The SOLiD™ Software Suite Data Analysis and Management

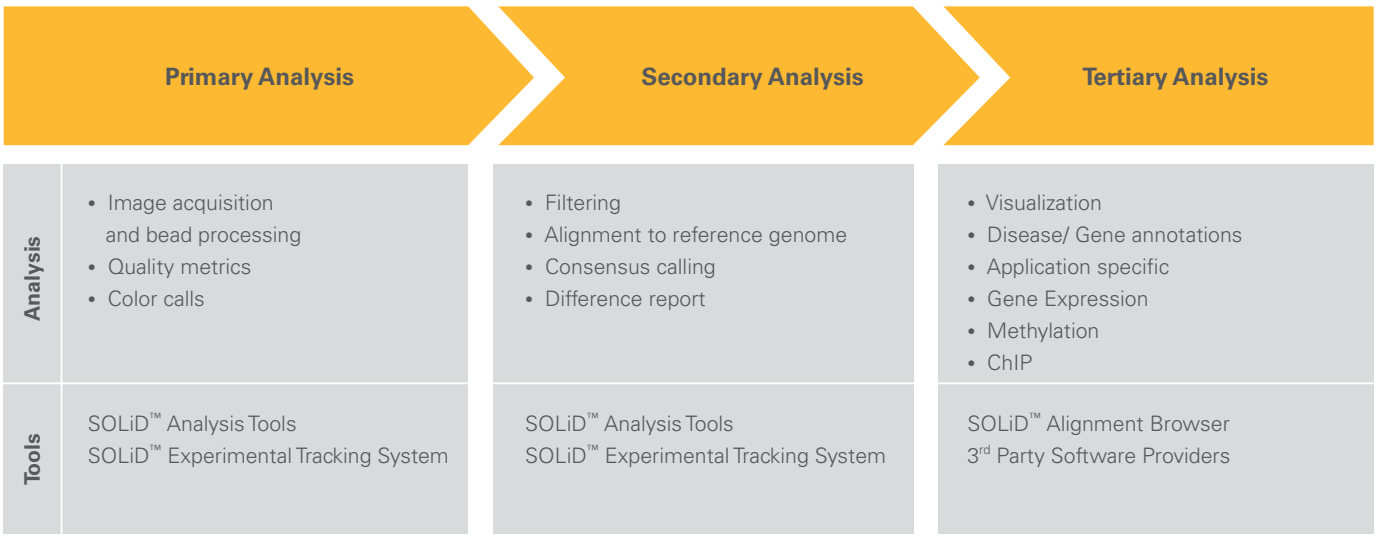

Figure 1. Data analysis may be segmented into primary, secondary and tertiary analysis. Primary and secondary analyses include universal processes for data generation, collection and processing. Tertiary analyses include application specific clustering, parsing and visualization tools.

Introduction

Next generation sequencing technologies are expanding the boundaries of traditional genetic analysis and enabling new applications such as whole genome resequencing, hypothesis neutral gene expression, ChIP and methylation analysis. The average next generation experiment will generate terabytes of information making data analysis and management critical to the success of any project. The SOLiD™ System comes complete with the necessary computing power and analysis software to complete primary (image acquisition and quality control) and secondary (alignment to a reference genome, base calling, and SNP identification) analysis of fragment and mate-paired experiments.

SOLiD Software Suite

Applied Biosystems provides a suite of analysis and management applications for data generated using the SOLiD System (Figure 1). The SOLiD Analysis Tools (SAT) process the array image, perform data filtering, calculate quality values, align to a reference genome, and generate base calls. The SOLiD Experiment Tracking System (SETS) is a web-based integrated application that enables real-time remote monitoring and visualization of analysis reports. The SOLiD Alignment Browser (SAB), which is based on an open source platform, allows visualization of sequences aligned to a reference genome. All three software applications are provided as part of the SOLiD System in addition to a compute cluster with sufficient power to support real-time data analysis on the instrument.

SOLiD Analysis Tools

The SOLiD Analysis Tools generate three types of data – image data, primary analysis data and secondary analysis data. Primary analysis data consists of image alignment, color and quality value (QV) calls, and an intensity record of each color channel. Some files may be further transformed into various reports (e.g., a quality value report is generated from the quality value file which can then be viewed in SETS).

During secondary analysis, a reference sequence is converted into color space, and the data are aligned to the reference sequence. A consensus sequence may then be constructed from the sequencing reads. Comparison of the consensus sequence to a reference genome enables the identification of SNPs and structural variations.

Results from SAT can be easily transferred to the tertiary analysis tools. Customers start with reference sequences in base space and the final output of the SOLiD System is a consensus sequence file and a differences (SNP) file relative to the base space reference provided by the customer. These files and other reports are accessible via SETS or the Linux file system.

SOLiD Experiment Tracking System

The SOLiD™ Experiment Tracking System is a web-based application that enables users to view real-time data and completed run analysis reports from the SOLiD Analysis Tools. SETS provides an easy to navigate interface (Figure 2) that allows laboratory scientists to track the progress of their experiments and to modify the analysis settings.

A unique feature of SETS is that users can remotely observe a run in-progress by using any available RSS feed reader, such as Internet Explorer or Firefox. Users can monitor a run on the office desktop or offsite. Remote monitoring enables users to be alerted to issues in real-time which reduces unnecessary downtime and reagent waste. Detailed information regarding various data files and reports can be found in the SETS Getting Started Guide.

SOLiD Alignment Browser

Applied Biosystems developed the SOLiD Alignment Browser to visually display color space reads aligned to a reference genome (Figure 3). SAB is based on Apollo, an open source platform which allows easy data manipulations and tertiary analysis tool development. SAB has been demonstrated to run on Windows, Mac and Linux systems.

Tertiary Analysis – Application-specific Analysis Tools

The SOLiD System is an open platform with respect to downstream analysis solutions. The SETS output files listed above use standard formats to facilitate analysis in commonly used application specific tools. Applied Biosystems is enabling the development of new analysis solutions through the SOLiD Software Community Program. All the analysis tools provided are open source so the analysis pipeline can be customized by the user.

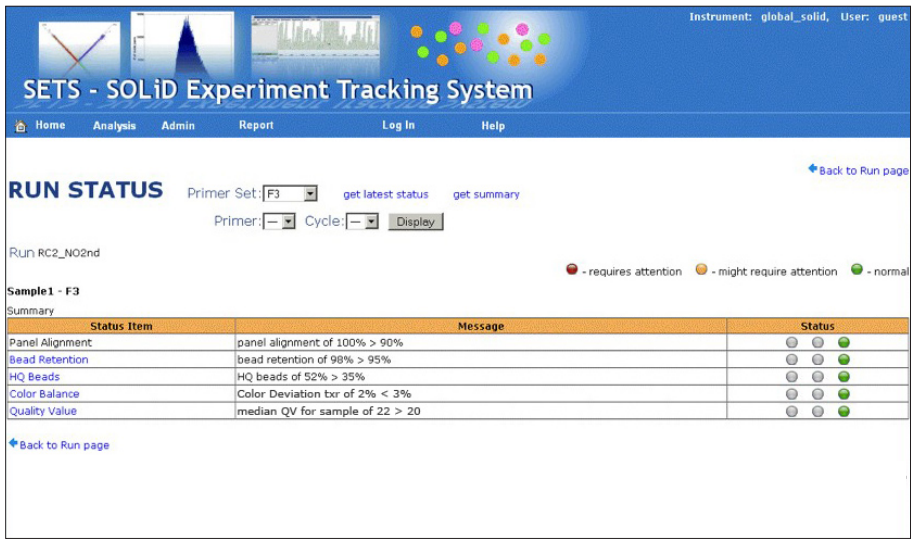

Figure 2. SOLiD™ Experiment Tracking System (SETS) provides an interface to view results from in-progress or recently completed analyses. Users can also modify analysis settings and manage user profiles.

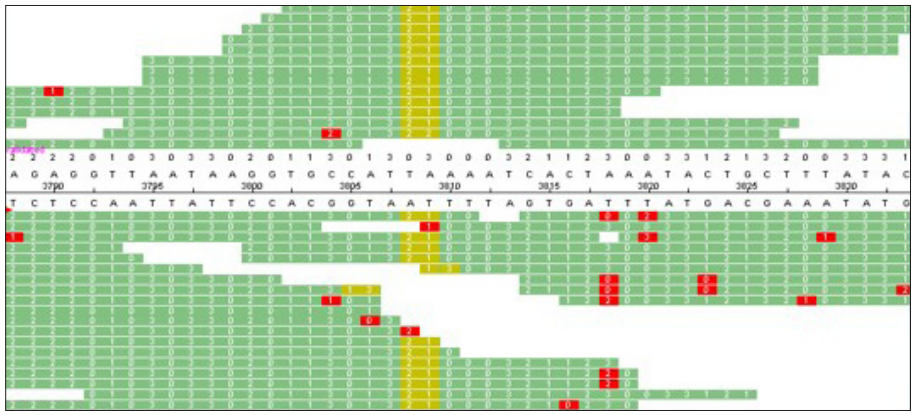

Figure 3. A snapshot of sequencing reads aligned in the SOLiD™ Alignment Browser (SAB). 0, 1, 2, and 3 represent the four colors used in 2-base encoding. Positions matching the color space reference are represented in green. Single color space mismatches are shown in red and adjacent color space mismatches are represented in yellow.

Post SOLiD System Data Analysis/ Management Recommendations

The enormous volume of data generated in a next generation experiment precludes long term storage on the SOLiD System. It is recommended that SOLiD image data be kept only until the completion of primary analysis. Applied Biosystems recommends the following hardware for additional analysis after transferring the data from the SOLiD System.

Recommended Hardware for Post SOLiD System Analyses

1. 4 rack-mounted computers, each computer with dual quad-core CPUs, 8GB RAM, 160 GB local hard drive, and 1GbE NICs
2. Shared disk system with 9TB RAID5 direct-attached storage
3. 1GbE network switch

TABLE 1. Average file sizes for various analyses

	Image Data Size†	Primary Analysis Data*	Secondary Analysis Data
1 slide - 1 tag (fragment library)	1.8 - 2.4 TB	< 100 GB	10-20 GB
1 slide - 2 tags (mate-paired library)	3.5 – 4.0 TB	< 100 GB	30-40 GB
2 slides - 1 tags (fragment library)	5.0 TB	< 100 GB	20-40 GB
2 slides - 2 tags (mate-paired library)	8.2 TB	< 100 GB	60-80 GB

Average file sizes for various analyses under the following assumptions: 7 ligation cycles for each sequencing primer, 4 images per cycle, 1 for each dye. A full slide contains more than 2,000 panels.
† There is no need to store Image Data.
* The size displayed here includes all primary analysis files.

TABLE 2. Data output files

	File name	.ext	File content
primary analysis files	Raw reads file	.csfasta	Info on calls in color space
	QV quality value file	_QV.qual	Quality value for each color space sequenced
	Reads summary file	.stats	Statistics summarizing the number of reads collection in each panel on a slide
	Scaled Intensity Value Files (optional)	_intensity.scaled[CY3[CY3|CY5|FTC|TXR].fasta	Intensity values
secondary analysis files	Consensus sequence file	.fasta	Consensus sequence in base space
	Differences file	.txt	Locations where the sequence data show discrepancies to the ref sequence
	Matching file	.csfasta.ma.#.#	Sequence data mapped back to the ref sequence in color space
	Gff or gff.2 file	.gff or .gff.2	Sequence data mapped back to the ref sequence in .gff format, reads in color space and info regarding mismatches and quality values.

4. Power for the cluster with a power distribution unit (PDU) and uninterruptible power supply (UPS)
5. Linux operating system

Primary analysis data may be transferred from the SOLiD System to an external USB drive. Backing up the necessary primary analysis data files should take less than 2 hours (assume USB 2.0 transfer speed at 10-20MB/second). The combined size of these files is less than 100GB (Table 1). Alternatively, all of data are also available in a binary file format (.spcf) and the combined size of these files should also be less than 100GB. Secondary analysis results may also be transferred onto an external USB drive or through a network onto another computer server (a hard drive is recommended). Above is a list of all output file types provided by the SOLiD Software Suite (Table 2).

Conclusion

The SOLiD Analysis Tools and the SOLiD Experiment Tracking System allow researchers to acquire and process data in real-time while the SOLiD Alignment Browser allows further visualization and discrimination of experimental results without the need for additional computing power. The SOLiD Software Suite allows maximum throughput and flexibility for all researchers' needs and provides a robust and efficient way for scientists to analyze results for applications such as targeted resequencing, gene expression, microRNA discovery, ChIP and whole genome sequencing studies.

Printed in the USA, 1/2008, Publication 139AP08-01

Headquarters
850 Lincoln Centre Drive | Foster City, CA 94404 USA
Phone 650.638.5800 | Toll Free 800.345.5224
www.appliedbiosystems.com

International Sales
For our office locations please call the division headquarters or refer to our Web site at
www.appliedbiosystems.com/about/offices.cfm