SOLiD™ Software— Data Analysis and Management

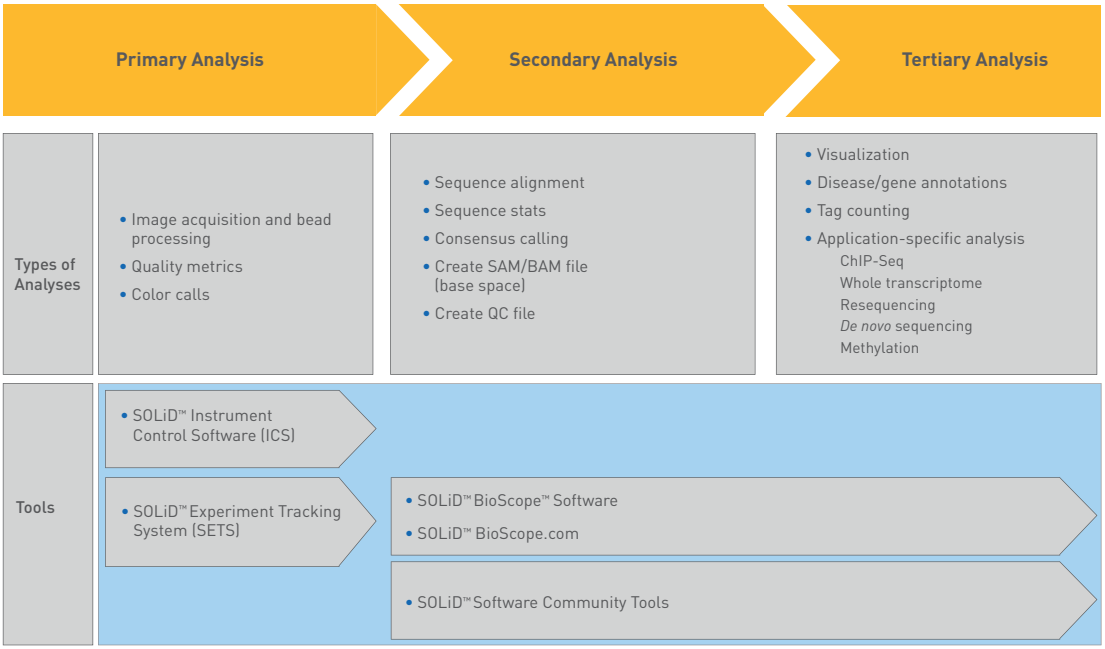

Figure 1. Data analysis may be segmented into primary, secondary, and tertiary analysis. Primary analysis includes universal processes for data generation, collection, and raw processing. Both secondary and tertiary analyses are application-specific. Secondary analysis processes application-specific data at the sequence level, while tertiary analysis generates biological interpretation specific to an application.

Introduction

The average next-generation sequencing experiment generates terabytes of information, making data analysis and management critical to the success of any project. Software for the SOLiD™ System platform is composed of tools for complete primary, secondary, and tertiary analysis, which include image acquisition, quality control, base calling, and alignment to a reference genome. For additional biological interpretation tools, the SOLiD™ Software Community Program provides access to a growing number of academic and commercial software options for SOLiD™ sequencing analysis.

SOLiD™ Software

A suite of Applied Biosystems® analysis and management applications are available for data generated using the SOLiD™ System (Figure 1). The SOLiD™ Instrument Control Software (ICS) provides easy-to-use, walk-away operation and data analysis.

The SOLiD™ Experiment Tracking System (SETS) is a web-based, integrated application that enables real-time remote monitoring, visualization of analysis reports, and the ability to reanalyze data. SETS allows sequencing results to be exported automatically to an off-instrument compute cluster for further analysis. SOLiD™ BioScope™ Software provides a validated single framework, which enables scalable and flexible application support and supports mapping, resequencing, and whole-transcriptome pipelines.

SOLiD™ Instrument Control Software (ICS)

ICS provides automated, walk-away operation and data analysis, an easy-to-use graphical user interface, and a guided wizard program for easy experimental setup.

SOLiD™ Experiment Tracking System (SETS)

SETS is a web-based, easy-to-navigate interface that enables users to view real-time data and run analysis reports from the SOLiD™ Analysis Tools (Figure 2).

A unique feature of SETS is the ability to remotely observe a run in progress by using any available Internet browser, and receive email notification on system information. Remote monitoring enables users to be alerted to issues in real time, which helps reduce unnecessary downtime and reagent waste. Detailed information regarding various data files and reports can be found in the *SETS Getting Started Guide*.

ICS/SETS

ICS/SETS generates primary and secondary analysis data and automatically processes the image during the instrument run, performs data filtering, calculates quality values, and generates base calls. Primary analysis data consist 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).

SOLiD™ BioScope™ Software

This software provides a simple Web interface for generating instructions for running application-specific sequence analysis tools. The BioScope™ Software framework enables the user to perform off-instrument secondary and tertiary analyses, and it allows configurable bioinformatics workflows for resequencing (mapping, SNP-finding (DiBayes), copy number variations, inversions, small indels, large indels), whole-transcriptome analysis (mapping, fusion splicing, counting, and UCSC WIG Files creation), and ChIP-Seq (mapping). Data outputs can be seamlessly visualized in browsers such as UCSC or Ensembl Genome Browsers and the Broad Institute Integrative Genomics Viewer (IGV).

SOLiD™ BioScope.com

The new SOLiD™ BioScope.com solution is a cloud computing solution for users of SOLiD™ System data who are looking for an alternative to buying and maintaining the compute infrastructure typically required for NGS data analysis. With our

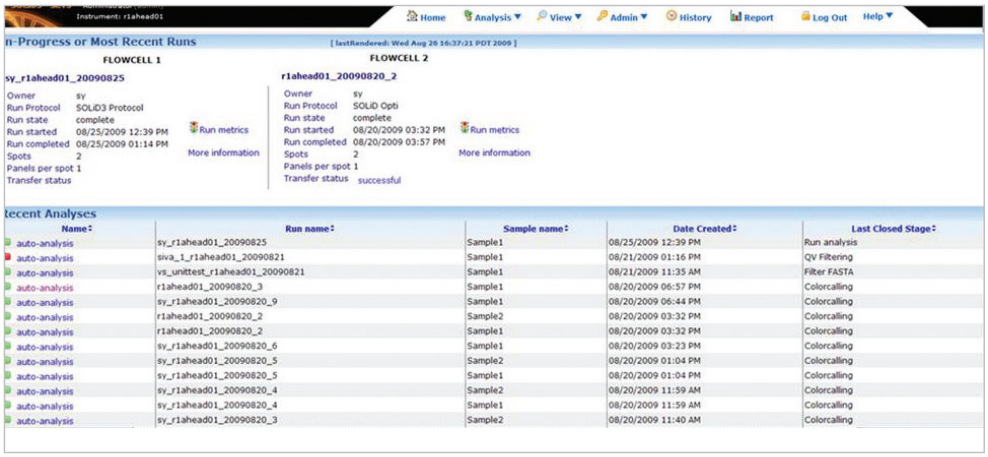

Figure 2. The SOLiD™ Experiment Tracking System (SETS) provides an interface to view results from in-progress or recently completed analyses. Users can also modify analysis settings and manage user profiles.

Figure 3. Visualization of SOLiD™ System sequencing reads using the Broad Institute Integrative Genomics Viewer. The 1000 Genomes Project public data release for high-throughput sequencing of NA19240 (Chr16:69,746,167-69,746,371) is shown, visualized using the Integrative Genomics Viewer, a high-performance visualization tool from the Broad Institute. Coverage plots and aligned read for the SOLiD™ System are depicted.

SOLiD™ BioScope.com solution, your compute costs are based on the amount of CPU time required for your SOLiD™ System data analysis, allowing you to leverage virtually limitless compute resources to temporarily accommodate increased demand (www.SOLiDBioScope.com).

SOLiD™ Software Community

The SOLiD™ Software Community supports life scientists and independent software vendors in the development and potential commercialization of bioinformatics applications for next-generation DNA sequencing platforms.

As part of this initiative, Applied Biosystems has included new sample data sets and open-source software tools for the SOLiD™ System at info.appliedbiosystems.com/solidsoftwarecommunity. The goal of this program is to directly address the challenges associated with analyzing and managing the vast amounts of research data generated by this ultrahigh-throughput sequencing technology.

Post-SOLiD™ Data Analysis

The enormous volume of data generated in a next-generation experiment precludes long-term storage on the SOLiD™ System. For both data storage and application-specific downstream analysis, an off-instrument solution is recommended. The combined size of the primary analysis flat file results is about 600 GB per slide run (Table 1). Table 2 shows all output file types provided by SOLiD™ Software.

Conclusion

ICS and SETS allow researchers to acquire and process data in real time. SOLiD™ BioScope™ Software provides a robust and efficient way for scientists to run workflows for such applications as whole-genome and targeted resequencing, as well as ChIP-Seq.

Table 1. Average file sizes for various analyses.

	Image Data Size*	Primary Analysis Data†	Secondary Analysis Results in BAM Format‡
1 slide—tag (fragment 50 bp)	1.9 TB	0.8 TB	0.6 TB

Average file sizes for various analyses under the following assumptions: 10 ligation cycles for each sequencing primer, 4 images per cycle, 1 for each dye. A full slide contains more than 2,350 panels.
* Minimum space needed for images.
† Minimum space needed for primary analysis results (spch, csfasta, QV.qual).
‡ The size of the analysis result correlates directly with the throughput.

Table 2. Data output files.

	File Type/Format	File Name Extension	File Content
Primary Analysis Files	Raw reads file	.csfasta	Color space reads
	QV quality value file	–QV.qual	Quality value for each color space sequenced
	Reads summary file	.stats	Statistics summarizing the number of reads collected in each panel on a slide
	Scaled Intensity Values File (optional)	–intensity.scaled [CY3|CY5|FTC|TXR].fasta	Color space reads
Secondary Analysis Files	Mapping file	.csfasta.ma	Sequence data mapped back to the reference sequence with quality values
	BAM	.bam	BAM (Binary Alignment Map) format is a generic format for storing large numbers of nucleotide sequence alignments

Publication CO12820 0510

applied biosystems™
by life technologies™

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 website at
www.appliedbiosystems.com/about/offices.cfm