**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 (1 page)**

NextGen sequencing technologies are fast approaching the ‘$1,000 genome’ target (1), 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 (2) and open up 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 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 (3 pages)**

**NextGen Technologies**

The current mainstream NextGen platforms produce millions of short (50bp – 400bp) sequence reads. Each of the three main platforms, namely, Illumina/Solexa (4), Roche/454 (5) and ABI/SOLiD (6) 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. Also attracted by huge base pair yields, large genome centers, core facilities and commercial contract-sequencing enterprises across the globe have already adopted this new technology (Figure 2) with smaller labs and molecular diagnostics facilities participating in growing numbers.

Sample preparation involves multiple steps and can take 2–4 days to complete,depending on the platform. ‘Bar codes’ – unique identifier sequences – can also be used to analyze multiple samples within the same separate flow-cell lanesor compartments. These ‘bar codes’ are ligated to individual samples which are then pooled and sequenced and later separated out based on their barcode.Incremental innovations through process streamlining, automation and chemistryrefinements will continue to reduce costs and improve data handling.



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, which is due to launch commercially in 2010, has a mean DNA synthesisrate of approximately 4 bases per second, with a maximum read length of 4,000 bp (REF).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>).

**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}.

**ELUCIDATING DNA-PROTEIN**

**INTERACTIONS THROUGH**

**CHROMATIN**

**IMMUNOPRECIPITATION**

**SEQUENCING**

**Mardis 2008**

**{Mardis, 2008 #23}**

The association betweenDNAand proteins is a

fundamental biological interaction that plays a

key part in regulating gene expression and controlling

the availability of DNA for transcription,

replication, and other processes. These

interactions can be studied in a focused manner

using a technique called chromatin immunoprecipitation

(ChIP) (43). ChIP entails a series

of steps: (*a*) DNA and associated proteins are

chemically cross-linked; (*b*) nuclei are isolated,

lysed, and the DNA is fragmented; (*c*) an antibody

specific for the DNA binding protein

(transcription factor, histone, etc.) of interest is

used to selectively immunoprecipitate the associated

protein:DNA complexes; and (*d* ) the

chemical crosslinks between DNA and protein

are reversed and the DNA is claimed for downstream

analysis

Expression profiling (a.k.a. RNA-seq or transcriptome analysis)

1. 't Hoen PAC, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RHAM, de Menezes RX, et al. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. Nucl Acids Res. 2008;36(21):e141-.

The hippocampal expression profiles of wild-type mice and mice transgenic for deltaC-doublecortin-like kinase were compared with Solexa/Illumina deep sequencing technology and five different microarray platforms. With **Illumina's digital gene expression assay, we obtained [~]2.4 million sequence tags per sample, their abundance spanning four orders of magnitude. Results were highly reproducible, even across laboratories. With a dedicated Bayesian model, we found differential expression of 3179 transcripts with an estimated false-discovery rate of 8.5%. This is a much higher figure than found for microarrays.** The overlap in differentially expressed transcripts found with deep sequencing and microarrays was most significant for Affymetrix. The changes in expression observed by deep sequencing were larger than observed by microarrays or quantitative PCR. Relevant processes such as calmodulin-dependent protein kinase activity and vesicle transport along microtubules were found affected by deep sequencing but not by microarrays. **While undetectable by microarrays, antisense transcription was found for 51% of all genes and alternative polyadenylation for 47%. We conclude that deep sequencing provides a major advance in robustness, comparability and richness of expression profiling data and is expected to boost collaborative, comparative and integrative genomics studies.**(7)

In yeast, RNA-Seq was used to demonstrate a much larger, more complex transcriptome than had been expected. 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 (8).

1. 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.

The identification of untranslated regions, introns, and coding regions within an organism remains challenging. We developed a quantitative sequencing-based method called RNA-Seq for mapping transcribed regions, in which complementary DNA fragments are subjected to high-throughput sequencing and mapped to the genome. We applied RNA-Seq to generate a high-resolution transcriptome map of the yeast genome and demonstrated that most (74.5%) of the nonrepetitive sequence of the yeast genome is transcribed. We confirmed many known and predicted introns and demonstrated that others are not actively used. Alternative initiation codons and upstream open reading frames also were identified for many yeast genes. We also found unexpected 3'-end heterogeneity and the presence of many overlapping genes. These results indicate that the yeast transcriptome is more complex than previously appreciated.

Complex and large-scale structural genetic variation between individual humans in the form of insertions, deletions and inversions from a few thousand to millions of base pairs in length have been elucidated in a high-resolution map of human structural variation (9).

CNV Kidd et al

Genetic variation among individual humans occurs on many different scales, ranging from gross alterations in the human karyotype to single nucleotide changes. Here we explore variation on an intermediate scale—particularly insertions, deletions and inversions affecting from a few thousand to a few million base pairs. We employed a clone-based method to interrogate this intermediate structural variation in eight individuals of diverse geographic ancestry. Our analysis provides a comprehensive overview of the normal pattern of structural variation present in these genomes, refining the location of 1,695 structural variants. We find that 50% were seen in more than one individual and that nearly half lay outside regions of the genome previously described as structurally variant. We discover 525 new insertion sequences that are not present in the human reference genome and show that many of these are variable in copy number between individuals. Complete sequencing of 261 structural variants reveals considerable locus complexity and provides insights into the different mutational processes that have shaped the human genome. These data provide the first high-resolution sequence map of human structural variation—a standard for genotyping platforms and a prelude to future individual genome sequencing projects.

Human genetic structural variation, including large (more than 1 kilobase pair (kbp)) insertions, deletions and inversions of DNA, is common[1](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2424287#R1)–[9](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2424287#R9). These differences are thought to encompass more polymorphic base pairs than single nucleotide differences[5](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2424287#R5),[6](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2424287#R6),[9](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2424287#R9),[10](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2424287#R10). The importance of structural variation to human health and common genetic disease has become increasingly apparent[11](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2424287#R11)–[14](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2424287#R14). However, only a small fraction of copy-number variant (CNV) base pairs have been determined at the sequence level[15](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2424287#R15). Most genome-wide approaches for detecting CNVs are indirect, depending on signal intensity differences to predict regions of variation. They therefore provide limited positional information and cannot detect balanced events such as inversions. Because the human genome reference assembly is now viewed as a patchwork of structurally variant sequence[1](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2424287#R1),[2](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2424287#R2), it is expected that sequencing projects of other individuals would reveal previously uncharacterized human euchromatic sequence, in a similar manner to comparisons between the Celera and International Human Genome Project assemblies[16](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2424287#R16)–[18](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2424287#R18). We implemented an approach to construct clone-based maps of eight human genomes with the aim of systematically cloning and sequencing structural variants more than 8 kbp in length. We present a validated, structural variation map of these eight human genomes of Asian, European and African ancestry, identify 525 regions of previously uncharacterized ‘novel sequence’, and provide sequence resolution of 261 selected regions of structural variation in the human genome.

Next generation sequencing is replacing other technologies in many applications. For example, NextGen sequences (CHiP-seq) have been shown to be comparable to CHiP-chip microarrays (REF) and

**GENE EXPRESSION:**

**SEQUENCING THE**

**TRANSCRIPTOME**

transcriptome analysis (Mortazavi et al.).

* **Chiang DY, Getz G, Jaffe DB, OKelly MJ, Zhao X, et al. (2009)** [**High-resolution mapping of copy-number alterations with massively parallel sequencing.**](http://www.ncbi.nlm.nih.gov/pubmed/19043412) Nat Methods 6:99-103
  + Cancer results from somatic alterations in key genes, including point mutations, copy-number alterations and structural rearrangements. A powerful way to discover cancer-causing genes is to identify genomic regions that show recurrent copy-number alterations (gains and losses) in tumor genomes. Recent advances in sequencing technologies suggest that massively parallel sequencing may provide a feasible alternative to DNA microarrays for detecting copy-number alterations. Here we present a statistical analysis of the power to detect copy-number alterations of a given size; (ii) SegSeq, an algorithm to segment equal copy numbers from massively parallel sequence data; and (iii) analysis of experimental data from three matched pairs of tumor and normal cell lines. We show that a collection of approximately 14 million aligned sequence reads from human cell lines has comparable power to detect events as the current generation of DNA microarrays and has over twofold better precision for localizing breakpoints (typically, to within approximately 1 kilobase).
* **Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, et al. (2008)** [**Identification of genetic variants using bar-coded multiplexed sequencing.**](http://www.ncbi.nlm.nih.gov/pubmed/18794863) Nat Methods 5:887-93.
  + We developed a generalized framework for multiplexed resequencing of targeted human genome regions on the Illumina Genome Analyzer using degenerate indexed DNA bar codes ligated to fragmented DNA before sequencing. Using this method, we simultaneously sequenced the DNA of multiple HapMap individuals at several Encyclopedia of DNA Elements (ENCODE) regions. We then evaluated the use of Bayes factors for discovering and genotyping polymorphisms. For polymorphisms that were either previously identified within the Single Nucleotide Polymorphism database (dbSNP) or visually evident upon re-inspection of archived ENCODE traces, we observed a false positive rate of 11.3% using strict thresholds for predicting variants and 69.6% for lax thresholds. Conversely, false negative rates were 10.8-90.8%, with false negatives at stricter cut-offs occurring at lower coverage (<10 aligned reads). These results suggest that >90% of genetic variants are discoverable using multiplexed sequencing provided sufficient coverage at the polymorphic base.
* **Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS (2009)** [**Genome-Wide Analysis In Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling.**](http://www.ncbi.nlm.nih.gov/pubmed/19213877) Science
  + Techniques for systematically monitoring protein translation have lagged far behind methods for measuring mRNA levels. Here, we present a ribosome profiling strategy, based on deep sequencing of ribosome protected mRNA fragments, that enables genome-wide investigation of translation with sub-codon resolution. We used this technique to monitor translation in budding yeast under both rich and starvation conditions. These studies defined the protein sequences being translated and found extensive translational control both for determining absolute protein abundance and for responding to environmental stress. We also observed distinct phases during translation involving a large decrease in ribosome density going from early to late peptide elongation as well as widespread, regulated initiation at non-AUG codons. Ribosome profiling is readily adaptable to other organisms, making high-precision investigation of protein translation experimentally accessible.

**DISCOVERING NONCODING**

**RNAs**

One of the most exciting areas of biological

research in recent years has been the discovery

and functional analysis of noncoding RNA

(ncRNA) systems in different organisms. First

described in plants, ncRNAs are providing new

insights into gene regulation in animal systems

as well, as recognized by the awarding of the

Nobel Prize in Medicine and Physiology to Andrew

Fire and Craig Mello in 2006. Perhaps the

most profound impact of next-generation sequencing

technology has been on the discovery

of novel ncRNAs belonging to distinct classes

in an extraordinarily diverse set of species (3, 8,

9, 18, 22, 29, 41, 55). In fact, this approach has

been responsible for the discovery of ncRNA

classes in organisms not previously known to

possess them (41). These discoveries are being

coupled with an ever-expanding comprehension

of the functions embodied by these unique

RNA species, including gene regulation by a

variety of mechanisms. In this regard, studying

the roles of specific microRNAs (miRNAs) in

cancer is helping to uncover certain aspects of

the disease (10, 44)

**ANCIENT GENOMES**

**RESURRECTED**

Attempts to characterize fossil-derived DNAs

have been limited by the degraded state of the

sample, which in the past permitted only mitochondrial

DNA sequencing and typically involvedPCRamplification

of specific mitochondrial

genome regions (1, 15, 23, 24, 35, 40).

The advent of next-generation sequencing has

for the first time made it possible to directly

sample the nuclear genomes of ancient remains

from the cave bear (34), mammoth (37), and

the Neanderthal (17, 33).

**METAGENOMICS EMERGES**

Characterizing the biodiversity found on Earth

is of particular interest as climate changes

reshape our planet. DNA- or RNA-based approaches

for this purpose are becoming increasingly

powerful as the growing number

of sequenced genomes enables us to interpret

partial sequences obtained by direct sampling

of specific environmental niches. Such investigations

are referred to as metagenomics, and

are typically aimed at answering the question:

who’s there? Conventionally, this question is

addressed by isolating DNA from an environmental

sample, amplifying the collective of 16S

ribosomal RNA (rRNA) genes with degenerate

PCR primer sets, subcloning the PCR products

that result, and classifying the taxa present

according to a database of assigned 16S rRNA

sequences. As an alternative, DNA (or RNA) is

isolated, subcloned, and then sequenced to produce

a fragment pool representative of the existing

population. These sequences can then be

translated in silico into protein fragments and

compared with the existing database of annotated

genome sequences to identify community

members. In both approaches, deep sequencing

of the population of subclones is necessary to

obtain the full spectrum of taxa present, and is

limited by potential cloning bias that can result

from the use of bacterial cloning. By sampling

RNA sequences from a metagenomic isolate,

one can attempt to reconstruct metabolic pathways

that are active in a given environment (45,

47). **Several early metagenomic studies utilized**

**DNA sequence sampling by capillary sequencing**

**to investigate an acidophilic biofilm (49), an**

**acid mine site (12) and the Sargasso Sea (52).**

**Although these studies defined metagenomics**

**as a scientific pursuit, they were limited in the**

**breadth of diversity that could be sampled owing**

**to the expense of the conventional sequencing**

**process. By contrast, the rapid, inexpensive,**

**and massive data production enabled by nextgeneration**

**platforms has caused a recent explosion**

**in metagenomic studies. These studies**

**include previously sampled environments such**

**as the ocean (2, 19, 42) and an acid mine site**

**(13), but soil (14, 27) and coral reefs (54) also**

**were studied by Roche/454 pyrosequencing.**

NIH Human Microbiome Project is only one of several international efforts designed to take advantage of metagenomic analysis to study human health

Broadly, the project has set the following goals:

* Determining whether individuals share a core human microbiome
* Understanding whether changes in the human microbiome can be correlated with changes in human health
* **Developing the new technological and bioinformatic tools needed to support these goals**
* Addressing the ethical, legal and social implications raised by human microbiome research.

. Supporting these characterization

efforts will be a large-scale project to sequence

hundreds of isolated microbial genomes that

are known symbionts of humans as references

(**http://www.genome.gov/25521743**). The

early successes in human microbiome characterization

and the apparent interplay between

the human host and its microbial census have

resulted in the inclusion of a Human Microbiome

Initative in the NIH Roadmap (**http://**

**nihroadmap.nih.gov/hmp/**). The associated

funding opportunities, with the advantages

offered by next-generation sequencing instrumentation,

should initiate a revolution in our

understanding of how the human microbiome

influences our health status.

Research is already underway to develop bioinformatics tools for assembling metagenomics data (<http://nihroadmap.nih.gov/hmp/fundedresearch.asp>).

Mihai P. 2008 <http://crisp.cit.nih.gov/crisp/CRISP_LIB.getdoc?textkey=7571986&p_grant_num=1R01HG004885-01&p_query=&ticket=77401286&p_audit_session_id=365863008&p_keywords>=

For example, the assembly of environmental data is being performed with software originally intended for homogeneous DNA sources, such as clonal bacterial populations or inbred eukaryotes. **These programs are ill-suited to the assembly of heterogeneous microbial communities and numerous "hacks" have been necessary to produce the assemblies published to date. This proposal aims to fill the need for specialized software for assembling and finding genes in metagenomic datasets. A particular focus will be on developing tools for uncovering genomic variation within the assemblies of microbial communities.**

The proposal further aims to investigate the impact of experimental design and choice of sequencing technology on the ability to assemble and analyze metagenomic data, through the development of software for simulating bacterial populations and emulating a variety sequencing strategies. … **The current proposal provides scientists with components of the software infrastructure that will be essential for genomic studies of the human microbiome.**

The research applications of NextGen sequencing have expanded rapidly to encompass resequencing, de novo assembly, transcriptomics, metagenomics and the detection of rare variants and chromosomal aberrations. 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 (10). 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 (11), Genomics Workbench (CLC Bio), Seqman NGen (DNAStar), NextGene (Soft Genetics), MAQ (12) (13) and Shrimp http://compbio.cs.toronto.edu/shrimp). De novo assemblers include Edina (14), EULER-SR (15), SHARCGS (16), SSAKE (17),Velvet (18), 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 result of this will be speeding up the transition to a mature technology, where few applications are deemed suitable for a wide number of applications. *In some cases, extensive comparisons may be needed to determine that there are suitable for a particular task however to determine this investigators can benefit from having all the applications at their fingertips.*

Mate-paired reads, now possible with all of the major platforms,

are anticipated to have a major impact on the overall success of

*de novo* assembly with short reads, and several algorithms have been

already developed that take advantage of these38–41.

{Shendure, 2008 #346}

Sequence viewers

Voelkerding 2009

Recentlyreleased commercial software for alignment and de novo assemblyincludes packages from DNAStar ([www.dnastar.com](http://www.dnastar.com)), SoftGenetics([www.softgenetics.com](http://www.softgenetics.com)), and CLC bio ([www.clcbio.com](http://www.clcbio.com)) that featuredata viewers that allow the user to see read alignments, coveragedepth, genome annotations, and variant analysis. Fig. 5presentssome examples of NGS data viewed in 2 different software systems.



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 (18), Mira v2.9.25 (11), Genomics Workbench (CLC Bio) v1.2, Seqman NGen (DNAStar) 1.1, NextGene (Soft Genetics) 1.0 and MAQ v 0.6.8 (12) (13). 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**

, including how key variables will be defined and measured and the general approach to analysis of the data.

This section should include discussion of control for possible confounders and assessment of effect modification, when appropriate, as well as a justification of the proposed sample size.

Discussion of the strengths and weaknesses of the proposed study

e) 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

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.

4. Bennett S. Solexa Ltd. Pharmacogenomics. 2004;5(4):433-8.

Solexa Ltd is developing an integrated system, based on a breakthrough single molecule sequencing technology, to address a US$2 billion market that is expected to grow exponentially alongside and as a consequence of further technological enhancements. The system, software and consumables will initially be sold to research organizations, pharmaceutical companies and diagnostic companies that will sequence large regions of genomic DNA, including whole genomes, at costs several orders of magnitude below current levels. Solexa expects to launch its first product in 2006, and as it continues to make time and cost efficiencies, additional products will be launched into the expanding markets that will have broad applications in basic research through to healthcare management.

5. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437(7057):376-80.

6. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, et al. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. Science. 2005;309(5741):1728-32.

We describe a DNA sequencing technology in which a commonly available, inexpensive epifluorescence microscope is converted to rapid nonelectrophoretic DNA sequencing automation. We apply this technology to resequence an evolved strain of Escherichia coli at less than one error per million consensus bases. A cell-free, mate-paired library provided single DNA molecules that were amplified in parallel to 1-micrometer beads by emulsion polymerase chain reaction. Millions of beads were immobilized in a polyacrylamide gel and subjected to automated cycles of sequencing by ligation and four-color imaging. Cost per base was roughly one-ninth as much as that of conventional sequencing. Our protocols were implemented with off-the-shelf instrumentation and reagents.

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.

Human cancers often carry many somatically acquired genomic rearrangements, some of which may be implicated in cancer development. However, conventional strategies for characterizing rearrangements are laborious and low-throughput and have low sensitivity or poor resolution. We used massively parallel sequencing to generate sequence reads from both ends of short DNA fragments derived from the genomes of two individuals with lung cancer. By investigating read pairs that did not align correctly with respect to each other on the reference human genome, we characterized 306 germline structural variants and 103 somatic rearrangements to the base-pair level of resolution. The patterns of germline and somatic rearrangement were markedly different. Many somatic rearrangements were from amplicons, although rearrangements outside these regions, notably including tandem duplications, were also observed. Some somatic rearrangements led to abnormal transcripts, including two from internal tandem duplications and two fusion transcripts created by interchromosomal rearrangements. Germline variants were predominantly mediated by retrotransposition, often involving AluY and LINE elements. The results demonstrate the feasibility of systematic, genome-wide characterization of rearrangements in complex human cancer genomes, raising the prospect of a new harvest of genes associated with cancer using this strategy.

8. 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.

The identification of untranslated regions, introns, and coding regions within an organism remains challenging. We developed a quantitative sequencing-based method called RNA-Seq for mapping transcribed regions, in which complementary DNA fragments are subjected to high-throughput sequencing and mapped to the genome. We applied RNA-Seq to generate a high-resolution transcriptome map of the yeast genome and demonstrated that most (74.5%) of the nonrepetitive sequence of the yeast genome is transcribed. We confirmed many known and predicted introns and demonstrated that others are not actively used. Alternative initiation codons and upstream open reading frames also were identified for many yeast genes. We also found unexpected 3'-end heterogeneity and the presence of many overlapping genes. These results indicate that the yeast transcriptome is more complex than previously appreciated.

9. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, et al. Mapping and sequencing of structural variation from eight human genomes. Nature. 2008;453(7191):56-64.

Genetic variation among individual humans occurs on many different scales, ranging from gross alterations in the human karyotype to single nucleotide changes. Here we explore variation on an intermediate scale--particularly insertions, deletions and inversions affecting from a few thousand to a few million base pairs. We employed a clone-based method to interrogate this intermediate structural variation in eight individuals of diverse geographic ancestry. Our analysis provides a comprehensive overview of the normal pattern of structural variation present in these genomes, refining the location of 1,695 structural variants. We find that 50% were seen in more than one individual and that nearly half lay outside regions of the genome previously described as structurally variant. We discover 525 new insertion sequences that are not present in the human reference genome and show that many of these are variable in copy number between individuals. Complete sequencing of 261 structural variants reveals considerable locus complexity and provides insights into the different mutational processes that have shaped the human genome. These data provide the first high-resolution sequence map of human structural variation--a standard for genotyping platforms and a prelude to future individual genome sequencing projects.

10. 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.

The detection of mutant spectra within a population of microorganisms is critical for the management of drug-resistant infections. We performed ultra-deep pyrosequencing to detect minor sequence variants in HIV-1 protease and reverse transcriptase (RT) genes from clinical plasma samples. We estimated empirical error rates from four HIV-1 plasmid clones and used them to develop a statistical approach to distinguish authentic minor variants from sequencing errors in eight clinical samples. Ultra-deep pyrosequencing detected an average of 58 variants per sample compared with an average of eight variants per sample detected by conventional direct-PCR dideoxynucleotide sequencing. In the clinical sample with the largest number of minor sequence variants, all 60 variants present in > or =3% of genomes and 20 of 35 variants present in <3% of genomes were confirmed by limiting dilution sequencing. With appropriate analysis, ultra-deep pyrosequencing is a promising method for characterizing genetic diversity and detecting minor yet clinically relevant variants in biological samples with complex genetic populations.

11. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WEG, Wetter T, et al. Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs. Genome Res. 2004;14:1147-59.

12. Li H, Durbin R. MAQ: Mapping and Assembly with Qualities. 2007 [updated 2007; cited 2008 November 11th]; Available from: <http://maq.sourceforge.net/>.

13. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008;18(11):1851-8.

New sequencing technologies promise a new era in the use of DNA sequence. However, some of these technologies produce very short reads, typically of a few tens of base pairs, and to use these reads effectively requires new algorithms and software. In particular, there is a major issue in efficiently aligning short reads to a reference genome and handling ambiguity or lack of accuracy in this alignment. Here we introduce the concept of mapping quality, a measure of the confidence that a read actually comes from the position it is aligned to by the mapping algorithm. We describe the software MAQ that can build assemblies by mapping shotgun short reads to a reference genome, using quality scores to derive genotype calls of the consensus sequence of a diploid genome, e.g., from a human sample. MAQ makes full use of mate-pair information and estimates the error probability of each read alignment. Error probabilities are also derived for the final genotype calls, using a Bayesian statistical model that incorporates the mapping qualities, error probabilities from the raw sequence quality scores, sampling of the two haplotypes, and an empirical model for correlated errors at a site. Both read mapping and genotype calling are evaluated on simulated data and real data. MAQ is accurate, efficient, versatile, and user-friendly. It is freely available at <http://maq.sourceforge.net>.

14. Hernandez D, Francois P, Farinelli L, Osteras M, Schrenzel J. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res. 2008;18(5):802-9.

Novel high-throughput DNA sequencing technologies allow researchers to characterize a bacterial genome during a single experiment and at a moderate cost. However, the increase in sequencing throughput that is allowed by using such platforms is obtained at the expense of individual sequence read length, which must be assembled into longer contigs to be exploitable. This study focuses on the Illumina sequencing platform that produces millions of very short sequences that are 35 bases in length. We propose a de novo assembler software that is dedicated to process such data. Based on a classical overlap graph representation and on the detection of potentially spurious reads, our software generates a set of accurate contigs of several kilobases that cover most of the bacterial genome. The assembly results were validated by comparing data sets that were obtained experimentally for Staphylococcus aureus strain MW2 and Helicobacter acinonychis strain Sheeba with that of their published genomes acquired by conventional sequencing of 1.5- to 3.0-kb fragments. We also provide indications that the broad coverage achieved by high-throughput sequencing might allow for the detection of clonal polymorphisms in the set of DNA molecules being sequenced.

15. Chaisson MJ, Pevzner PA. Short read fragment assembly of bacterial genomes. Genome Res. 2008;18(2):324-30.

In the last year, high-throughput sequencing technologies have progressed from proof-of-concept to production quality. While these methods produce high-quality reads, they have yet to produce reads comparable in length to Sanger-based sequencing. Current fragment assembly algorithms have been implemented and optimized for mate-paired Sanger-based reads, and thus do not perform well on short reads produced by short read technologies. We present a new Eulerian assembler that generates nearly optimal short read assemblies of bacterial genomes and describe an approach to assemble reads in the case of the popular hybrid protocol when short and long Sanger-based reads are combined.

16. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. Genome Res. 2007;17(11):1697-706.

The latest revolution in the DNA sequencing field has been brought about by the development of automated sequencers that are capable of generating giga base pair data sets quickly and at low cost. Applications of such technologies seem to be limited to resequencing and transcript discovery, due to the shortness of the generated reads. In order to extend the fields of application to de novo sequencing, we developed the SHARCGS algorithm to assemble short-read (25-40-mer) data with high accuracy and speed. The efficiency of SHARCGS was tested on BAC inserts from three eukaryotic species, on two yeast chromosomes, and on two bacterial genomes (Haemophilus influenzae, Escherichia coli). We show that 30-mer-based BAC assemblies have N50 sizes >20 kbp for Drosophila and Arabidopsis and >4 kbp for human in simulations taking missing reads and wrong base calls into account. We assembled 949,974 contigs with length >50 bp, and only one single contig could not be aligned error-free against the reference sequences. We generated 36-mer reads for the genome of Helicobacter acinonychis on the Illumina 1G sequencing instrument and assembled 937 contigs covering 98% of the genome with an N50 size of 3.7 kbp. With the exception of five contigs that differ in 1-4 positions relative to the reference sequence, all contigs matched the genome error-free. Thus, SHARCGS is a suitable tool for fully exploiting novel sequencing technologies by assembling sequence contigs de novo with high confidence and by outperforming existing assembly algorithms in terms of speed and accuracy.

17. Warren RL, Sutton GG, Jones SJ, Holt RA. Assembling millions of short DNA sequences using SSAKE. Bioinformatics. 2007;23(4):500-1.

Novel DNA sequencing technologies with the potential for up to three orders magnitude more sequence throughput than conventional Sanger sequencing are emerging. The instrument now available from Solexa Ltd, produces millions of short DNA sequences of 25 nt each. Due to ubiquitous repeats in large genomes and the inability of short sequences to uniquely and unambiguously characterize them, the short read length limits applicability for de novo sequencing. However, given the sequencing depth and the throughput of this instrument, stringent assembly of highly identical sequences can be achieved. We describe SSAKE, a tool for aggressively assembling millions of short nucleotide sequences by progressively searching through a prefix tree for the longest possible overlap between any two sequences. SSAKE is designed to help leverage the information from short sequence reads by stringently assembling them into contiguous sequences that can be used to characterize novel sequencing targets. Availability: <http://www.bcgsc.ca/bioinfo/software/ssake>.

18. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18:821-9.

**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.