
GeneSifter: Next Generation Data Management and Analysis for Next Generation Sequencing

Dale Baskin, N. Eric Olson, Laura Lucas, Todd Smith¹

Abstract

Next generation sequencing technology is rapidly changing the way laboratories and researchers approach the management and analysis of biological information. The sheer volume of data produced by emerging technologies threatens to overwhelm many researchers,^{1,2} and labs increasingly find themselves in the business of trying to build and run data centers rather than devoting resources to scientific discovery.

GeneSifter® Lab Edition provides labs and researchers with next generation tools specifically designed to address the evolving needs of next generation sequencing, including:

- A cloud computing model that frees labs from the burden of building and maintaining expensive data centers.
- The ability to track experimentally relevant information throughout an entire project within a single unified system.
- A framework for bioinformaticians to build automated analysis pipelines that can be shared across the research community.
- A platform which allows scientists to run complex application-specific data analysis pipelines on large data sets with the press of a button.
- A universally accessible, OS-independent system that can be used to distribute and visualize data from anywhere.

Introduction

Next generation sequencing technology (NextGen) enables accelerated scientific discovery using a broad array of applications, including both RNA and DNA applications, and promises to transform genomic,

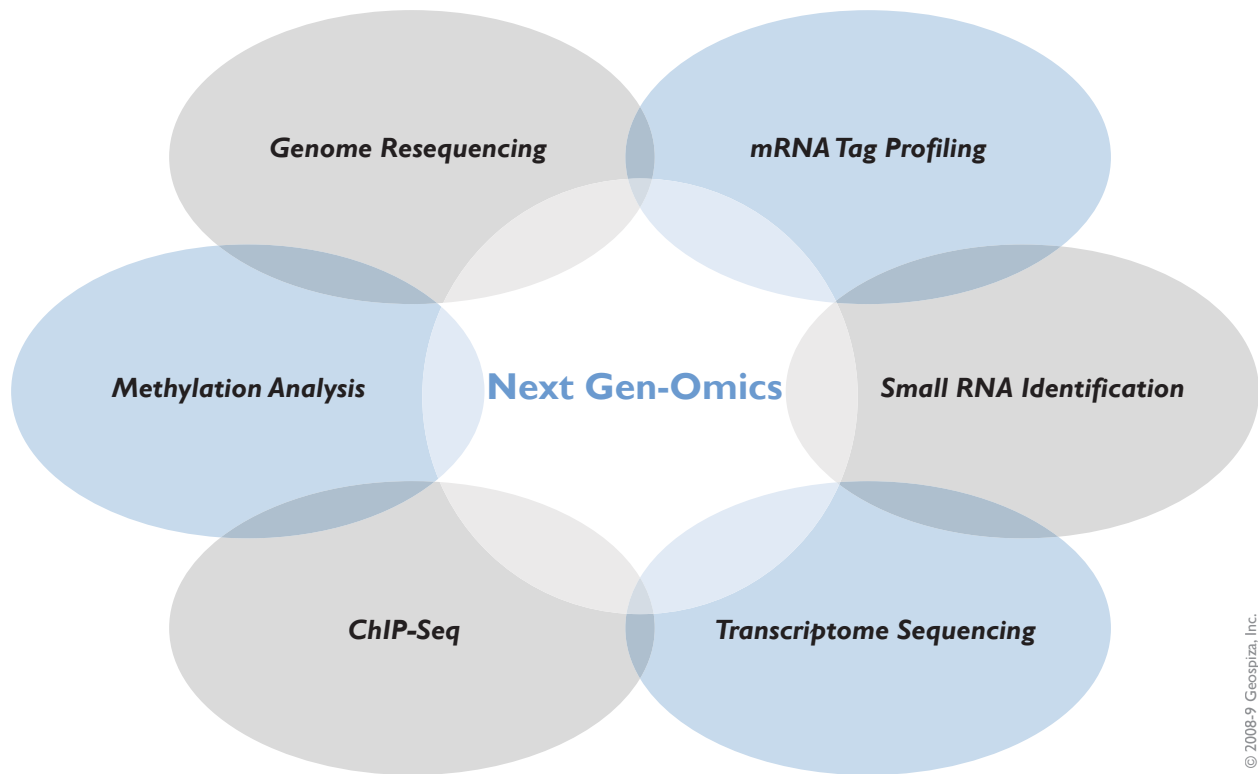
transcriptomic, and epigenomic research.³ Instruments such as the Genome Analyzer from Illumina, the SOLiD system from Applied Biosystems, and the Genome Sequencer from 454 Life Sciences allow individual labs or lab groups to produce data in quantities unthinkable outside a major genome center just a few years ago.

The rapid rate of NextGen technology development has forced researchers to rethink data management strategies from the ground up. Single instrument runs can produce terabytes of data; data analysis for a small group of samples may involve analyzing entire genomes or screening hundreds of millions of sequence reads; in a world of network computing, many labs have reverted to shipping data to researchers through the mail on portable hard drives. As one expert in the NextGen field recently stated, “Next generation sequencing has been, and continues to outpace, Moore’s law for computing. If you’re waiting for your computer to catch up with your data, you’re going to be waiting a long time.”⁴

Adding to this challenge is the diverse application portfolio supported by NextGen, and the recognition that the goal of NextGen sequencing is not necessarily the generation of DNA sequence; many applications, such as transcriptome analysis or digital gene expression, use sequence data as quantitative data points to measure other information⁵. These myriad applications require distinct laboratory workflows and correspondingly distinct bioinformatics pipelines to filter, assemble, organize, and visualize NextGen data in a way that provides meaningful biological information to researchers.

Traditional approaches to data management, storage, distribution, analysis and visualization are insufficient to manage the quantity and complexity of NextGen data.

¹Geospiza, Inc., 100 West Harrison, North Tower, Suite 330, Seattle, WA 98119
Correspondence should be addressed to Laura Lucas (laural@geospiza.com)
May 21, 2009



© 2008-9 Geospiza, Inc.

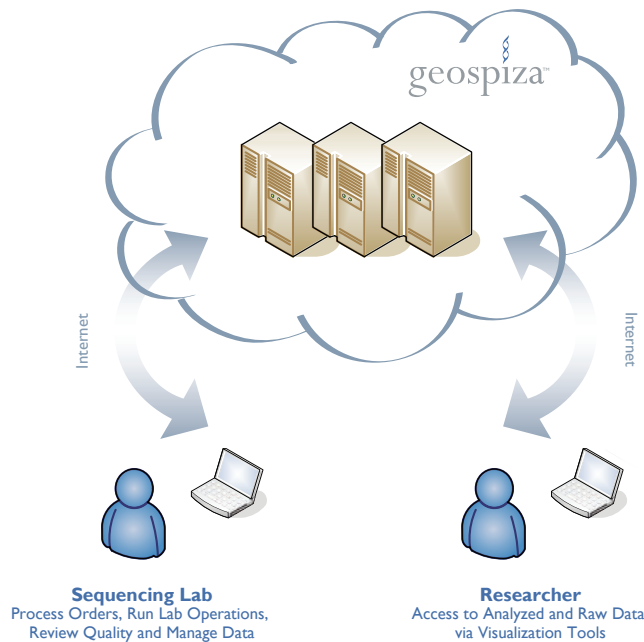
Figure 1: Next Generation sequencing facilitates many applications, including both DNA- and RNA-based applications. Many of these applications use DNA sequences as quantitative data points to measure other information.

Attacking the Problem

Geospiza recognized that NextGen technology would present an entirely new set of challenges surrounding information management and data analysis. Simply scaling current approaches with existing systems and software by orders of magnitude would be insufficient to meet these needs. A next generation system to address these challenges would need to satisfy several critical requirements:

- Allow biologists to concentrate on doing biology.
- Track experimentally relevant data from NextGen projects within a single unified system.
- Free labs and researchers from the burden of having to build and maintain the data centers required to support NextGen technology.
- Give bioinformaticians a framework in which to easily develop complex, application-specific data analysis pipelines that could be automated and made available to researchers.
- Give researchers the ability to rapidly translate massive amounts of raw NextGen data into well-characterized, biologically meaningful results.
- Facilitate efficient distribution, sharing and visualization of experimental results using universally accessible, platform independent tools.

Geospiza worked closely with leading researchers in academia, industry, and government to design and build a system that would be flexible enough to address the continuously evolving needs of NextGen technology while achieving the stated goals. The result was GeneSifter Lab Edition.



© 2008-9 Geospiza, Inc.

Figure 2: A centralized, internet-based data center provides IT infrastructure and system access to both labs and users.

Computing in the Cloud

A key design feature of GeneSifter is its ability to run as a cloud computing application.⁶ This is a critical shift from a traditional desktop or local server strategy, and provides distinct advantages to both researchers and laboratories.

Universal web access: All users interact with the system through a standard web interface, providing universal, platform independent access to the system from anywhere and at any time. Researchers, whether down the hall or across the continent, can initiate projects, track samples, and visualize and download data and analyzed results online. Labs can manage samples, track complex application-specific workflows, manage instrumentation, analyze, and distribute data in the same manner. This model is especially powerful for core labs or shared resource labs with distributed users; some current GeneSifter labs routinely support hundreds of researchers on multiple continents.

System architecture: Running GeneSifter as a web-based application decouples the need to build, run, and maintain a significant portion of the computing infrastructure required to support NextGen sequencing from the individual

laboratory. Servers and storage are located in a secure data center, and a single data center is capable of servicing the needs of numerous labs. The economy of scale provided by this model reduces the overhead costs for all labs, and industry standard security protocols insure data integrity. Additionally, processing and storage capacity can be quickly adjusted to meet the expanding needs of users. Although GeneSifter is optimized to run as a web-based application, it can also be deployed as a local service by organizations that want to engage in significant development work or connect to proprietary data sources.

End-to-End Workflow Management

GeneSifter tracks all phases of NextGen experiments from end-to-end. By spanning the entire continuum of the experimental process, data can be collected, stored, analyzed and permanently associated with samples within a single electronic lab notebook. It is easy to visualize this experimental continuum by breaking it into five distinct phases.

Sample entry/ordering: When new samples enter the lab they are entered into the system using web-based order forms. Forms can be customized to collect experimentally

relevant information for virtually any type of sample or experiment; forms are easily built using a web interface that requires no programming knowledge, reducing dependence on dedicated IT staff and eliminating drawn out development cycles associated with software. Data in order forms is dynamically linked to subsequent laboratory workflows that insure application-specific treatment of samples in the lab.

Laboratory workflow management: Different types of NextGen experiments require distinct application-specific laboratory workflows which may span a period of days, or even weeks, to complete. GeneSifter allows users to rapidly develop application-specific workflows using a web interface, and virtually any workflow with any number of distinct steps can be implemented. As samples progress through the lab, technicians record completed workflow steps to track progress and comply with laboratory SOPs. Workflow steps can be configured to record experimentally relevant information (such as QC data) either as user input or attached files, and data become permanently associated with samples in a relational database.

Instrument management: All current NextGen platforms have unique instrument configurations for sample processing. GeneSifter is instrument-aware and can

schedule runs on all platforms using instrument-specific vocabulary and physical sample arrangement. The underlying architecture of the system is such that additional instrument configurations can be added as platforms evolve. Recognizing that many NextGen labs will continue to use established methods for some applications, such as microarrays or capillary electrophoresis, the system supports major microarray and CE platforms as well.

Data storage and distribution: Following completion of an instrument run, experimentally useful data such as sequence reads and quality files are uploaded into the system using an automated file transfer tool which encrypts data for security. The system associates instrument results with all previously recorded lab data for each sample. Since GeneSifter is a web-based system data may be made available to researchers immediately, no matter where they are located. A researcher simply logs into their account in order to view or download data from their samples.

Data analysis: Raw sequence data is run through automated, application-specific analysis pipelines to produce well characterized, biologically meaningful data sets such as gene lists or variant reports. Researchers have immediate access to analyzed data and visualization tools. An in depth examination of data analysis follows.

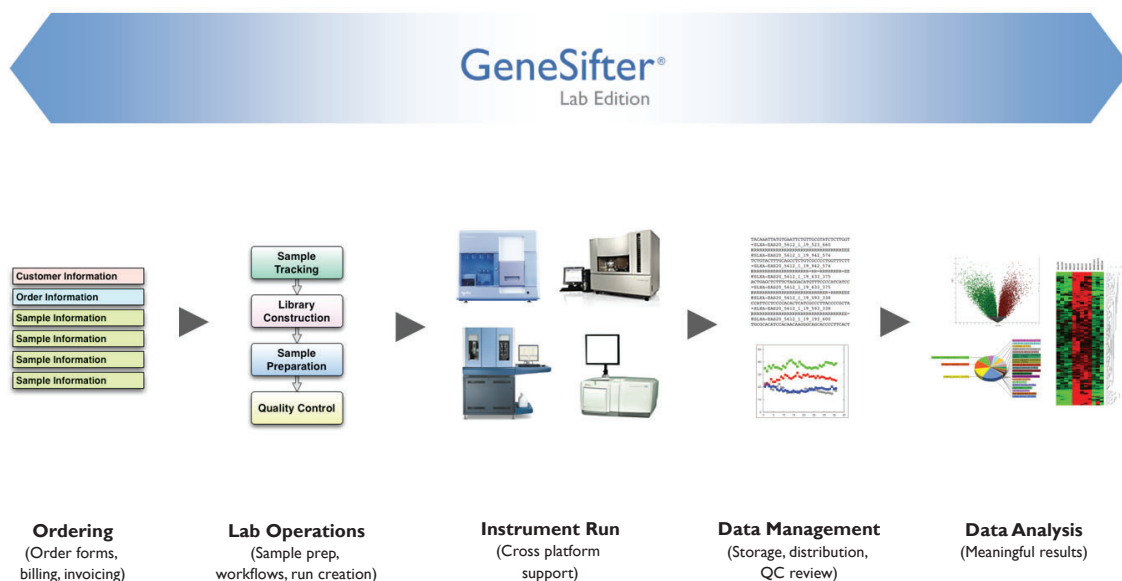


Figure 3: GeneSifter manages NextGen projects through the entire experimental continuum.

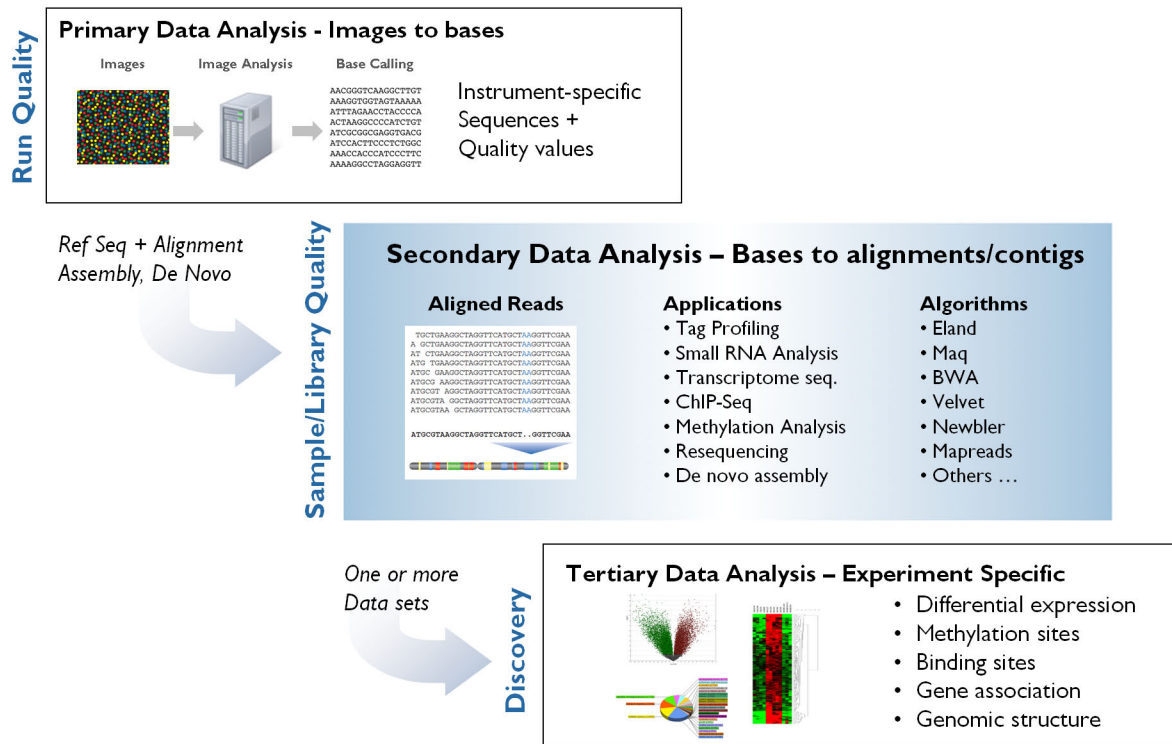


Figure 4: NextGen data analysis can be divided into Primary, Secondary, and Tertiary analysis.

NextGen Data Analysis

Data analysis is of particular importance given that it has rapidly become the rate limiting step for many NextGen projects. At a very basic level, all NextGen platforms provide application-agnostic data in the form of DNA sequences. While the volume of data is impressive, having millions of sequence reads is virtually useless without some meaningful characterization of what those sequences represent. Also, as noted previously, many NextGen applications actually use sequences as quantitative data points by measuring the frequency of their occurrence within the sample.

To better understand the process of NextGen analysis it is helpful to classify the process into three stages: primary, secondary, and tertiary, as highlighted above.

Primary analysis (instrument): Primary data analysis involves converting image data to sequence data. The sequence data can be in familiar “ACGT” sequence space or less familiar color space (SOLiD) or flow space (454). Primary data analysis is normally performed by instrument vendor software and is the first place where quality assessment of a sequencing run occurs.

Secondary analysis (application): Secondary data analysis converts primary data into well-characterized, biologically meaningful data sets – such as gene lists or variant reports – that will be further used to develop or test a scientific hypothesis. This step often involves aligning sequences from the primary analysis to reference data. Reference data can be complete genomes, subsets of genomic data such as expressed genes, or individual chromosomes.

Reference data are chosen in an application-specific manner and sometimes multiple reference data sets will be used in an iterative fashion.

Secondary analysis usually depends on complex, application-specific analysis pipelines that may include dozens of discrete steps. The ability to automate these pipelines is a key factor in quickly moving from raw data to publishable results.

Tertiary analysis (experiment): Tertiary analysis is the phase at which well-characterized data sets may be combined to compare and contrast results in order to reach an experimental conclusion. This may involve a simple activity such as viewing data in a tool like a genome browser so that the frequency of tags can be used to identify promoter sites, patterns of variation, or structural differences. In other experiments, like digital gene expression, it may involve comparing expression levels in a similar fashion to microarray experiments. The ability to perform tertiary analysis is directly correlated to a robust secondary analysis process.

NextGen presents a significant challenge in that most researchers are typically interested in performing tertiary analysis in order to answer a scientific question. Unfortunately, the process of converting primary data into a biologically meaningful data set is complex. Secondary analysis pipelines are often complicated, multi-step algorithms executed by overworked bioinformaticians. GeneSifter significantly mitigates this intermediate step by automating the process of secondary analysis.

Automated Analysis Pipelines:

The value of this automation is significant. A NextGen application pipeline is a multi-step process that converts millions of short reads into application-specific data summaries, reports and specific file formats necessary for additional downstream analysis. Through automation, bioinformaticians are able to make complex pipelines available to researchers without the burden of providing hands on support. Conversely, researchers have ready access to a multitude of pipelines independent of a bioinformatician.

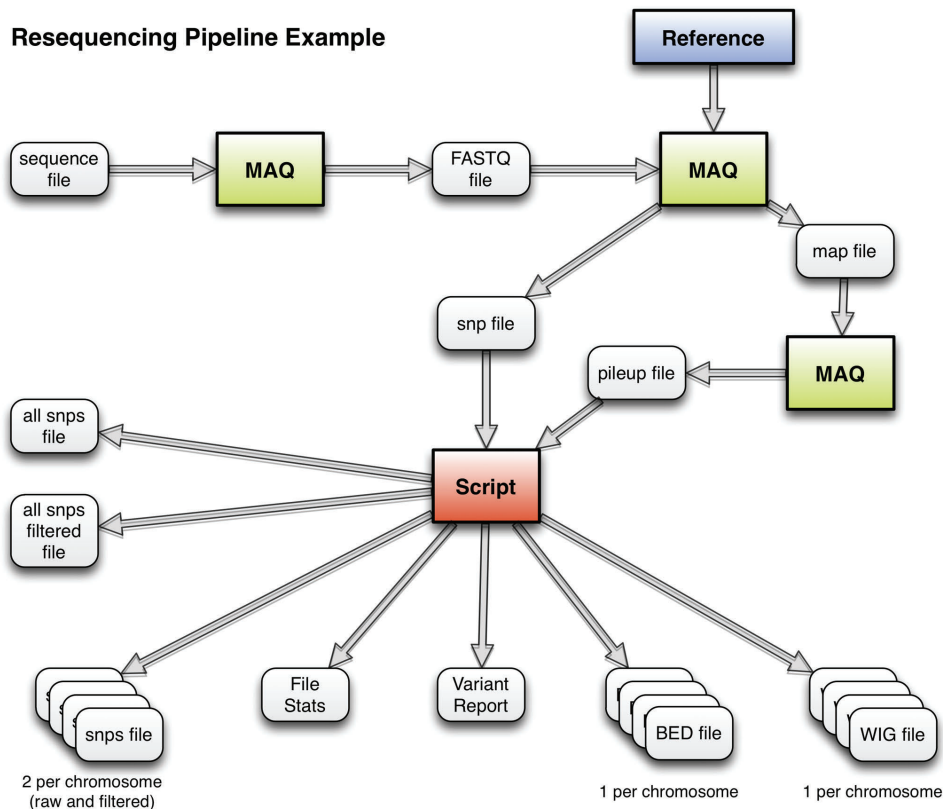


Figure 5: Example of an automated analysis pipeline for DNA resequencing applications. A researcher can run this entire pipeline to create a well characterized data set with the click of a button.

and reports frequently provide links to sources such as Entrez Gene¹¹, miRBase¹² or the UCSC Genome Browser¹³. Reports are always downloadable directly to a user's local computer, and when appropriate standard formats such as BED and WIG files are available as well.

Universal access to data via the internet presents distinct advantages to laboratories as well as to individual researchers. As data sets become more complex, and scientists increasingly depend on service or core labs to generate data, an efficient and automated mechanism for distribution and analysis of data is a key requirement for continued success.

Return on Investment

In addition to the overall efficiency and synergy achieved through using GeneSifter, a demonstrable ROI can be measured:

IT infrastructure: The economy of scale of a hosted/cloud-based IT infrastructure significantly lowers capital and operational costs for most labs, often by tens of thousands of dollars per year, and sometimes more. Since additional capacity can be added in real-time as needed, a hosted infrastructure also eliminates costly under and over planning mistakes, and labs do not have to manage a farm of rapidly depreciating assets. Additionally, time to get up and running on a hosted system is measured in days rather than weeks or months, allowing labs to immediately focus on producing results. Web-based access to all system features permits rapid access to and sharing of data between geographically separated groups.

Automated data analysis: Bioinformatics support for many NextGen projects has proven to be an intensive, hands-on process. The ability to automate complex analysis pipelines frees up bioinformaticians to spend time developing additional tools for researchers, while allowing researchers to leverage existing pipelines more easily. Depending on the size of a lab, this automation has the potential to free up many thousands of dollars per year in bioinformatics resources and personnel, and has the added benefit of delivering

data shortly after a run is complete rather than after a wait of many weeks. The ability to share pipelines with other members of the research community encourages collaboration and reduces duplicated effort between groups.

Conclusion

NextGen technology is accelerating the pace of discovery and revolutionizing life science research. This rapid change has brought about a fundamental shift in the way that biologists will approach information management.

The strategy of scaling existing data analysis approaches is not sufficient to meet NextGen challenges. With the exception of a few large research centers, NextGen researchers will increasingly rely on cloud computing networks and applications to track, store, analyze, and share research data. GeneSifter provides this capability in an integrated, universally accessible environment, allowing scientists to focus on doing science rather than constantly reinventing tools and infrastructure to support science.

As a cloud-based system, GeneSifter can adapt in real time to meet the expanding needs of users. Data center infrastructure can be updated in anticipation of researcher demand, and software updates can be applied system-wide for all users in a single process. Since all interaction with the system is web based there are no desktop clients to update or synchronize, and users have immediate access to the most up to date tools possible. All these processes are transparent to end users, permitting them to focus on their research instead of their IT infrastructure.

GeneSifter has the flexibility to evolve as rapidly as the technology it supports. As novel applications emerge, new workflows and analysis pipelines can be easily created by labs and bioinformaticians, and bioinformaticians can share pipelines with the research community to reduce duplicated effort. Access to automated analysis pipelines allows researchers to rapidly translate raw data into well characterized data sets – the Achilles heel of NextGen analysis – and quickly proceed to testing their scientific hypothesis, ultimately speeding up time to publication.

Appendix

Supported Platforms

- Applied Biosystems: Capillary Electrophoresis and SOLiD Sequencers
- Illumina: Genome Analyzer and iScan System for BeadArrays
- Roche: Genome Sequencer FLXSystem and NimbleGen Array System
- Affymetrix: GeneChip Microarray System
- Agilent: DNA Microarray Platform
- Helicos: Single Molecule Sequencer

Supported Applications:

- Digital Gene Expression
- mRNA Expression
- RNA-Seq
- Small RNA
- Tag Profiling
- ChIP-Seq
- Resequencing
- Variant Analysis
- Mutation Discovery

References

1. 2008. Prepare for the deluge. *Nat. Biotechnol.* 10, 1099.
2. 2008. Byte-ing off more than you can chew. *Nat. Methods* 5, 577.
3. Wold B., Myers R.M., 2008. Sequence census methods for functional genomics. *Nat. Methods* 5, 19-21.
4. Michael Rhodes, speaking at NIH, 10 Dec., 2008.
5. Wang Z., Gerstein M., Snyder M., 2008. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, doi:10.1038/nrg2484.
6. Bell G., Hey T., Szalay A., 2009. Computer science. Beyond the data deluge. *Science* 323, 1297-1298.
7. UCSC, "UCSC Genome Bioinformatics Home", <http://genome.ucsc.edu/>
8. Li H., Ruan J., Durbin R., 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 18, 1851-1858.
9. Langmead B., Trapnell C., Pop M., Salzberg S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25.
10. Applied Biosystems, "SOLiD System Color Space Mapping Tool (mapreads)", <http://solidsoftwaretools.com/gf/project/mapreads/>
11. NCBI, "Entrez Gene Home", <http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene>
12. Griffiths-Jones S., Sain: H.K., van Dongen S., Enright A.J., 2008. miRBase: tools for microRNA genomics. *Nucleic Acides Res.* 36, D154-D158.
13. Kent W.J., Sugnet C.W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M., Haussler D., 2002. The human genome browser at UCSC. *Genome Res.* 12, 996-1006. <http://genome.ucsc.edu/>