High priority

**06-HG-101 New computational and statistical methods for the analysis of large data sets from next-generation sequencing technologies.**

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 (ref). A $5,000 genome will likely be attained by mid-2009 by Comparative Genomics while other major industry players (ref) are steadily increasing run yield and reducing cost per Mbase to $2 and less (ref). The overall development trajectory of the industry suggests that, within a few years, differences between the leading sequencing technologies will diminish to the point that they can be used interchangeably to address the same biological problems. Bioinformatics tools are being developed to harness the large volumes and novel kinds of next generation sequencing data for applications such as reference or de novo alignment, and variant and copy number prediction. The performance of these bioinformatics tools varies widely in terms of data volume capacity, number of reads aligned/assembled, error rates and bias. A performance-based comparison of these tools is an important precondition for mitigating systemic bias in next generation data analysis.

Ultimately, we want to be able to compare the genomic content between individuals, tissues or individual cells to determine the causes and factors influencing diseases and other pathologies in humans and other organisms of vital economic or social importance (ref??). In pharmaceuticals development, development pipelines are drying up now that the readily accessible therapy candidates (the so-called “ low-hanging fruit”) have already been discovered. Real-time analysis of complete human genomes marks a qualitative shift in modern medicine from population-based research to individual-based research, opening up new avenues for therapies for complex diseases and rejuvenating the pharmaceuticals and therapeutics industries. The currently emerging bottleneck in the shift to this new paradigm is data analysis and interpretation (refs). This area requires a confluence of high performance computing, bioinformatics algorithms development, the 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 and by developing new tools to visualize analysis results and incorporate diverse external genomic feature sets.

1 Develop Aqwa to manage the analysis of next gen sequencing data

Base calling, alignment/assembly pipeline

Assembly comparison and combination

2. Improve alignment

Evaluate all current ones

Develop new alignment algorithms

Distributed (Grid, cluster)

3. Analysis of variants

Comparison of SNP calling algorithms

Pipeline for SNP verification

HapMap

dbSNP

4. Integration with viewer

Large genome feature sets

Concurrent, high-multiple data sets

Integrate with Spotfire, Genespring information

**Background and Significance (3 pages)**

NextGen sequencing technologies are fast approaching the $1,000 genome target (ref). A $5,000 genome will likely be attained by mid-2009 by Comparative Genomics while ABI/SOLiD (ref), Illumina/Solexa (ref) and Roche/454 (ref) are steadily increasing run yield and reducing cost per Mbase to $2 and less (ref). The different next generation sequencing technologies have their particular performance and error profile – for example, long reads prone to homopolymer error for 454, and short reads susceptible to substitution errors for Solexa reads – and are commonly used in distinct applications, such as 454 reads for *de novo* sequencing and Solexa/SOLiD for analysis of variants. Combinations of reads from different technologies can also be used filling in the gaps in 454 assemblies. One problem

The overall development trajectory of the industry suggests that, within a few years, differences between the leading sequencing technologies will diminish to the point that they can be used interchangeably to address the same biological problems. Applications of next generation sequencing now encompass the main realms of sequence analysis hitherto occupied by a variety of technologies such as microarrays in CHiP-chip applications (ref – GWAS/transcriptome? – Jennifer paper on Wiki) and transcriptome analysis(Mortazavi? Other refs), SAGE (transcriptome ref), de Novo sequencing (ref). Next generation sequencing also provides novel applications such as ultra-deep sequencing for detection of rare variants (HIV, NextGen ppt on rare variants paper).

Bioinformatics tools are being developed to deal with the large volumes and novel kinds of next generation sequencing data, both for alignment against a reference sequence (Eland, MAQ, Shrimp, MIRA, Genomics Workbench, Seqman Ngen, Nextgen) and de novo alignment (Velvet, Euler). The performance of these bioinformatics tools varies widely in terms of data volume capacity, number of reads aligned/assembled, error rates and bias. A performance-based comparison of these tools is an important precondition for mitigating systemic bias in next generation data analysis.

Ultimately, we want to be able to compare the genomic content of individuals with other individuals or tissues or individual cells to determine the causes and factors influencing diseases and other pathologies in humans and other organisms of vital economic or social importance (ref??). In pharmaceuticals development, development pipelines are drying up now that the readily accessible therapy candidates (the so-called “ low-hanging fruit”) have already been discovered. Real-time analysis of complete human genomes will mark a qualitative shift in modern medicine from population-based research to individual-based research, opening up new avenues for therapies for complex diseases and rejuvenating the pharmaceuticals and therapeutics industries. The currently emerging bottleneck in the shift to this new paradigm is data analysis and interpretation (refs). This area requires a confluence of high performance computing, bioinformatics algorithms development, the integration of diverse biological data sources and novel data representation and visualization technologies (ref).

**Preliminary studies**

Evaluation of next gen assemblers

Aqwa project pipeline & viewer

**Research design and methods**

Software design strategy