**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 {NHGRI, 2008 #35}, an effort to develop sequencing technologies that are inexpensive and efficient enough to sequence entire genomes as a routine part of biomedical research and health care. 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 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.



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

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, integration of diverse biological data sources and novel data representation and visualization technologies (ref).

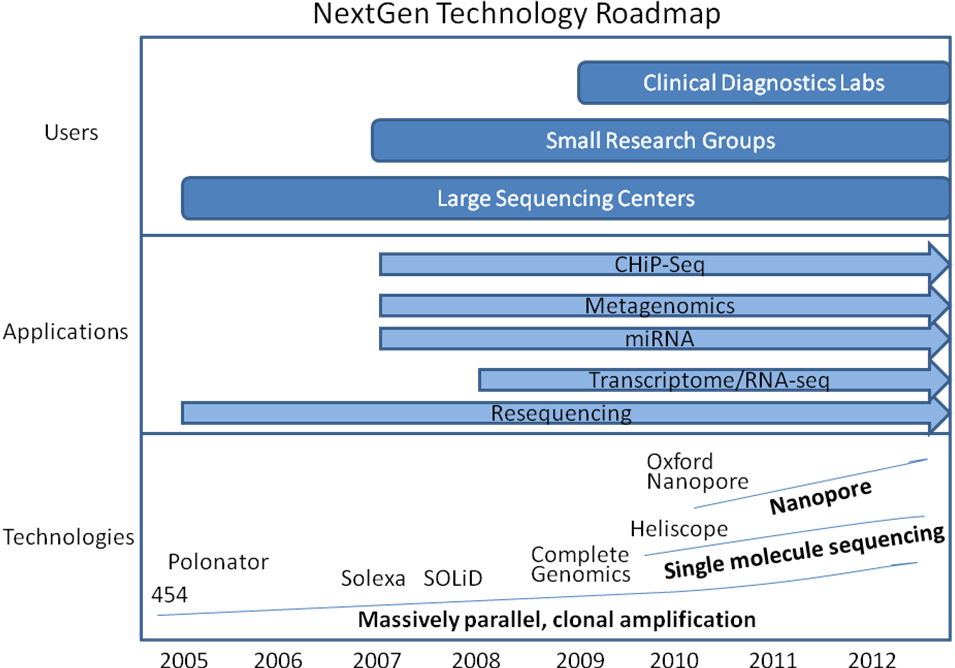
The specific aims of this study are designed to meet these requirements by providing a tool for managing large sequence projects from sequence generation to bioinformatics analysis and by developing new tools to visualize analysis results and incorporate diverse external genomic feature sets.

**Background and Significance (3 pages)**

**NextGen Technologies**

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 because of this can be used complementarily or for distinct applications, such as 454 reads for *de novo* sequencing and Solexa/SOLiD for analysis of variants.

Next generation sequencing (NGS) technologies produce millions of short (50bp – 400bp) sequence reads. These new sequencing methodologies, 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. Using NGS data to obtain an accurately assembled sequence contig is a still a very challenging problem. The first generations of NGS assembly programs have varying applicability, from bacterial to human data sets, and differing error rate, all of which may lead to suboptimal assemblies.



**3G Technologies**

New single-molecule sequencing technologies in development maydecrease sequencing time, reduce costs, and streamline samplepreparation. Real-time sequencing by synthesis is being developedby VisiGen (<http://www.visigenbio.com>) and Pacific Biosciences(<http://www.pacificbiosciences.com>). VisiGen’s approachuses DNA polymerase modified with a fluorescent donor molecule.Attached to a glass slide surface, the polymerase directs strandextension from primed DNA templates. Nucleotides are modifiedwith fluorescent acceptor molecules, and light energy is usedduring incorporation to invoke fluorescence resonance energytransfer between polymerase and nucleotide fluorescent moieties,the latter being in the -phosphate position and cleaved awayduring incorporation. The company envisions its platform willconsist of a massively parallel array of tethered DNA polymerasesthat will generate 1 x 106 bp of sequence per second.



Pacific Biosciences performs single-molecule real-time sequencingand uses phospholinked fluorescently labeled dNTPs. DNA sequencingis performed in thousands of reaction wells 50–100 nmin diameter that are fabricated with a thin metal cladding filmdeposited on an optical waveguide consisting of a solid, transparentsilicon dioxide substrate. Each reaction well is a nanophotonicchamber in which only the bottom third is visualized, producinga detection volume of approximately 20 zL (20 x 10–21 L). DNA polymerase/template complexes are immobilized to thewell bottoms, and 4 differently labeled dNTPs are added. Asthe DNA polymerase incorporates complementary nucleotides, eachbase is held within the detection volume for tens of milliseconds,orders of magnitude longer than the amount of time it takesfor a nucleotide to diffuse in and out of the detection volume.Laser excitation enables the incorporation events in individualwells to be captured through the optical waveguide, with thefluorescent color detected reflecting the identity of the dNTPincorporated. For sequencing, Pacific Biosciences uses a modifiedphi29 DNA polymerase that has enhanced kinetic properties forincorporating the system’s phospholinked fluorescentlylabeled dNTPs. In addition, phi29 DNA polymerase is highly processive,with strand-displacement activity. By taking advantage of theseproperties, Pacific Biosciences has demonstrated sequencingreads exceeding 4000 bases when a circularized single-strandedDNA molecule is used as template. In this configuration, thephi29 DNA polymerase carries out multiple laps of DNA strand-displacementsynthesis around the circular template. The mean DNA-synthesisrate was determined to be approximately 4 bases/s. The observederrors, including deletions, insertions, and mismatches canbe addressed by developing a consensus sequence read derivedfrom the multiple rounds of template sequencing. Further refinementof the chemistries and platform instrumentation are ongoing,with a 2010 target date for commercial launch([92](http://www.clinchem.org/cgi/content/full/55/4/641#R92))([93](http://www.clinchem.org/cgi/content/full/55/4/641#R93))([94](http://www.clinchem.org/cgi/content/full/55/4/641#R94)).

Farther out toward the horizon is sequencing based on monitoringthe passage of DNA molecules through nanopores 2–5 nmor greater in diameter. Nanopores are being fabricated in inorganicmembranes (solid-state nanopores), assembled from protein channelsin lipid membranes, or configured in polymer-based nanofluidicchannels. In some configurations, current is applied acrossnanopore membranes to drive the translocation of negativelycharged DNA molecules through pores while monitoring changesin membrane electrical conductance measured in the picoampererange. NABsys (<http://www.nabsys.com>) is pursuing a combinationof nanopores with sequencing by hybridization in which single-strandedDNA molecules are hybridized with a library of hexamers of knownsequence. The hybridized DNA is interrogated through a nanopore,with the current changes being different in regions of hexamerhybridization. The patterns of hybridization are used to mapannealing regions and determine sequence. Oxford Nanopore Technologies(<http://www.nanoporetech.com>) is developing nanopore-based sequencingthat uses an -hemolysin protein channel in reconstituted lipidbilayers. The nanopores are situated in individual array wells,and single DNA molecules are introduced into the wells and progressivelydigested by exonuclease. The released single-nucleotide basesenter the nanopore and alter the electrical current, creatinga characteristic current change for each individual base([95](http://www.clinchem.org/cgi/content/full/55/4/641#R95))([96](http://www.clinchem.org/cgi/content/full/55/4/641#R96)).Although the technology is seemingly futuristic, considerableNHGRI funding is being directed toward a variety of nanoporetechnologies under development as part of the goal of achievingthe $1000 genome. For further descriptions of nanopore technologies,the reader is referred to recent reviews([97](http://www.clinchem.org/cgi/content/full/55/4/641#R97))([98](http://www.clinchem.org/cgi/content/full/55/4/641#R98)).



**NextGen Applications**

|  |
| --- |
| **A Clinical Future for NGS** |

From the impact that NGS has made at the basic-research level,we can anticipate its translation into molecular diagnostics.Key issues that will need to be addressed in this transitionwill include complexity of technical procedures, robustness,accuracy, and cost. By all these measures, NGS platforms willbenefit from continued process streamlining, automation, chemistryrefinements, cost reductions, and improved data handling. Thecost of NGS is currently substantial in terms of the investmentin capital equipment (from approximately $600 000 for the Roche/454Life Science, Illumina, and Applied Biosystems SOLiD platformsto $1.35 million for the HeliScope platform) and costs of sequencingreagents (from approximately $3500–$4500 for the Illumina,Applied Biosystems, and Roche/454 platforms to $18 000 for theHeliScope platform). Nonetheless, the cost per base is substantiallylower than for Sanger sequencing, and combined with the tremendousoutput, it is straightforward to see why genome centers, corefacilities, and commercial contract-sequencing enterprises havereadily adopted this new technology.

Work flow considerationsinclude the fact that preparation of a sample library requiresmultiple molecular biology steps and 2–4 days to complete, **or longer for SOLiD**,depending on the platform. In addition to the required molecularbiology expertise, data analysis requires expertise in bioinformaticsfacilitated by a knowledge of Linux operating systems. Leveragingthe high-throughput capacity of NGS platforms can be facilitatedby analyzing multiple samples with separate flow-cell lanesor compartments. In addition, unique identifier sequences or"bar codes" can be ligated to individual samples, which cansubsequently be pooled and sequenced. After sequencing, sequencesof individual samples are derived by data deconvolution([81](http://www.clinchem.org/cgi/content/full/55/4/641#R81))([82](http://www.clinchem.org/cgi/content/full/55/4/641#R82))([83](http://www.clinchem.org/cgi/content/full/55/4/641#R83)).

CLINICAL DIAGNOSTICS

Alongside the expansion of applications in basic research, NGS is being beginning to be adopted by large clinical diagnostics laboratories for applications requiring deep sequence coverage and high-sensitivity such as rare HIV drug resistant variant detection {Wang, 2007 #34}

The transition of NGS into clinical diagnostics is in the earlystages of development in large reference laboratories and isbeing leveraged for applications that require large amountsof sequence information, relative quantification, and high-sensitivitydetection. Examples that meet these criteria include the aforementioneddetection of mutations in tumor cells from biopsies or in thecirculation. In the area of mitochondrial disorders, NGS canbe used to sequence the entire 16.5-kb mitochondrial genome,determine mutation heteroplasmy percentage, and analyze nucleargenes whose protein products affect mitochondrial metabolism—allin a single analytical run. In the authors’ laboratory,sequencing of mycobacterial genomes is ongoing as an approachto refine organism identification and support clinical epidemiologicinvestigations. HIV quasi-species detection and relative quantificationhave been demonstrated and can be used to monitor emerging drugresistance([84](http://www.clinchem.org/cgi/content/full/55/4/641#R84)). For human genetics, there is an increasing needto analyze multiple genes that, when mutated, lead to overlappingphysical findings and clinical phenotypes. For example, 16 differentgenes are implicated in the pathogenesis of hypertrophic cardiomyopathy([85](http://www.clinchem.org/cgi/content/full/55/4/641#R85))([86](http://www.clinchem.org/cgi/content/full/55/4/641#R86)).For a comprehensive diagnostic evaluation in such settings,it will be necessary to sequence upwards of 100 000 to 200 000bp. The coupling of NGS with the genomic-enrichment techniquesdescribed above offers a promising approach to this technicalchallenge.

Recently, investigative groups led by Y.M. Dennis Lo and StephenQuake have applied NGS to the detection of fetal chromosomalaneuploidy([87](http://www.clinchem.org/cgi/content/full/55/4/641#R87))([88](http://www.clinchem.org/cgi/content/full/55/4/641#R88)). Prior work had demonstrated that cell-freefetal nucleic acids (DNA and RNA) are present in maternal bloodduring pregnancy, along with maternally derived cell-free nucleicacids. Several analytical approaches that use cell-free fetalnucleic acids have been developed to determine fetal aneuploidy,including the analysis of placental mRNA derived from the chromosomesof interest (e.g., chromosome 21) and the determination of relativechromosomal dosage via digital PCR analysis of a large numberof target chromosomal loci compared with reference chromosomalloci([89](http://www.clinchem.org/cgi/content/full/55/4/641#R89))([90](http://www.clinchem.org/cgi/content/full/55/4/641#R90))([91](http://www.clinchem.org/cgi/content/full/55/4/641#R91)). Building on the concept of relative chromosomedosage, the Lo and Quake groups have independently shown thefeasibility of converting cell-free DNA from maternal bloodinto an Illumina library, followed by sequencing and mappingthe reads to the reference human genome. Counting the numberof reads that map to each chromosome allows the relative dosageof each chromosome to be ascertained. If fetal aneuploidy ispresent, the number of sequence reads mapping to the affectedchromosome would be expected to be statistically overrepresentedin the data set. This expectation was confirmed in trisomy 21pregnancies, with additional supporting evidence obtained fortrisomy 18 and 13 pregnancies. These studies open a new avenuefor assessing fetal aneuploidy and provide a foundation forNGS-based analysis of cell-free DNA in both nonpathologic andpathophysiological states.

**NextGen Bioinformatics**

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. 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 medicine, enabling research for therapies for complex diseases.

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).

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}.

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).

**Preliminary studies**

**1. Evaluation of next gen assemblers**

Next generation sequencing (NGS) technologies produce millions of short (50bp – 400bp) sequence reads. These new sequencing methodologies, 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. Using NGS data to obtain an accurately assembled sequence contig is a still a very challenging problem. The first generations of NGS assembly programs have varying applicability, from bacterial to human data sets, and differing error rates (Figure 1), all of which may lead to suboptimal assemblies. 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.

**METHODS**

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.

**Aqwa project pipeline & viewer**

Need for a viewer

Voelkerding 2009

<http://www.clinchem.org/cgi/content/full/55/4/641#R15>

NGS software features vary with the application and in generalmay include alignment, de novo assembly, alignment viewing,and variant-discovery programs. In addition some NGS statisticaldata-analysis tools are being developed (such as JMP Genomics;SAS Institute). Software packages available for alignment andassembly to a reference sequence include Zoom([71](http://www.clinchem.org/cgi/content/full/55/4/641#R71)), MAQ([67](http://www.clinchem.org/cgi/content/full/55/4/641#R67)),Mosaik([72](http://www.clinchem.org/cgi/content/full/55/4/641#R72)), SOAP([73](http://www.clinchem.org/cgi/content/full/55/4/641#R73)), and SHRiMP (<http://compbio.cs.toronto.edu/shrimp/>),which supports SOLiD color-space analysis. Software for de novoassembly includes Edina([70](http://www.clinchem.org/cgi/content/full/55/4/641#R70)), EULER-SR([74](http://www.clinchem.org/cgi/content/full/55/4/641#R74)), SHARCGS([75](http://www.clinchem.org/cgi/content/full/55/4/641#R75)), SSAKE([69](http://www.clinchem.org/cgi/content/full/55/4/641#R69)),Velvet([76](http://www.clinchem.org/cgi/content/full/55/4/641#R76)), and SOAPdenovo (<http://soap.genomics.org.cn/>). 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.



BIOINFORMATICS IS THE BOTTLENECK

Voelkerding 2009

The past few years have witnessed the emergence of NGS technologiesthat share a common basis, massively parallel sequencing ofclonally amplified DNA molecules. In 2008, the first NGS platformbased on single-molecule DNA sequencing was launched. On thehorizon are real-time single-molecule DNA-sequencing technologiesand approaches based on nanopores. NGS has had a substantialimpact on basic genomics research in terms of scale and feasibility.Over the next several years, NGS is anticipated to transitioninto clinical-diagnostics use. Essential elements to make thistransition successful will be the requirement of streamliningthe processes, especially sample preparation, coupled with improvementsin technology robustness and characterization of accuracy throughvalidation studies. **The large amounts of sequence-data outputwill pose a bioinformatics challenge for the clinical laboratory.In addition to data processing, the interpretation of sequencingresults will require further characterization of the genomicvariation present in the regions analyzed.** Although considerablework lies ahead to implement NGS into clinical diagnostics,the potential applications are exciting and numerous.

BIOINFORMATICS TIME & BUDGETS ARE UNDERESTIMATED

Both for research and clinical diagnostics, high throughput sequencing requires high throughput analysis. Although the falling cost of NGS sequencing made it accessible to a widening pool of users, 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.

**Research design and methods**

Software design strategy

#### Analysis and Tools

* **Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, et al. (2009)** [**ABySS: A parallel assembler for short read sequence data.**](http://www.ncbi.nlm.nih.gov/pubmed/19251739) Genome Res.
  + Widespread adoption of massively parallel DNA sequencing instruments has prompted the recent development of de novo short read assembly algorithms. A common shortcoming of the available tools is their inability to efficiently assemble vast amounts of data generated from large-scale sequencing projects, such as the sequencing of individual human genomes to catalog natural genetic variation. To address this limitation, we developed ABySS (Assembly By Short Sequences), a parallelized sequence assembler. As a demonstration of the capability of our software, we assembled 3.5 billion paired-end reads from the genome of an African male publicly released by Illumina Inc. Approximately 2.76 million contigs >/=100bp in length were created with an N50 size of 1499bp, representing 68% of the reference human genome. Analysis of these contigs identified polymorphic and novel sequences not present in the human reference assembly, which were validated by alignment to alternate human assemblies and to other primate genomes.
* **Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, et al. (2009)** [**PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data.**](http://www.ncbi.nlm.nih.gov/pubmed/19236709) Genome Biol 10:R23.
  + Personal-genomics endeavors, such as the "1000 Genomes Project", are generating maps of genomic structural variants (SVs) by analyzing ends of massively sequenced genome-fragments. To process these we developed Paired-End Mapper (PEMer; <http://sv.gersteinlab.org/pemer>). This comprises a parallelizable analysis pipeline, compatible with several next-generation sequencing platforms; simulation-based error models, yielding confidence-values for each SV; and a back-end database. The simulations demonstrated high SV-reconstruction efficiency for PEMer's coverage-adjusted multi-cutoff scoring-strategy and showed its relative insensitivity to base-calling errors.
* **Malhis N, Butterfield YS, Ester M, Jones SJ (2009)** [**Slider--maximum use of probability information for alignment of short sequence reads and SNP detection.**](http://www.ncbi.nlm.nih.gov/pubmed/18974170) Bioinformatics 25:6-13.
  + A plethora of alignment tools have been created that are designed to best fit different types of alignment conditions. While some of these are made for aligning Illumina Sequence Analyzer reads, none of these are fully utilizing its probability (prb) output. In this article, we will introduce a new alignment approach (Slider) that reduces the alignment problem space by utilizing each read base's probabilities given in the prb files. RESULTS: Compared with other aligners, Slider has higher alignment accuracy and efficiency. In addition, given that Slider matches bases with probabilities other than the most probable, it significantly reduces the percentage of base mismatches. The result is that its SNP predictions are more accurate than other SNP prediction approaches used today that start from the most probable sequence, including those using base quality.
* **Fejes AP, Robertson G, Bilenky M, Varhol R, Bainbridge M, et al. (2008)** [**FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology.**](http://www.ncbi.nlm.nih.gov/pubmed/18599518) Bioinformatics 24:1729-30.
  + Next-generation sequencing can provide insight into protein-DNA association events on a genome-wide scale, and is being applied in an increasing number of applications in genomics and meta-genomics research. However, few software applications are available for interpreting these experiments. We present here an efficient application for use with chromatin-immunoprecipitation (ChIP-Seq) experimental data that includes novel functionality for identifying areas of gene enrichment and transcription factor binding site locations, as well as for estimating DNA fragment size distributions in enriched areas. The FindPeaks application can generate UCSC compatible custom 'WIG' track files from aligned-read files for short-read sequencing technology. The software application can be executed on any platform capable of running a Java Runtime Environment. Memory requirements are proportional to the number of sequencing reads analyzed; typically 4 GB permits processing of up to 40 million reads. AVAILABILITY: The FindPeaks 3.1 package and manual, containing algorithm descriptions, usage instructions and examples, are available at <http://www.bcgsc.ca/platform/bioinfo/software/findpeaks> Source files for FindPeaks 3.1 are available for academic use.
* **Erlich Y, Mitra PP, delaBastide M, McCombie WR, Hannon GJ.(2008).** [**Alta-Cyclic: a self-optimizing base caller for next-generation sequencing.**](http://www.ncbi.nlm.nih.gov/pubmed/18604217?ordinalpos=4&itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_DefaultReportPanel.Pubmed_RVDocSum) Nat Methods. 2008 Aug;5(8):679-82.
  + Next-generation sequencing is limited to short read lengths and by high error rates. We systematically analyzed sources of noise in the Illumina Genome Analyzer that contribute to these high error rates and developed a base caller, Alta-Cyclic, that uses machine learning to compensate for noise factors. Alta-Cyclic substantially improved the number of accurate reads for sequencing runs up to 78 bases and reduced systematic biases, facilitating confident identification of sequence variants.

#### Improvements to Sample Preparation and Sequencing Chemistry

* **Sorber K, Chiu C, Webster D, Dimon M, Ruby JG, et al. (2008)** [**The long march: a sample preparation technique that enhances contig length and coverage by high-throughput short-read sequencing.**](http://www.ncbi.nlm.nih.gov/pubmed/18941527) PLoS ONE 3:e3495.
  + High-throughput short-read technologies have revolutionized DNA sequencing by drastically reducing the cost per base of sequencing information. Despite producing gigabases of sequence per run, these technologies still present obstacles in resequencing and de novo assembly applications due to biased or insufficient target sequence coverage. We present here a simple sample preparation method termed the "long march" that increases both contig lengths and target sequence coverage using high-throughput short-read technologies. By incorporating a Type IIS restriction enzyme recognition motif into the sequencing primer adapter, successive rounds of restriction enzyme cleavage and adapter ligation produce a set of nested sub-libraries from the initial amplicon library. Sequence reads from these sub-libraries are offset from each other with enough overlap to aid assembly and contig extension. We demonstrate the utility of the long march in resequencing of the Plasmodium falciparum transcriptome, where the number of genomic bases covered was increased by 39%, as well as in metagenomic analysis of a serum sample from a patient with hepatitis B virus (HBV)-related acute liver failure, where the number of HBV bases covered was increased by 42%. We also offer a theoretical optimization of the long march for de novo sequence assembly.
* **Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ. (2008)** [**A large genome center's improvements to the Illumina sequencing system.**](http://www.ncbi.nlm.nih.gov/pubmed/19034268?ordinalpos=5&itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_DefaultReportPanel.Pubmed_RVDocSum) Nat Methods. 2008 Dec;5(12):1005-10.
  + The Wellcome Trust Sanger Institute is one of the world's largest genome centers, and a substantial amount of our sequencing is performed with 'next-generation' massively parallel sequencing technologies: in June 2008 the quantity of purity-filtered sequence data generated by our Genome Analyzer (Illumina) platforms reached 1 terabase, and our average weekly Illumina production output is currently 64 gigabases. Here we describe a set of improvements we have made to the standard Illumina protocols to make the library preparation more reliable in a high-throughput environment, to reduce bias, tighten insert size distribution and reliably obtain high yields of data.
* **Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ.(2009)** [**Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G C)-biased genomes.**](http://www.ncbi.nlm.nih.gov/pubmed/19287394?ordinalpos=1&itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_DefaultReportPanel.Pubmed_RVDocSum) Nat Methods. 2009 Mar 15.
  + Amplification artifacts introduced during library preparation for the Illumina Genome Analyzer increase the likelihood that an appreciable proportion of these sequences will be duplicates and cause an uneven distribution of read coverage across the targeted sequencing regions. As a consequence, these unfavorable features result in difficulties in genome assembly and variation analysis from the short reads, particularly when the sequences are from genomes with base compositions at the extremes of high or low G+C content. Here we present an amplification-free method of library preparation, in which the cluster amplification step, rather than the PCR, enriches for fully ligated template strands, reducing the incidence of duplicate sequences, improving read mapping and single nucleotide polymorphism calling and aiding de novo assembly. We illustrate this by generating and analyzing DNA sequences from extremely (G+C)-poor (Plasmodium falciparum), (G+C)-neutral (Escherichia coli) and (G+C)-rich (Bordetella pertussis) genomes.

#### Sequence Capture

* **Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C. (2009)** [**Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing.**](http://www.ncbi.nlm.nih.gov/pubmed/19182786?ordinalpos=1&itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_DefaultReportPanel.Pubmed_RVDocSum) Nat Biotechnol. 2009 Feb;27(2):182-9
  + Targeting genomic loci by massively parallel sequencing requires new methods to enrich templates to be sequenced. We developed a capture method that uses biotinylated RNA 'baits' to fish targets out of a 'pond' of DNA fragments. The RNA is transcribed from PCR-amplified oligodeoxynucleotides originally synthesized on a microarray, generating sufficient bait for multiple captures at concentrations high enough to drive the hybridization. We tested this method with 170-mer baits that target >15,000 coding exons (2.5 Mb) and four regions (1.7 Mb total) using Illumina sequencing as read-out. About 90% of uniquely aligning bases fell on or near bait sequence; up to 50% lay on exons proper. The uniformity was such that approximately 60% of target bases in the exonic 'catch', and approximately 80% in the regional catch, had at least half the mean coverage. One lane of Illumina sequence was sufficient to call high-confidence genotypes for 89% of the targeted exon space.

#### DNA / RNA / Transcriptome Sequencing

* **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.

#### Reviews and Primers

* **Voelkerding KV, Dames SA, Durtschi JD (2009)** [**Next-Generation Sequencing: From Basic Research to Diagnostics.**](http://www.ncbi.nlm.nih.gov/pubmed/19246620) Clin Chem
  + For the past 30 years, the Sanger method has been the dominant approach and gold standard for DNA sequencing. The commercial launch of the first massively parallel pyrosequencing platform in 2005 ushered in the new era of high-throughput genomic analysis now referred to as next-generation sequencing (NGS). CONTENT: This review describes fundamental principles of commercially available NGS platforms. Although the platforms differ in their engineering configurations and sequencing chemistries, they share a technical paradigm in that sequencing of spatially separated, clonally amplified DNA templates or single DNA molecules is performed in a flow cell in a massively parallel manner. Through iterative cycles of polymerase-mediated nucleotide extensions or, in one approach, through successive oligonucleotide ligations, sequence outputs in the range of hundreds of megabases to gigabases are now obtained routinely. Highlighted in this review are the impact of NGS on basic research, bioinformatics considerations, and translation of this technology into clinical diagnostics. Also presented is a view into future technologies, including real-time single-molecule DNA sequencing and nanopore-based sequencing. SUMMARY: In the relatively short time frame since 2005, NGS has fundamentally altered genomics research and allowed investigators to conduct experiments that were previously not technically feasible or affordable. The various technologies that constitute this new paradigm continue to evolve, and further improvements in technology robustness and process streamlining will pave the path for translation into clinical diagnostics.

Walsh 2003

Fig. 8. Disruptive technology industry roadmapping process

