√
	Standard workflows	X	X	X	X	X	X	√	√	√	√	√	√	√	√
	Customisable workflows	X	X	X	X	X	X	X	√	√	√	X	√	√	√
	Drag and drop workflows	X	X	X	X	X	X	X	√	X	√	√	√	√	√
	Cluster execution	X	X	X	X	X	X	X	√	√	√	√	√	√	√
	Grid execution	X	X	X	X	X	X	X	X	√	X	X	X	X	√
	Loops, conditional branching	X	X	X	X	X	X	X	X	√	√	√	X	√	√
	Data management	X	X	X	X	X	X	X	X	X	X	X	√	√	√
	Workflow search	X	X	X	X	X	X	X	X	X	X	√	√	√	√
	Web interface	X	X	X	X	X	X	X	X	X	X	X	X	√	√
	Web service	X	X	X	X	X	X	X	X	√	√	X	X	√	√
	Plug 'n play applications	X	X	X	X	X	X	X	X	X	X	√	√	√	√
	Annotation	X	X	X	X	X	X	X	X	X	X	√	√	√	√
	Group access and sharing	X	X	X	X	X	X	X	X	X	X	√	√	√	√

Figure 2. An overview of workflow and viewer functionalities of currently available workflow tools and genome viewers shows that no existing tool provides all of the required functionality for an end-to-end solution.

Specific Aim 3: Development of a NextGen workflow and visualization tool

As bioscience becomes increasingly a quantitative analysis activity, workflow tools enable users to accomplish two main tasks: 1) automating well-defined, repetitive processes and 2) exploring data with ad-hoc analyses (16). **The few studies on common bioinformatics tasks and usability of bioinformatics tools (17) identified an urgent need among bioscience researchers for workflow-based tools.** There are over 200 major Internet biological data sources however these sites are mostly simple GUIs (Graphical User Interfaces) with limited data compatibility between them despite the fact that researchers often need to combine the outputs of multiple sites to generate bioinformatics analyses. **These resources may be underexploited if users feel too much time is used navigating the sites, selecting among appropriate sources, downloading and uploading files. Workflow tools can solve this problem by generating large and complex systems from collections of programs, data sources and even structured data services. However, the majority of bioinformatics workflow tools only partially realize the potential of the available data and application resources (Figure 2).** With notable exceptions (18), there has been limited progress in connecting different sites with the client as the intermediary. Another approach is to use a central site as a service directory lookup such as BioMOBY and TAVERNA with the limitation that service discovery relies upon the accurate and uniform description of biological data types and relations between them, for which there is no commonly-accepted ontology (19-21) or language, despite some developments (22).

Data visualization and interpretation become paramount as the bioinformatics challenge shifts from mastering the basic tools to gaining biological insights from huge amounts of data. Three commercial software packages by DNAStar, SoftGenetics and CLC Bio provide data viewers that allow the user to see read alignments, coverage depth, genome annotations, and variant analysis. However, they lack the capability for viewing data sets as large as a whole human chromosome and show poor performance even on sub-chromosome data sets. The three major publicly available genome viewers, UCSC Genome Browser, Ensembl genome browser and GBrowse, are based on the traditional client-server model where the user's requested data is reloaded as an

image file delivered from the server. Java-based applications such as Apollo are more interactive but lack a concerted approach to data sharing. Most do not allow the user to filter the displayed data set based on biological criteria, although some newer applications such as IGV (http://www.broad.mit.edu/igv/) allow for limited filtering of the displayed features. Figure 2 lists the capabilities that a fully functional workflow tool and genome viewer must possess and shows the gaps in functionality of the currently available tools. We propose to provide all of this functionality in an integrated workflow tool and genome viewer.

The Approach

We propose to address the abovementioned bioinformatics challenges, namely the development of 1) improved NextGen sequence assembly workflows, 2) optimized genomic variation and expression workflows, and 3) a NextGen workflow and visualization tool with the following approaches:

Specific Aim 1: Develop and implement an optimized NextGen assembly workflow

We propose carrying out an objective evaluation of current NextGen assemblers/aligners using artificial data sets based on human biological samples in which each read's position is known *a priori* in order to accurately compare results between different algorithms. Performance criteria will be established before testing based on the particular difficulties of assembling short reads derived from human genomic material. One or more optimized assembly workflows maximizing the performance criteria will be created as push-button tools to generate assemblies and associated quality control information. These workflows will also be customizable by the biologist/researcher to suit particular desired quality metrics or to meet any necessary tradeoffs between different quality metrics.

A combination of different assemblies may provide more reliable estimates of genetic aberrations by flagging dubious assembly regions that are not represented in a majority of the different assemblies. Conversely, regions that are matched identically by a majority of the different algorithms might be accorded greater confidence with regard to their predicted SNPs, indels and breakpoints. We will attempt to prove or disprove these hypotheses in the second step of specific aim 1 by using artificial reads generated from approximately 100 human sequence samples selected from the SRA (Short Read Archive http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi). **A comparison study of this scale has never before been attempted.** Samples from individuals with sequences from all three NextGen sequencing platforms will be selected based on coverage depth, the presence of family member samples and the availability of additional genomic analyses such as microarrays in order to detect anomalies and discriminate between differing sequence assembly results.

Specific Aim 2: Develop and implement NextGen genomic variation and expression analysis workflows

To meet the need for comprehensive diagnostic evaluations of genomic variation in multiple genes, we propose the development of discovery and annotation workflows for three kinds of variation: SNPs, CNVs and large-scale genomic variation (chromosomal indels, inversions and dislocations). We will conduct three consecutive studies using biological samples from several human diseases to determine optimal methods for the prediction, filtering and verification of these different genomic variations. **The novelty of our approach lies in the fact that these workflows will include defined quality control/assurance algorithms, and the use of pedigree information and statistical techniques for predicting heterozygote/homozygote calls.** We will develop a methodology for estimating erroneous SNP calls and predicting homozygote and heterozygote genotypes using Roche/454 reads derived from Nimblegen exome capture samples (Figure 3). The initial experimental subjects are an extended family of 8 individuals, including 4 with a heritable neurodegenerative disorder. Initial SNP filtering and annotation will be carried out using read quality, read coverage, presence of read coverage in other sample members, presence in dbSNP, HapMap and the Venter and Watson genomes. We will validate our approach with a sample of 1,000 individuals on a smaller subset of captured genomic regions. **This analysis is being carried out in partnership with the MIHG (Miami Institute for Human Genomics). Further joint studies with the MIHG are anticipated involving the analysis of SNPs, large-scale variation and copy number variation in metabolic disorders, cancer, and cardiovascular disease.**

We propose enabling the integration into the variation analysis workflows of existing algorithms for determining the relationship between genomic variation and disease. These include GWAS studies and combinations of SNP, CNV and gene expression data with disease phenotypes in different human tissues and for different disease states. Many of these methods have not yet been applied to NextGen data and comparing the results of several methods will allow us to confirm associations between cis- or trans-acting genomic variations, expression phenotypes and disease.

In addition to the variation workflows, we propose the development of an optimized workflow for expression analysis. **This workflow will be integrated with current expression analysis packages such as ERANGE (4) and will include new expression analysis algorithms to improve quantization of transcript counts (Figure 3).** One approach that will be evaluated for inclusion in the workflow is the use of 'standard gene sets' or other transcribed regions with stable copy numbers across tissue types that can be used to calibrate the relative expression levels of genes and estimate absolute copy numbers per cell between different tissue samples.

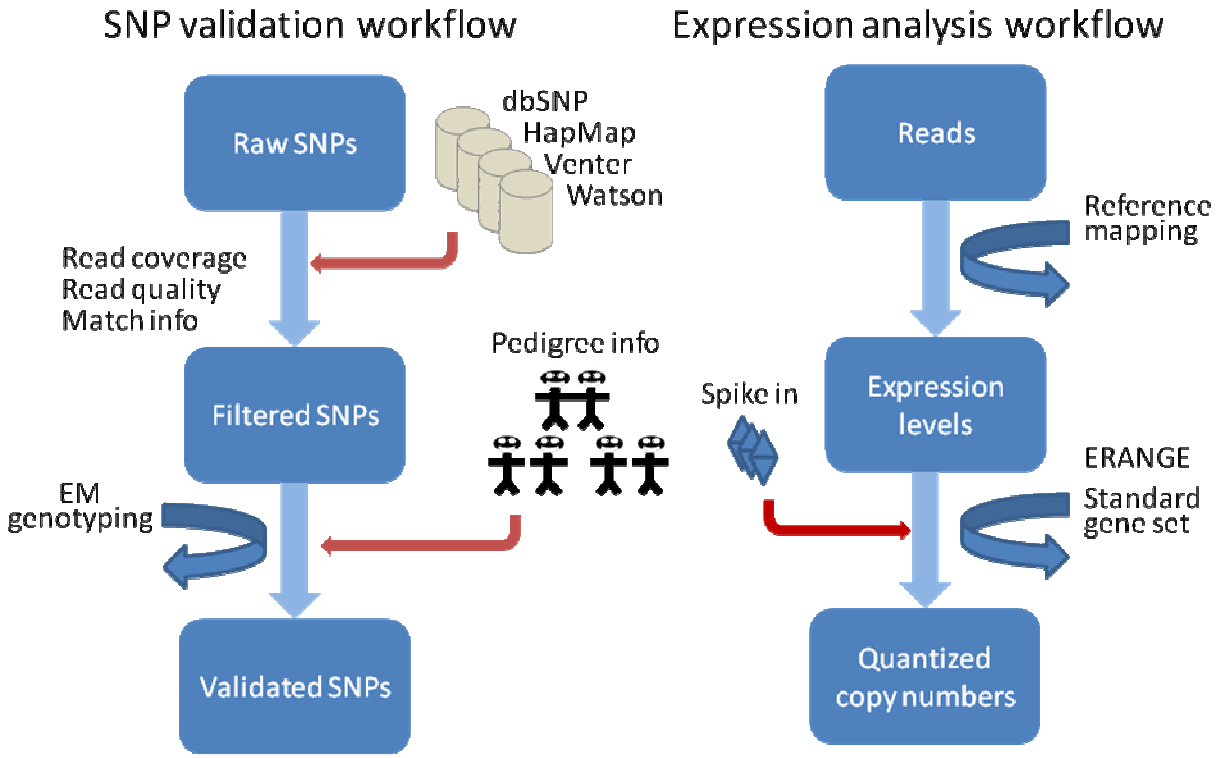

Figure 3. Prototype SNP validation and expression analysis workflows.

Another unique feature of our approach is that, as part of the reporting and visualization of results, data filters will be designed based on user requirements to extract result subsets and provide genome-level views of the results integrated with external genomic features. Results will also be exportable to downstream analysis applications (Cytoscape, Genespring, R, etc.).

Specific Aim 3: Develop a NextGen workflow and visualization tool

Based on the requirements in Aims 1 and 2, we propose the implementation of a novel bioinformatics tool providing end-to-end integrated NextGen analysis workflows, reporting and real-time visualization of huge genomic data sets. The tool, named Aqwa (Automated Query and Workflow Agent), will provide pre-optimized workflows for NextGen assembly, genomic variation and expression analysis, and will also allow users to create their own customized workflows. **The software development process will implement a user-centric approach including extensive user testing at each project milestone and intervening iterations. Our emphasis on user testing and user-centric development, which differs from all previous bioinformatics workflow tools, is designed to ensure that the user interface is as intuitive as possible.**

Aqwa is designed so that only a basic familiarity with web page navigation and drag and drop user interfaces is required of the user. **Unlike other existing workflow systems, Aqwa will also support the addition of new Linux-platform applications in a "plug 'n play" and scalable manner in order to flexibly adapt to changing computational needs and rapidly evolving bioinformatics challenges.** The following list of requirements and functional criteria encompasses the functionality envisaged for Aqwa.

Functional Requirements

1. Low barriers to usage
    a. Web access
    b. User-friendly, intuitive interface
    c. “Plug ‘n play” applications for rapid deployment
    d. Searchable project annotations
2. Workflows
    a. Predefined workflows (transcriptome, variation annotation, gene networks, file conversion utilities, ID conversion utilities, etc.)
    b. Customizable workflows
    c. Drag ‘n drop workflows
    d. Persistent data and workflow configurations
    e. Loops, conditional branching
3. Reports - customizable report extraction from workflow output
4. Views
    a. Customizable genomic views of report data
    b. Interactive display with rich context menu
    c. Integrated, extensible genomic features
    d. Multiple feature views – nucleotide level to aggregate high-level view
    e. Fast view update
    f. User can filter view based on data
    g. Genomic feature-level annotation by user
5. Sharing
    a. User-defined groups with customizable permissions
    b. Workflow, report and view sharing among groups
6. Input/Output and execution
    a. Import external biological data and genomic features into workflow
    b. Integration with external software (e.g., Cytoscape, R, GeneSet Analyzer)
    c. Programmatic remote access (API and Web Service)
    d. Cluster execution
    e. Grid execution
7. Maintain state (action history)
8. Data management – direct user access to input and output files

Information regarding data provenance is retained in the system to identify the source of data throughout the workflow such as the owner, author application, creation and modification dates, and content type. In addition, a log is kept of all project changes and updates. The user can also annotate the project at all levels and search these annotations. Aqwa is similar to the laboratory notebook paradigm employed by the BCJ (Bioinformatics Computational Journal) workflow tool but has a wider the range of functionality. **Aqwa’s functional requirements largely encompass those for a proposed ‘genome wiki’ (23) intended to facilitate cooperative genome annotation by a community of experts, reflecting Aqwa’s utility to the wider genomics community as an accurate, continually updated source of genome annotation.** Aqwa’s

genome viewer functionality incorporates the AJAX-enabled JBrowse genome viewer (http://jbrowse.org) to provide a fast, fluid and responsive genome browser interface (Figure 4).

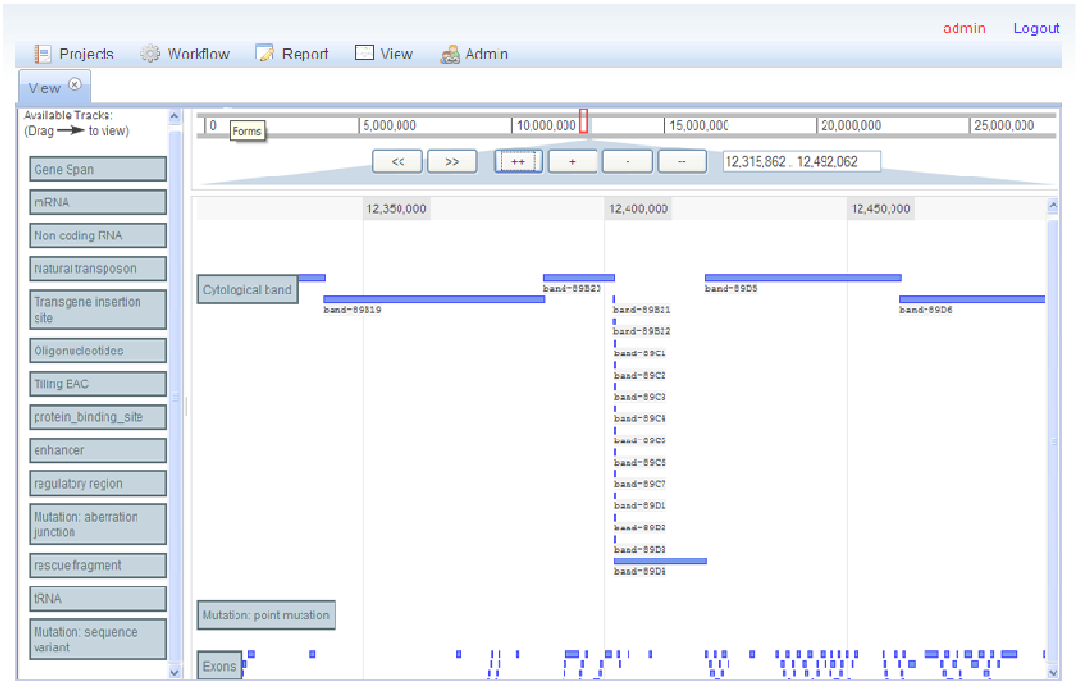

Figure 4. Aqwa's Jbrowse-based allows drag 'n drop feature selection and real-time zooming and panning.

From the user's perspective, designing customizable workflows can be a daunting prospect due to the complexity of inputs for bioinformatics applications. Help information is displayed for each application linked to the University of Miami's online bioinformatics information services portal (http://bio.ccs.miami.edu/ibis) and an automatic syntax checker ensures that application inputs are sufficient and correct. In addition, each application object contains methods that automatically derive its input arguments from the resources and outputs of preceding applications in the workflow. In the web interface, each application in a workflow is represented as a block with required input parameters and outputs (Figure 5).

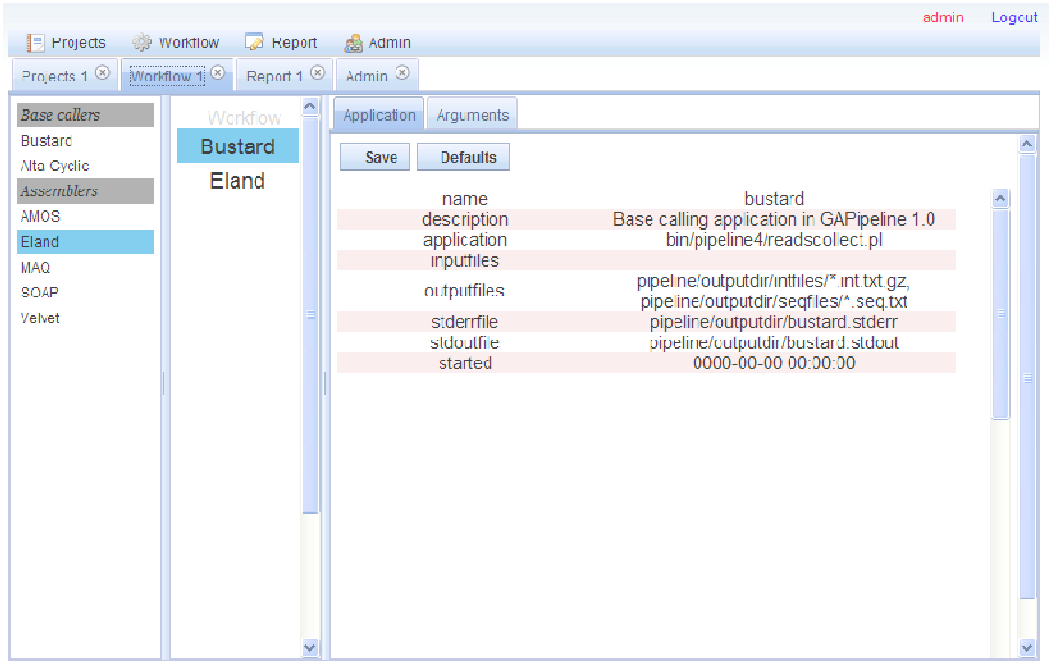

Figure 5. Workflows are created by dragging applications from the left toolbar into the center pane. Users can choose to use the default automated argument settings or manually configure applications in the right pane.

For advanced users, Aqwa provides a file manager which allows direct access and drag ‘n drop manipulation of the file system of each workflow (Figure 6), rather than solely an abstract data interface such as in the BCJ. This is an optional function, which provides the benefits of fine control of workflows for power users without the burden of increased complexity for less advanced users.

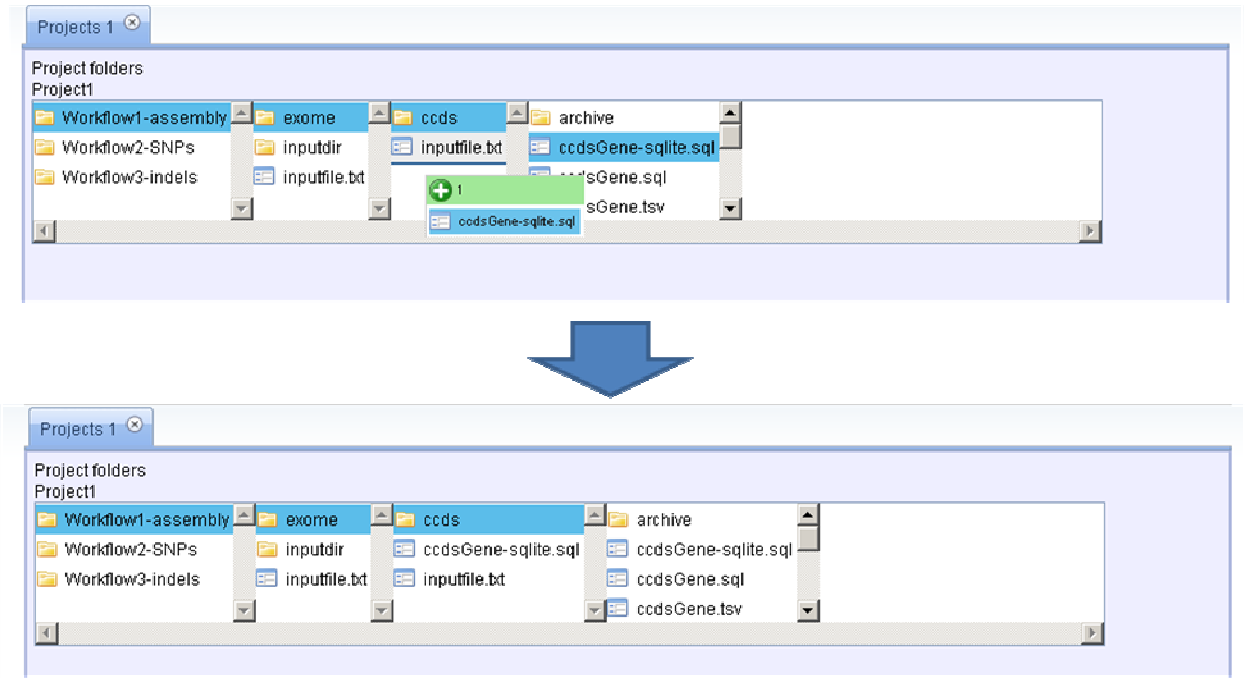

Figure 6. Aqwa’s file manager allows drag ‘n drop manipulation of workflow file systems.

The various activities involved in Aqwa’s development are detailed in the following summarized software design document.

Architectural design

Guided by the software engineering strategy of *separation of concerns* for ease of development and maintenance, the system is essentially embodied by the Model-View-Controller design pattern. The Model represents the data objects of the system (i.e., database tables on the server), the View is the display the user sees representing the Model and the Controller takes care of processing user-initiated events like mouse clicks and key presses to change the Model and the representation of the Model in the View (Figure 7). The Controller is composed of several MVC components that interact with a single Model to accomplish specific tasks such as authentication, project management and workflow. By adhering to a loose coupling between the system components, this component-based approach promotes ease of maintenance and reuse of architectural elements. Each controller component is represented at the file level by a plugin directory containing Javascript class files, templates and other resources required to perform one particular function. The core files of the Controller, found in the plugins/core directory, complete the task of loading all of the required modules, which can then load the rest of the system. This involves, for each plugin:

1. Locating the plugin descriptor files 'info.json' in the plugin directory
2. Determining whether the requirements for the plugin are already loaded
3. Loading the plugin
4. Registering the plugin and its version number

Once a plugin is loaded it may request information from the backend using asynchronous dojo.xhrGet/xhrPost requests. The additional resources included in each plugin (templates, css files, json files, images, etc.) represent the 'View' portion of the application.

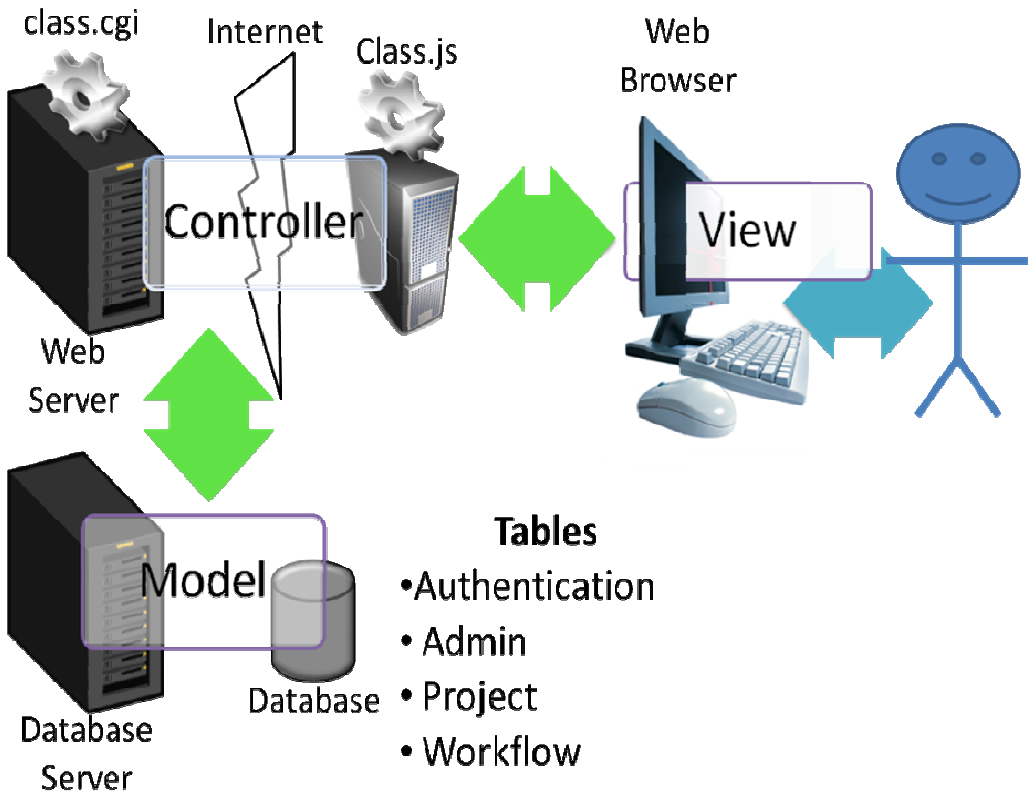

Figure 7. Breakdown of Model-View-Controller architectural design. The user interacts with the View, sending AJAX messages to the Web server via the particular MVC component (e.g., 'Workflow.js' represents the browser-side Controller component and it communicates with the server-side Controller component 'workflow.cgi'). The Controller effects any necessary transactions with the Model (i.e., changes in database tables) and changes the state of the View accordingly.

User Interface (View)

The central metaphor of the user interface is that of a project folder containing one or more workflows (customized or standard). Each workflow contains one or more applications and their input data. Each application has inputs, resources (which are similar to inputs but are not specified as command line arguments) and outputs (Figure 4). The user interface is a web application implemented in Javascript and based on the Dojo Javascript toolkit (http://www.dojotoolkit.org) which contains a rich assortment of web application components and utilities. Implementing the View as a web interface also allows for the eventual incorporation of three-dimensional graphical viewers directly inside the View for molecular dynamics simulation outputs. In addition, the user will be able to customize the View by specifying the default viewer for a particular data type. When the user elects to view a file with the specified data type, a copy of the file is downloaded to the client machine and the associated application is invoked with this input.

Data design (Model)

The database tables residing on the server and their relationships make up the Model. These include tables for user authentication, project ownership and permissions, workflow content and execution status, and data provenance. Conversion of data into a unified data model in a standard format, such as in Pegasys, is not part of the functional requirements of the system, although the system will present different export and processing options for data based on its type. Individual applications are represented by JSON, similar to the resource definition XML files commonly used in systems such as GPIPE/PISE and Pegasys. The data design is distinct from the API (Application Programming Interface), which is presented as a web service with relatively stable inputs and outputs.

Low-level design (Controller)

The Controller, or so-called 'business logic' of the application, which interacts with the View and Model is composed of Javascript classes on the client and corresponding Perl modules on the server. The choice of Perl as the backend programming language was influenced by its large user community, abundance of

bioinformatics tools such as BioPerl (http://www.bioperl.org), ease of accessibility for novice programmers and wide range of applications from text manipulation to system administration.

The system will use existing or novel syntactic structures and algebraic operators for describing bioinformatics workflows to achieve fully customizable workflows with forks, conditional statements and loops, and construct treelike workflows composed of multiple workflows linked together with logical commands. JSON (JavaScript Object Notation) is used as the data interchange format because, unlike XML, JSON-encoded objects need no additional parsing to define them at the object level, which allows them to be easily chained together in complex workflows. Following the evaluation of existing workflow control methods and their applicability to the Aqwa system, a core set of operators will be implemented in the release version of Aqwa.

The first two versions of Aqwa will support execution on the local server or on a cluster using the PBS (Portable Batch System) queue scheduler. An abstraction layer separates the workflows and the execution method for ease of extension in later versions to execution on a grid and eventually as web service requests. The latter two execution modes will require a more sophisticated pipe component and conditional operator due to the need to check for service availability before jobs can be dispatched for execution on a remote host. Failure management of jobs executed on a local server or cluster is accomplished by job monitoring scripts and wrappers to distinguish between error and normal exit modes. Failure management of jobs executed as web services or on a grid may require the development of additional tools.

The project uses an iterative/incremental development model, starting with a simple implementation of the basic functions (workflow, data management, reporting and genome view) and iteratively enhancing at each build with design modifications and new functional capabilities until the release version. Each iteration includes an examination of both the functional and quality requirements, the latter defined by user feedback and user testing (17) and interaction with other stakeholders. The alpha version will be used by a selected group of ‘power users’ within the University of Miami and the source code will be freely available for academic users. For the beta release, the user group will be expanded to all NextGen data users within the University of Miami and registered external users. Extensive user testing will be carried out at regular intervals and the results of the tests will be used to inform any additions or changes to the system’s functional requirements.

Timeline and Milestones

The project timeline and milestones are shown in Figure 8, with projected stage start and end dates in months. The project has four stages – the inception, elaboration, construction and transition stages. The inception stage (-3 to 0 months) has already been completed. This stage included identifying the stakeholders, determining the project requirements and implementation options, and conducting preliminary studies for specific aims 1 and 2. This stage also involved the preparation of this research plan and a detailed timeline.

In the elaboration stage (0 to 6 months) we will elaborate Aqwa’s project design and architecture, and develop prototype assembly, genomic variation and expression analysis workflows. The majority of the work for specific aim 1 will be done in this stage and work will begin in earnest on specific aim 2. Prototypes of the developing workflows will be implemented in the Aqwa alpha version (Milestone 1) at the midpoint of this stage. Significant risks will also be identified and risk mitigation procedures put in place and the first of six, four-month user testing cycles will begin in this stage.

During the one-year construction stage, the software design will be further refined through interaction with testers into the Aqwa beta version (Milestone 2). Specific Aim 1 will be completed at the midpoint of the construction stage and specific aim 2 will be completed at the end of the stage. Three user testing cycles will be carried out and the bulk of the software implementation work will be accomplished during this stage.

The transition stage will begin with the Aqwa gamma release (Milestone 3), which will incorporate all of the analysis workflows in developed in specific aims 1 and 2 and will be thus complete enough to transition to the user community as the release version. The goal of this stage is to ensure that the requirements have been met to the satisfaction of the stakeholders. During this stage, all remaining user and developer documentation will be completed and any defects will be identified and corrected.

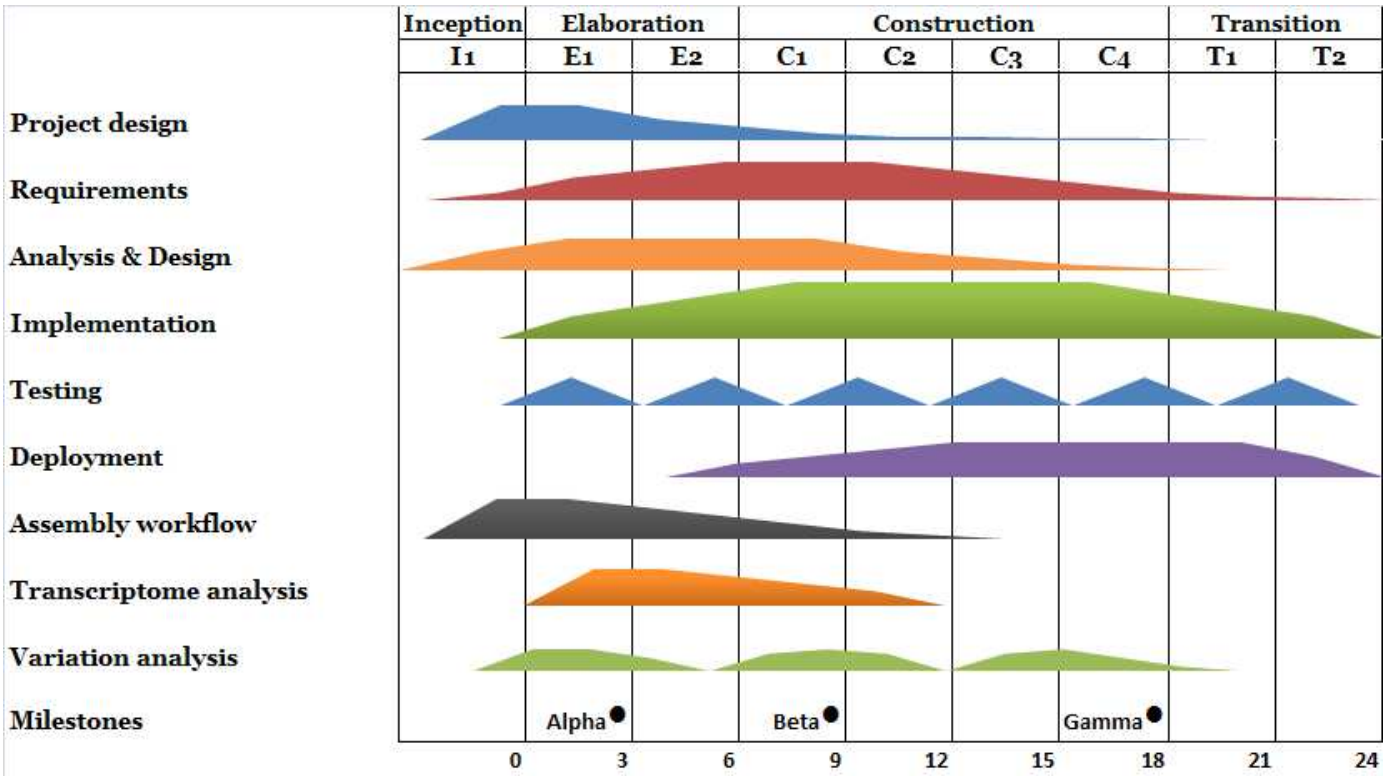

Figure 8. Project timeline and milestones by month from project start. Following a 3-month project Inception stage (currently underway), user testing will take place at regular intervals during the 6-month Elaboration, 1-year Construction and 6-month Transition stages.

As part of the inception stage, preliminary data for all three specific aims have been obtained:

Specific Aim 1: Develop and implement an optimized NextGen assembly workflow

We have started investigating the performance of the following short read assembly tools: Eland (GAPipeline v0.30, Illumina), Velvet v0.7.16, Mira v2.9.25, Genomics Workbench (CLC Bio) v1.2, Seqman NGen (DNAStar) 1.1, NextGene (Soft Genetics) 1.0 and MAQ v 0.6.8. The input data sets are: human mtDNA, human whole-genome mRNA, E. coli DNA, Herpes simplex and bacteriophage PhiX. The assemblers are being assessed for assembly speed and capacity in terms of the maximum number of reads that can be effectively assembled using relatively high-end computer hardware. We are comparing the assemblies produced by the different programs and determining a consensus based on read identity and divergence from the relevant reference sequence. Thus far in our analysis, there are significant differences in the sequence capacities of the different reference alignment and *de novo* short read assembly tools, and the reference aligner outputs show significant differences in reads matching against the reference sequence in particular locations.

Specific Aim 2: Develop and implement NextGen genomic variation and expression analysis workflows

The Nimblegen exome capture of an eight-person pedigree sample set has been successfully concluded and sequencing of up to 15-fold average coverage is currently underway. Based on the initial shallow sequencing data, the framework of a SNP validation workflow has been completed, including filtering based on read quality, read depth and presence in dbSNP followed by the incorporation of pedigree information such as the presence/absence of non-SNP calls at identical chromosomal locations in parent-offspring trios. We will shortly evaluate the use of an expectation-maximization approach to predicting heterozygotes and homozygotes. In the final stage of this experiment, we will validate our approach with a sample of 1,000 individuals on a smaller subset of the captured genomic regions.

Specific Aim 3: Develop a NextGen workflow and visualization tool

A working prototype of Aqwa (pre-alpha) will soon be deployed, providing most of the core functionality of the project Aqwa alpha version. Power users are currently being recruited for the initial testing phase to assess the usability of the interface and the suitability of the current requirements specification.

RESOURCE SHARING

The Center for Computational Science is committed to the ideals of collaborative research. Intellectual property and data generated under this project will be administered in accordance with both University and NIH policies, including the NIH Data Sharing Policy and Implementation Guidance of March 5, 2003. The Aqwa source code will be made available as an open-source resource for educational, research and non-profit purposes under the terms of the GNU General Public License (GPL).

Ownership of sole or joint inventions developed under the project will be owned by the institution(s) employing the inventor(s). Inventors shall be determined by U.S. Patent law, Title 35 SC. University and Participating investigators/institutions will disclose any inventions developed under the project and such inventions will be reported and managed as provided by NIH policies. Sole inventions will be administered by the institution employing the inventor. Joint inventions shall be administered based on mutual consultation between the parties.

Similar procedures will be followed for copyrights. Materials generated under the project will be disseminated in accordance with University/Participating institutional and NIH policies. Depending on such policies, materials may be transferred to others under the terms of a material transfer agreement. Such access will be provided using web-based applications, as appropriate. Publication of data shall occur during the project, if appropriate, or at the end of the project, consistent with normal scientific practices.

Research data which documents, supports and validates research findings will be made available after the main findings from the final research data set have been accepted for publication. Such research data will be redacted to prevent the disclosure of personal identifiers. Should any intellectual property arise which requires a patent, we will ensure that the technology (materials and data) remains widely available to the research community in accordance with University policies and the NIH Principles and Guidelines document."

MILLER
SCHOOL OF MEDICINE
UNIVERSITY OF MIAMI

April 20, 2009

John R. Gilbert, Ph.D.
Directory, Center for Genome Technology
Miami Institute for Human Genomics
Leonard M. Miller School of Medicine
1120 NW 14th Street, CRB-814 (M-860)
Miami, Florida 33136

Sawsan Khuri, PhD
Bioinformatics Lead Scientist
Assistant Research Professor, Dr. John T. Macdonald Foundation
Department of Human Genetics,
1120 NW 14th Street, CRB 926
Miami, FL 33136

Dear Sawsan:

This letter is to confirm my support and that of the Miami Institute for Human Genomics for your project on the development of optimized NextGen bioinformatics pipelines for sequence assembly, gene expression and genomic variation analysis, and Aqwa, the NextGen workflow and visualization tool. The Miami Institute for Human Genomics has an urgent need for such a tool to support the activities of our NextGen sequencing core facility and for use in our research programs on metabolic disorders, cancer, and cardiovascular disease.

The Miami Institute for Human Genomics focuses on biomedical research in the following areas: Alzheimer's disease, amyotrophic lateral sclerosis, age-related macular degeneration, autism, Asperger disorder, Charcot-Marie-Tooth disease, familial spastic paraparesis/paraplegia, hereditary spastic paraparesis/paraplegia, multiple sclerosis, Parkinson's disease, thrombotic storm, tuberculosis and trichotillomania. Of particular relevance to this proposal are our plans to greatly expand our sequencing facilities later this year. At that point, our NextGen bioinformatics needs will become even greater and Aqwa will help us meet these demands. It will also provide us with significant time and labor cost savings.

I look forward to working with you and your team on our upcoming NextGen sequencing projects.

Regards,

REFERENCES

1. Pohlhaus JR, Cook-Deegan RM. Genomics research: world survey of public funding. BMC Genomics. 2008;9:472. PMCID: 2576262.
2. Morozova O, Marra MA. Applications of next-generation sequencing technologies in functional genomics. Genomics. 2008;92(5):255-64.
3. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. Science. 2008;320(5881):1344-9.
4. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):621-8.
5. Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. Genome Res. 2007;17(8):1195-201. PMCID: 1933516.
6. Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, et al. A general approach to single-nucleotide polymorphism discovery. Nat Genet. 1999;23(4):452-6.
7. Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet. 2008;40(6):722-9.
8. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-scale copy number polymorphism in the human genome. Science. 2004;305(5683):525-8.
9. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. Nat Genet. 2004;36(9):949-51.
10. Chiang DY, Getz G, Jaffe DB, O'Kelly MJ, Zhao X, Carter SL, et al. High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat Methods. 2009;6(1):99-103. PMCID: 2630795.
11. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. 2007;315(5813):848-53. PMCID: 2665772.
12. Szymczak S, Nuzzo A, Fuchsberger C, Schwarz DF, Ziegler A, Bellazzi R, et al. Genetic association studies for gene expressions: permutation-based mutual information in a comparison with standard ANOVA and as a novel approach for feature selection. BMC Proc. 2007;1 Suppl 1:S9. PMCID: 2359872.
13. Zheng T, Wang S, Cong L, Ding Y, Ionita-Laza I, Lo SH. Joint study of genetic regulators for expression traits related to breast cancer. BMC Proc. 2007;1 Suppl 1:S10. PMCID: 2367474.
14. Olivier M. From SNPs to function: the effect of sequence variation on gene expression. Focus on "A survey of genetic and epigenetic variation affecting human gene expression". Physiol Genomics. 2004;16(2):182-3.
15. Pastinen T, Sladek R, Gurd S, Sammak Aa, Ge B, Lepage P, et al. A survey of genetic and epigenetic variation affecting human gene expression. Physiol Genomics. 2004;16(2):184-93.
16. Stein L. Creating a bioinformatics nation. Nature. 2002;417(6885):119-20.
17. Bolchini D, Finkelstein A, Perrone V, Nagl S. Better bioinformatics through usability analysis. Bioinformatics. 2009;25(3):406-12.
18. Bare JC, Shannon PT, Schmid AK, Baliga NS. The Firegoose: two-way integration of diverse data from different bioinformatics web resources with desktop applications. BMC Bioinformatics. 2007;8:456. PMCID: 2211326.
19. Baker PG, Brass A, Bechhofer S, Goble C, Paton N, Stevens R. TAMBIS--Transparent Access to Multiple Bioinformatics Information Sources. Proc Int Conf Intell Syst Mol Biol. 1998;6:25-34.
20. Baker PG, Goble CA, Bechhofer S, Paton NW, Stevens R, Brass A. An ontology for bioinformatics applications. Bioinformatics. 1999;15(6):510-20.
21. Stevens R, Baker P, Bechhofer S, Ng G, Jacoby A, Paton NW, et al. TAMBIS: transparent access to multiple bioinformatics information sources. Bioinformatics. 2000;16(2):184-5.
22. Rak R, Kurgan L, Reformat M. xGENIA: A comprehensive OWL ontology based on the GENIA corpus. Bioinformation. 2007;1(9):360-2. PMCID: 1891717.
23. Salzberg S. Genome re-annotation: a wiki solution? Genome Biology. 2007;8(1):102.

PHS 398 Checklist

OMB Number: 0925-0001
Expiration Date: 9/30/2007

1. Application Type:

From SF 424 (R&R) Cover Page. The responses provided on the R&R cover page are repeated here for your reference, as you answer the questions that are specific to the PHS398.

* Type of Application:

● New ○ Resubmission ○ Renewal ○ Continuation ○ Revision

Federal Identifier:

2. Change of Investigator / Change of Institution Questions

❑ Change of principal investigator / program director

Name of former principal investigator / program director:

Prefix:
* First Name:
Middle Name:
* Last Name:
Suffix:

❑ Change of Grantee Institution

* Name of former institution:

3. Inventions and Patents (For renewal applications only)

* Inventions and Patents: Yes ○ No ○

If the answer is "Yes" then please answer the following:

* Previously Reported: Yes ○ No ○

4. * Program Income

Is program income anticipated during the periods for which the grant support is requested?

○Yes ●No

If you checked "yes" above (indicating that program income is anticipated), then use the format below to reflect the amount and source(s). Otherwise, leave this section blank.

*Budget Period *Anticipated Amount ($) *Source(s)

5. Assurances/Certifications (see instructions)

In agreeing to the assurances/certification section 18 on the SF424 (R&R) form, the authorized organizational representative agrees to comply with the policies, assurances and/or certifications listed in the agency's application guide, when applicable. Descriptions of individual assurances/certifications are provided at: http://grants.nih.gov/grants/funding/424

If unable to certify compliance, where applicable, provide an explanation and attach below.

Explanation:

Attachments

CertificationExplanation_attDataGroup0

File Name	Mime Type