**06-HG-101\* New computational and statistical methods for the analysis of large data sets from next-generation sequencing technologies. (High priority grant)s**

The introduction of new methods for DNA sequencing has opened new avenues, including large-scale sequencing studies, metagenomics, transcriptomics, genetic network analysis, and determination of the relationship of sequence variation and phenotypes to disease, to address heretofore unapproachable problems in biomedical research. However, since the large amounts (terabases) of data generated overwhelm existing computational resources and analytic methods, urgent action is needed to enable the translation of this rich new source of genomic information into medical benefit. Contact: Dr. Lisa Brooks, 301 496-7531, [brooksl@mail.nih.gov](mailto:brooksl@mail.nih.gov)

(*See end of document for Research Plan structure and format*)

**Specific aims**

NextGen sequencing technologies are fast approaching the ‘$1,000 genome’ target {NHGRI, 2008 #35}, an effort to develop sequencing technologies that are inexpensive and efficient enough for biomedical research and health care labs to routinely sequence entire genomes. A $5,000 genome will likely be attained in mid-2009 by Comparative Genomics while other NextGen industry players are not far behind with steadily increasing run yield and reducing 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 to generate and confirm hypotheses that may begin to unlock the secrets of the cell {Mardis, 2008 #23} and open up new avenues for clinical diagnostics. Bioinformatics infrastructure – hardware, software and personnel – is the bottleneck in the development of this new paradigm {Schuster, 2008 #18;Mardis, 2008 #23}. Costly investments of high performance computing resources and skilled personnel are required to develop bioinformatics algorithms, integrate diverse biological data sources and devise ways to visualize and interpret these huge data sets. Most biomedical research and health care labs are unable to provide even the minimum of these requirements. The specific aims of this study are designed to meet these requirements by providing small to medium sized laboratories with a tool for managing NextGen sequencing projects, from short read generation to bioinformatics analysis, which provides data visualization and allows the integration of diverse external genomic feature sets. This tool, named Aqwa (Automated Query and Workflow Agent), is being developed from a user-centric perspective, in contrast to many bioinformatics resources where user experience is often an afterthought. Aqwa is designed to shorten the time required for non tech-savvy research and laboratory personnel to accomplish hitherto impossible or difficult bioinformatics tasks swiftly and routinely.



Figure 1. Aqwa: a High Throughput NGS Data Analysis and Visualization Tool

**Background and Significance**

**NextGen Technologies**

The current mainstream NextGen platforms produce millions of short (50bp – 400bp) sequence reads. Each of the three main platforms, namely, Illumina/Solexa {Bennett, 2004 #29}, Roche/454 {Margulies, 2005 #21} and ABI/SOLiD {Shendure, 2005 #22} have their own inherent problems, including significant sequencing error rates and systematic errors. Despite an initial capital outlay of $600 000 to $1.35 million, the cost per base is substantially lower than for Sanger sequencing. 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 2) and smaller labs and molecular diagnostics facilities participating in growing numbers.

Prior to sequencing, NextGen sample preparation varies considerably but usually involves multiple steps taking 2–4 days to complete,depending on the platform. ‘Barcodes’ – unique identifier sequences added to reads – can also be used to analyze multiple samples within the same separate flow-cell lanesor compartments. These barcodes are ligated to individual samples which are then pooled and sequenced and later separated out based on their barcode.Barcode-based multiplexing and other incremental innovations in process streamlining, automation and chemistryrefinements will continue to reduce costs and sequencing errors. The recent rollout of paired-end reads (a.k.a. mate-paired reads) – short reads that flank a region of known length in the sample sequences - by all of the major platforms has provided a major advance in de novo assembly and the correction of reference alignment errors by eliminating alignments that do not match the size of gap between the paired reads {Shendure, 2008 #346}. There is also a strong need for flexible and effective targeted capture methods for isolating reduced genomic subsets, such as genomic regions, or exons of candidate genes, implicated in disease prior to NextGen sequencing of multiple individuals. Different approaches have already shown proof-of-concept, such as 10,000 enrichment by hybridizing biotinylated BACs (bacterial artificial chromosomes) with targeted segments of genomic DNA {Bashiardes, 2005 #347}, microarray-based enrichment of several kilobase-sized human genomic regions {Okou, 2007 #348} and multiplex PCR amplification of 170 exons {Fredriksson, 2007 #356}.



Figure 2. NextGen Sequencing Technology Roadmap

**3G Technologies**

Third generation sequencing technologies are being developed to sequence single DNA molecules faster and cheaper with streamlined samplepreparation. Real-time sequencing by synthesis is being developedby 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 synthesisrate 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 nmin diameter). Different methods are being tested to create nanopores, including inorganicmembranes (solid-state nanopores), genetically engineered protein channelsby Oxford Nanopore Technologies(<http://www.nanoporetech.com>), polymer-based nanofluidicchannels, and a combinationof nanopores with sequencing by hybridization by NABsys (<http://www.nabsys.com>). As these technologies develop, they will present new bioinformatics problems to be solved and greater data infrastructure demands.

**NextGen Applications**

A growing variety of molecular methods has enabled the investigation of a broad range of biological phenomena by high-throughput DNA sequencing, including genetic variation, RNA expression, protein-DNA interactions and chromosome conformation {Shendure, 2008 #346}. Chromatin immuno-precipitation (ChIP) is used to investigate protein-DNA interactions, which play a key part in regulating gene expression and controlling the availability of DNA for transcription and replication. In the technique, DNA chemically cross-linked to associated proteins is fragmented and transcription factor-specific antibodies are used to immunoprecipitate selected protein:DNA complexes. The DNA is then processed by NextGen sequencing (ChIP-Seq). Expression profiling (a.k.a. RNA-seq or transcriptome analysis) is another popular NextGen application that has been demonstrated as robust and sensitive in comparison to five microarray platforms {'t Hoen, 2008 #364}. Furthermore, microarrays cannot detect antisense transcription, which was found in 51% of all genes. Transcriptomes for mouse brain, liver and skeletal muscle were mapped by NextGen deep sequencing {Mortazavi, 2008 #417}, 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.

In yeast, NextGen expression analysis has demonstrated a larger, more complex transcriptome than had been expected {Nagalakshmi, 2008 #255}. 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. Similarly, 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 {Kidd, 2008 #94}. Somatically acquired genomic rearrangements have been implicated in cancer development. Paired-end read pairs that did not align correctly with respect to each other on the reference human genome, were used to characterize 306 germline structural variants and 103 somatic rearrangements to the base-pair level of resolution {Campbell, 2008 #93}. The results demonstrate the feasibility of using NextGen sequencing for the systematic, genome-wide characterization of rearrangements in human cancer genomes. At the level of large-scale genomic variation, copy number variations (CNV ) remain difficult to measure although CNVs of 100 kilobases and greater contribute substantially to genomic variation between normal humans {Sebat, 2004 #367}{Iafrate, 2004 #366}. 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 NextGen CNV technique, only a small fraction of copy-number variant (CNV) base pairs had been determined at the sequence level {Cooper, 2007 #365}. NextGen CNV mapping allows the discovery cancer-causing genes in genomic regions that show recurrent copy-number alterations (gains and losses) in tumor genomes {Chiang, 2009 #368}.

Other applications include discovering non-coding RNAs (e.g., miRNAs), sequencing the nuclear genomes of extinct species and metagenomics, such as the characterization of changes in biodiversity due to climate changes. For metagenomics, the growing number of sequenced genomes enables us to interpret partial sequences obtained by direct sampling of specific environmental niches to determine which kinds of species are present. The rapid, inexpensive, and massive data production enabled by NextGen platforms has caused a recent explosion in metagenomic studies. The NIH Human Microbiome Project is one of several international efforts using metagenomic analysis to study human health and has developing the new technological and bioinformatic tools as one of its four stated goals (http://www.genome.gov/25521743). New bioinformatics tools for assembling metagenomics data (http://nihroadmap.nih.gov/hmp/fundedresearch.asp) are required to assembling and finding genes and genomic variation in heterogeneous metagenomic datasets, a task that cannot be handled by currently available software.

Alongside the profound impact of NextGen applications in basic research, high throughput sequencing is now being adopted by clinical diagnostics laboratories for applications requiring deep sequence coverage and high-sensitivity such as rare HIV drug resistant variant detection {Wang, 2007 #34}. As the focus in human genetics has shifted to multi-gene, complex diseases, there is an increasing need for comprehensive diagnostic evaluations of multiple genes, enhanced by sequence enrichment/capture methods. NextGen analysis of placental mRNA - counting the numberof reads that map to each chromosome – has been used to confirm trisomy 21pregnancies, with additional supporting evidence obtained fortrisomy 18 and 13 pregnancies (REF). Other novel applications include the sequencing of ancient DNA samples and large-scale metagenomic analysis of environmentally derived samples.

**NextGen Bioinformatics**

The anticipated growth of NextGen sequencing among clinical diagnostics labs requires streamlined sample preparation and improved robustness through characterization of accuracy in validation studies [6]. Biomedical research labs also require methods for mitigating systemic bias in next generation data analysis. 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. At present there are insufficient studies comparing the efficacy and applicability of the various tools. A objective, performance-based comparison of NextGen bioinformatics tools is an important step towards lowering the bioinformatics hurdle and allowing smaller labs to do big science. Bioinformatics tools are available for reference alignment, de novo assembly, variant-discovery and alignment viewing. Among the reference aligners are Eland (GAPipeline v0.30, Illumina), Mira {Chevreux, 2004 #16}, Genomics Workbench (CLC Bio), Seqman NGen (DNAStar), NextGene (Soft Genetics), MAQ {Li, 2007 #1} {Li, 2008 #4} and Shrimp http://compbio.cs.toronto.edu/shrimp). De novo assemblers include Edina {Hernandez, 2008 #262}, EULER-SR {Chaisson, 2008 #284}, SHARCGS {Dohm, 2007 #286}, SSAKE {Warren, 2007 #289},Velvet {Zerbino, 2008 #15}, and SOAPdenovo (http://soap.genomics.org.cn). Some NextGen statistical data-analysis tools are also available, such as JMP Genomics (<http://www.jmp.com/software/genomics>) . Despite the growing number of software packages available for NextGen data, obtaining an accurately assembled sequence contig is a still a very challenging problem. The currently available assembly/alignment programs 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. A performance-based comparison of these tools is an important precondition for mitigating systemic bias in next generation data analysis. One anticipated outcome of this will be hastening the transition to a mature technology, with fewer bioinformatics applications used for a wide number of applications. In some cases, extensive comparisons may be needed to determine that certain bioinformatics approaches are suitable for a particular task however to determine this investigators can benefit from having all the applications at their fingertips.

Other bioinformatics challenges include the development of quality value standards similar to that of phred in Sanger sequencing {Shendure, 2008 #346} and the development of NextGen sequence viewers. As the bioinformatics challenge shifts from mastering the tools to gaining biological insights using huge amounts of data, visualization and interpretation of short read data become paramount. The three commercial software packages by DNAStar, SoftGeneticsand CLC Bio contain data viewers that allow the user to see read alignments, coveragedepth, genome annotations, and variant analysis. However, they as yet lack the capability for viewing data sets as large as a whole human chromosome and show performance losses even on sub-chromosome data sets. The three major genome viewers – UCSC Genome Browser, Ensembl genome

UCSC Genome Browser, Ensembl, Apollo, IGV

In summary, high throughput sequencing requires high throughput analysis. Although the falling cost of NGS sequencing has allowed a widening pool of users to access the technology, the costs of downstream bioinformatics analysis resources is often underestimated. Constraints on grant revenues and a shortage of qualified personnel mean that bioinformatics resources are scarce in comparison to the increasing rate of generation of NGS sequence data. Even provided the necessary funding and personnel, building bioinformatics capacity is a time-consuming process which can extend many months after sequence data becomes available.

**Preliminary studies**

**1. Evaluation of next gen assemblers**

We compared several commonly used short read assembly tools and propose a method for reducing these errors by combining different assemblies for the final result.

**METHOD**

Human mtDNA and whole-genome mRNA short reads produced using the Illumina/Solexa Genome Analyzer I platform were used, as well as E. coli, Herpes simplex and bacteriophage PhiX. Seven commercial and open-source short read assemblers were first assessed for assembly capacity in terms of the maximum number of reads that can be effectively assembled using relatively high-end computer hardware. We investigated the performance of Eland (GAPipeline v0.30, Illumina), Velvet v0.7.16 {Zerbino, 2008 #15}, Mira v2.9.25 {Chevreux, 2004 #16}, Genomics Workbench (CLC Bio) v1.2, Seqman NGen (DNAStar) 1.1, NextGene (Soft Genetics) 1.0 and MAQ v 0.6.8 {Li, 2007 #1} {Li, 2008 #4}. Assemblies produced by the different programs were compared and a consensus determined based on read identity and divergence from the relevant reference sequence. The overall combination of assemblies was viewed for quality control purposes using a sequence viewer that we developed to handle the huge data depth and breadth of sample types.

**RESULTS**

Our results indicate that there are significant differences in the capabilities of the different reference and *de novo* short read assembly tools as shown in Figures 1, 2 and 3. The resulting assemblies showed significant differences in read matching against the reference sequence in particular locations. For human data, Seqman NGen, Genomics Workbench and NextGen showed better performance in terms of the number of reads assembled however this may result in less accurate contigs. A combination of different assemblies can 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 can be accorded greater confidence with regard to their predicted SNPs, indels and breakpoints. Based on this research, we are currently developing a new visualization and analysis tool to meet the needs of next generation sequencing data analysis.

**CONCLUSION**

Effective use of these technologies depends on the correct interpretation of differing assembly results based on a consensus of reads incorporated by the various assembly algorithms. Manual verification of the assembly combination using a Next Generation sequence viewer improved confidence in the resulting assembly and aided comprehension of the strengths and weaknesses of this approach.

**2. Aqwa: Automated Query and Workflow Agent**

Transcriptome

SNP verification

Genetic network analysis – GeneSet Analyzer (Nick)

**Research design and methods**

*How key variables will be defined and measured and the general approach to analysis of the data.*

*Discussion of the strengths and weaknesses of the proposed study*

*Realistic time line*

Software design strategy

1. Human-centered development process

User tests

The user testing was performed by having a person sitting beside the

subject acting as facilitator and observer, posing questions and taking

notes when appropriate in order to gain as much insight as possible into

the perception and reasoning of the users in trying to complete the tasks

(“think aloud” technique). The scenario, goals and tasks are summarized

below.

Stevens,R., Goble,C., Baker,P., Brass,A. (2001) A classification of tasks in bioinformatics.

Bioinformatics, 17, 180-188.

{Stevens, 2001 #340}

1. Stevens R, Goble C, Baker P, Brass A. A classification of tasks in bioinformatics. Bioinformatics. 2001;17(2):180-8.

MOTIVATION: This paper reports on a survey of bioinformatics tasks currently undertaken by working biologists. The aim was to find the range of tasks that need to be supported and the components needed to do this in a general query system. This enabled a set of evaluation criteria to be used to assess both the biology and mechanical nature of general query systems. RESULTS: **A classification of the biological content of the tasks gathered offers a checklist for those tasks (and their specialisations) that should be offered in a general bioinformatics query system**. This semantic analysis was contrasted with a syntactic analysis that revealed the small number of components required to describe all bioinformatics questions. Both the range of biological tasks and syntactic task components can be seen to provide a set of bioinformatics requirements for general query systems. These requirements were used to evaluate two bioinformatics query systems.

Bolchini 2009:

A lot of effort, he said, has been spent on development, data integration, data visualization in bioinformatics resources, but less on “the difficulties users may encounter while searching this data, making sense of this data,” said Bolchini.

**ABSTRACT**

**Motivation:** Improving the usability of bioinformatics resources

enables researchers to find, interact with, share, compare and

manipulate important information more effectively and efficiently.

It thus enables researchers to gain improved insights into biological

processes with the potential, ultimately, of yielding new scientific

results. Usability ’barriers’ can pose significant obstacles to a

satisfactory user experience and force researchers to spend unnecessary

time and effort to complete their tasks. The number of

online biological databases available is growing and there is an

expanding community of diverse users. In this context there is an

increasing need to ensure the highest standards of usability.

**Results:** Using ‘state-of-the-art’ usability evaluation methods, we

have identified and characterised a sample of usability issues

potentially relevant to web bioinformatics resources in general.

These specifically concern the design of the navigation and

search mechanisms available to the user. The usability issues we

have discovered in our substantial case studies are undermining

the ability of users to find the information they need in their daily

research activities. In addition to characterising these issues,

specific recommendations for improvements are proposed leveraging

proven practices from web and usability engineering. The

methods and approach we exemplify can be readily adopted by

the developers of bioinformatics resources.

**Supplementary Information**: additional data about the usability

methods and a summary of the results of the usability analyses

presented in the paper is provided in a supplementary information

1. file.
2. The focus of usability analysis is to detect, understand, mitigate,
3. and ultimately prevent usability problems. A usability problem
4. is an obstacle to a successful user experience, meaning by “successful”

the effectiveness (feasibility) and efficiency (spending an acceptable amount of time and energy) in carrying out tasks.

For example, the fact that it takes several frustrating and timeconsuming

trials for a user to eventually get the desired result

from a search in a biological database may be caused by a number

of usability problems inherent in the way the search functionality

has been conceived, designed and communicated to the

user. There is evidence of an increasing awareness – at least in

the research arena – of the need of usability studies in the development

of biomedical systems in general (Rose, 2005). There

have been some notable efforts to address the usability of bioinformatics

systems. The challenge of bringing an increased

awareness of usability and user-centered design to the development

of bioinformatics applications can be tackled in a number

of different ways. The Human-Centered Software Engineering

(HCSE) at Concordia University has worked on developing

integrated web-based interfaces to popular bioinformatics portals

in order to provide integrated access to web resources relevant to

a set of typical tasks (Javahery *et al.*, 2004). The Human-

Computer Interaction Lab at the University of Maryland is investigating

advanced visualisation techniques to access and manipulate

large multimedia information sets in biological databases

(Hochheiser *et al.*, 2003). Tackling the work context of

bioinformaticians, Joan Bartlett at McGill University has been

investigating the daily activities of bioinformatics researchers in

order to derive a list of typical information tasks that entail the

use of web-based resources to complete (Bartlett and Toms,

2005). Working in a similar vein, Robert Stevens at University

of Manchester has undertaken research aimed at characterising

and classifying tasks in bioinformatics and analysing their interrelationships

(Stevens *et al.*, 2001).

Although these contributions cover important aspects of improving

the user experience of biological databases little has

been done to analyse the underlying *design* characteristics of

web bioinformatics resources that can lead to potential usability

problems. Tackling design issues identifies the usability problems

at their source, and helps to prevent the emergence of problems

1. in current and future applications.

## Bolchini, 2009. Better Bioinformatics Through Usability Analysis

paper

<http://bioinformatics.oxfordjournals.org/cgi/content/abstract/btn633v1?maxtoshow=&HITS=10&hits=10&RESULTFORMAT=&fulltext=Bolchini&searchid=1&FIRSTINDEX=0&resourcetype=HWCIT>

article

<http://www.genomeweb.com/informatics/usability-analysis-could-bring-benefits-bioinformatics-portals-and-research>

Aqwa roadmap

External data – API-based query: more flexible than using site’s web interface, can tackle more specialized, custom queries

ID conversion - DAVID {Huang da, 2008 #301}

Technology roadmap

Hardware requirements

8-node (8-core each) cluster

Production server

Development server

**PHS 398 Research Plan structure and format**

**(part of SF 424 (R&R) Application for Federal Assistance)**

*1. Introduction to Application* ***XXX NOT REQUIRED*** *(for RESUBMISSION or REVISION only)*

2. Specific Aims

3. Background and Significance

4. Preliminary Studies / Progress Report

5. Research Design and Methods

Notes on Required Format ( <http://grants.nih.gov/grants/funding/424/SF424_RR_Guide_General_Adobe_VerA.doc>)

Text attachments should be generated using word processing software and then converted to PDF using PDF generating software. Additional tips for creating PDF files can be found at <http://era.nih.gov/ElectronicReceipt/pdf_guidelines.htm>.

When attaching a PDF document to the actual forms, please note you are attaching an actual document, not just pointing to the location of an externally stored document. Therefore, if you revise the document after it has been attached, you **must** delete the previous attachment and then reattach the revised document to the application form. Use the “**View Attachment**” button to determine if the correct version has been attached.

**Font:** Use an Arial, Helvetica, Palatino Linotype, or Georgia typeface

**Color:** Black

**Size:** 11 points or larger. (A Symbol font may be used to insert Greek letters or special characters; the font size requirement still applies.)

**Type density:** including characters and spaces, must be no more than 15 characters per inch.Type may be no more than six lines per inch.

**Page Margins:** Use standard paper size (8 ½" x 11). Use at least one-half inch margins (top, bottom, left, and right) for all pages. No information should appear in the margins, including the PI’s name and page numbers.

**Header/footer:** Do not include any information in a header or footer of the attachments. Page numbers for the footer will be system-generated in the complete application, with all pages sequentially numbered.

**Figures, Graphs, Diagrams, Charts, Tables, Figure Legends, and Footnotes**

You may use a smaller type size but it must be in a black font color, readily legible, and follow the font typeface requirement. Color can be used in figures; however, all text must be in a black font color, clear and legible.

**Acronyms/Abbreviations:** If terms are not universally known, spell out the term the first time it is used and note the appropriate abbreviation in parentheses. The abbreviation may be used thereafter.

**Separate Attachments**

Separate attachments have been designed for the Research Plan sections to maximize automatic validations conducted by the eRA system. When the application is received by the agency, all of the Research Plan sections will be concatenated in the appropriate order so that reviewers and agency staff will see a single cohesive Research Plan.

While each section of the Research Plan needs to eventually be uploaded separately, applicants are encouraged to construct the Research Plan as a single document, separating sections into distinct PDF attachments just before uploading the files. In this way the applicant can better monitor formatting requirements such as page limits. When validating for page limits, the eRA Commons will not count the white space created by breaking the text into separate files for uploading.

**Page Limits**

Although many of the sections of this application are separate text (PDF) attachments, page limitations referenced in these instructions and/or funding opportunity announcement must still be followed. Agency validations will include checks for page limits. Some accommodation will be made for sections that when combined must fit within a specified limitation. Note that while these computer validations will help minimize incomplete and/or non-compliant applications, they do not replace the validations conducted by NIH staff. Applications found not to comply with the requirements may lead to rejection of the application during agency validation or delay in the review process.

All applications and proposals for NIH and other PHS agency funding must be self-contained within specified page limitations. Unless otherwise specified in an NIH solicitation, Internet website addresses (URLs) may not be used to provide information necessary to the review because reviewers are under no obligation to view the Internet sites. Moreover, reviewers are cautioned that they should not directly access an Internet site as it could compromise their anonymity.