**2. Specific Aims (one page maximum; separate PDF attachment)**

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

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

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

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

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

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

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

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

1. NHGRI. NHGRI Seeks DNA Sequencing Technologies Fit for Routine Laboratory and Medical Use. 2008 [updated 2008; cited]; Available from: <http://www.genome.gov/27527585>.

2. Mardis ER. Next-Generation DNA Sequencing Methods. Annual Review of Genomics and Human Genetics. 2008;9(1):387-402.

3. Schuster SC. Next-generation sequencing transforms today's biology. Nat Meth. 2008;5(1):16-8.