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

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

**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 steadily increasing run yield and reducing cost per Mbase to $2 and less.

The bottleneck in the shift to this new paradigm lies in bioinformatics data analysis and interpretation (refs). This area requires a confluence of high performance computing, bioinformatics algorithms development, integration of diverse biological data sources and novel data representation and visualization technologies (ref).

The specific aims of this study are designed to meet these requirements by providing a tool for managing large sequence projects, from sequence generation to bioinformatics analysis, provide data visualization and incorporate diverse external genomic feature sets. A performance-based comparison of these tools is an important precondition for mitigating systemic bias in next generation data analysis.



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 [2], Roche/454 [3] and ABI/SOLiD [4] 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**

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 being adopted by clinical diagnostics laboratories for applications requiring deep sequence coverage and high-sensitivity such as rare HIV drug resistant variant detection [5]. Next generation sequencing is replacing other technologies in many applications. For example, NextGen sequences have been shown to be comparable to microarrays in CHiP-chip applications (REF) and transcriptome analysis (Mortazavi et al.). 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**

Bioinformatics and data interpretation are now the bottleneck in the shift to the new sequencing paradigm (REF). This requires a confluence of Linux-based high performance computing, bioinformatics algorithms development, integration of diverse biological data sources and novel data representation and visualization technologies (REF). Bioinformatics tools are available for reference alignment, de novo assembly, variant-discovery and alignment viewing. For example, Reference alignment tools include Eland, MAQ, Shrimp, MIRA, Genomics Workbench, Seqman Ngen, Nextgen (REFS) and several de novo assemblers are available, including Velvet (REF) and Euler (REF). Some NextGen statistical data-analysis tools are also available, such as JMP Genomics (REF). However, using NextGen data to obtain 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.

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]. A performance-based comparison of these tools is thus an important precondition for mitigating systemic bias in next generation data analysis.

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 [7], Mira v2.9.25 [8], Genomics Workbench (CLC Bio) v1.2, Seqman NGen (DNAStar) 1.1, NextGene (Soft Genetics) 1.0 and MAQ v 0.6.8 [9] [10]. 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**

Software design strategy

Aqwa roadmap

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; Available from: <http://www.genome.gov/27527585>.

2. Bennett, S. (2004) Pharmacogenomics **5**(4) 433-8.

3. Margulies, M., et al. (2005) Nature **437**(7057) 376-380.

4. Shendure, J., et al. (2005) Science **309**(5741) 1728-1732.

5. Wang, C., et al. (2007) Genome Res **17**(8) 1195-201.

6. Voelkerding, K.V., S.A. Dames, and J.D. Durtschi (2009) Clin Chem **55**(4) 641-658.

7. Zerbino, D.R. and E. Birney (2008) Genome Res. **18** 821-829.

8. Chevreux, B., et al. (2004) Genome Res. **14** 1147-1159.

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

10. Li, H., J. Ruan, and R. Durbin (2008) Genome Res **18**(11) 1851-8.