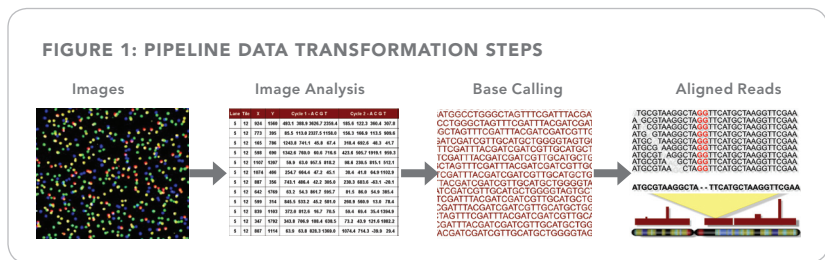# Genome Analyzer Data Analysis Software

Illumina has created a robust set of software tools to support the massive output of the Genome Analyzer. These tools provide an end-to-end solution from imaging and base calling to the analysis and visual representation of biologically relevant data.

INTRODUCTION

The massive quantity of sequence output by the Genome Analyzer has rendered traditional sequence analysis tools obsolete. In light of these challenges, Illumina has developed a software portfolio that rapidly takes users from raw data acquisition to publishable biologically meaningful results. This portfolio consists of the Analysis Pipeline and BeadStudio analysis software. Both packages represent customizable and comprehensive solutions for performing rigorous primary analysis. The graphical interface is intuitive for use by non-informaticians.

The Analysis Pipeline is responsible for performing primary data acquisition, determining base calls and confidence scores from the fluorescent signals on the Genome

HIGHLIGHTS OF SEQUENCING ANALYSIS SOFTWARE

- **Comprehensive**: End-to-end solution for sequencing data analysis

- **Powerful**: Customizable to meet the needs of any analysis workflow

- **Intuitive**: Graphical, easy-to-use interface

- **Extensible**: BeadStudio is home to a growing number of modules for specific applications

FIGURE 1: PIPELINE DATA TRANSFORMATION STEPS

Analyzer. Higher level analyses, such as identifying genome-wide binding sites in a ChIP-Seq experiment, are performed using modules within Illumina's BeadStudio analysis software. Both of these packages are described in more detail below.

GENOME ANALYZER PIPELINE

Illumina Pipeline software is an automated engine for transforming primary imaging output from the Genome Analyzer into discrete aligned strings of bases. A package of integrated algorithms perform the core primary data transformation steps: image analysis, intensity scoring, base calling, and alignment (*Figure 1*).

The Pipeline software runs on a standard Linux workstation, supporting the use of additional downstream informatics tools. The entire Pipeline is run in a scripted mode to easily create aligned sequence output from a run. Advanced users can customize the scripts to meet the needs of specific experimental designs.

Image Analysis

The first step in the primary analysis is interpreting the image data to identify distinct clusters and create digital intensity files describing the signal intensities of each cluster in each cycle. To ensure that the maximum number of clusters are used to generate the maximum amount of sequence from a given run, a sophisticated cluster identification algorithm is needed. Using algorithms inspired by those that astronomers use to identify stars, the Pipeline extracts tens of thousands of clusters per image.

Base Calling

Signal intensity profiles for each cluster are used to call bases. Determining the quality of each base call is crucial for downstream analysis. Confidence scores for each call are calculated by an implementation of a logarithmic Q value, based on several distinct predictors of data quality.

illumina®

Alignment

The entire set of called sequence reads are aligned to create a final sequence output file. The extremely fast algorithm for the alignment of Genome Analyzer data is optimized for SNP identification. Easy-to-interpret confidence scores are determined for all alignments. Unalignable reads are flagged for further investigation.

Aligning reads from a paired-end sequencing run incorporates the empirical measurement of median insert sizes to rank loci for read pairs that have more than one sequence match. The ability to identify strongly aligning fragments with aberrant insert sizes is useful for identifying possible chromosomal rearrangement.

BEADSTUDIO ANALYSIS SOFTWARE

BeadStudio analysis software supports secondary analysis with easy-to-use tools, an intuitive graphical interface, and graphical data analysis output. BeadStudio is easily extended by adding different modules and plug-ins. The ChIP-Seq Module is the first in a series of BeadStudio modules designed specifically to support Genome Analyzer sequencing applications.

FIGURE 2: ILLUMINA CHIP-SEQ ANALYSIS SOFTWARE

Multiple ChIP-Seq reads aligned to a small genomic region on the BeadStudio Genome Browser display show a clear binding event.

Genome Browser Display

The Pipeline completes the primary data analysis phase by providing sequence information in a graphical format. At this point, the sequence data is ready for use in traditional and novel informatics analysis tools, such as Illumina's BeadStudio ChIP-Seq Module.

ChIP-Seq Module

The ChIP-Seq (CS) Module allows users to derive biological meaning from aligned sequence reads in a chromatin immunoprecipitate sequencing experiment. Sequence data are readily transferred from Pipeline output into the CS module. A wizard interface facilitates fast secondary analysis.

Binding site events are called by an algorithm that detects peaks in the census count of individual sequence reads across the entire genome. Sites where the factor of interest was likely bound to DNA are those that are overrepresented in experimental samples (*Figure 2*).

The CS Module takes advantage of the existing BeadStudio framework to provide data output in tabular form or on the graphical genome browser for quick and easy scanning of results. Nearby genomic elements are also displayed in the integrated Illumina Genome Viewer to facilitate result interpretation.

CONCLUSION

Illumina provides a robust software portfolio to support the revolutionary Genome Analyzer. Data produced with Illumina Pipeline software are easily imported into other analysis tools for SNP discovery, gene expression studies, and newly emerging applications. Novel secondary analysis modules continue to emphasize the strength of the Genome Analyzer system in many applications beyond sequencing genomic DNA fragments. In concert, the Genome Analyzer, informatics hardware, and analysis software represent a true genome center in a box.

ADDITIONAL INFORMATION

Visit our website or contact us at the address below to learn more about Illumina Sequencing applications, systems, and software.

Illumina, Inc.
Customer Solutions
9885 Towne Centre Drive
San Diego, CA 92121-1975
1.800.809.4566 (toll free)
1.858.202.4566 (outside North America)
techsupport@illumina.com
www.illumina.com
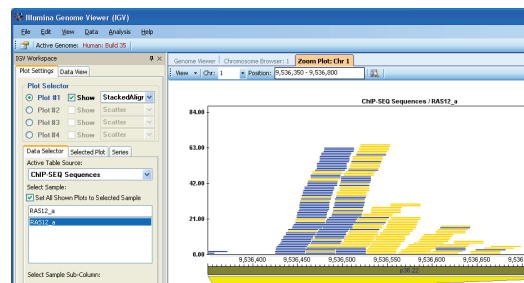

TABLE 1: HARDWARE REQUIREMENTS

Pipeline (Recommended)
- Linux OS (Red Hat recommended)
- 8 processors (or 4 dual-core)
- 32 GB RAM
- > 4 TB disk storage

ChIP-Seq Module (Required)
- Windows 64-bit OS
- > 4 GB RAM
- 30 GB free disk storage

FOR RESEARCH USE ONLY

illumina®