GenomeStudio® Data Analysis Software

Illumina has created a comprehensive suite of data analysis tools to support a wide range of genetic analysis assays. This single software package provides data visualization and results analysis for all Illumina assay platforms.

GenomeStudio Software Highlights

- **Broad Set of Tools**
  Analyze data generated from all Illumina platforms, sequencing and array.
- **Integrated Analysis**
  Combine data from more than one assay type in the same project.
- **Sophisticated Platform**
  Use high-performance algorithms and statistics calculations for a wide range of supported applications.
- **Open Architecture**
  Leverage plug-ins and application programming interfaces (APIs) to facilitate data export for secondary or tertiary analysis.

Introduction

Illumina leads the industry with a broad spectrum of innovative and powerful genetic assays. The Illumina GenomeStudio Data Analysis Software supports researchers' equally diverse needs for data analysis. Scientists using any of Illumina's platforms—Genome Analyzer™, MiSeq®, or HiSeq® 2500 systems for next-generation sequencing; HiScan® or iScan System for BeadArray technology—can use the highly visual and intuitive GenomeStudio software for primary analysis of data generated with Illumina assays (Figure 1).

Integrated Framework

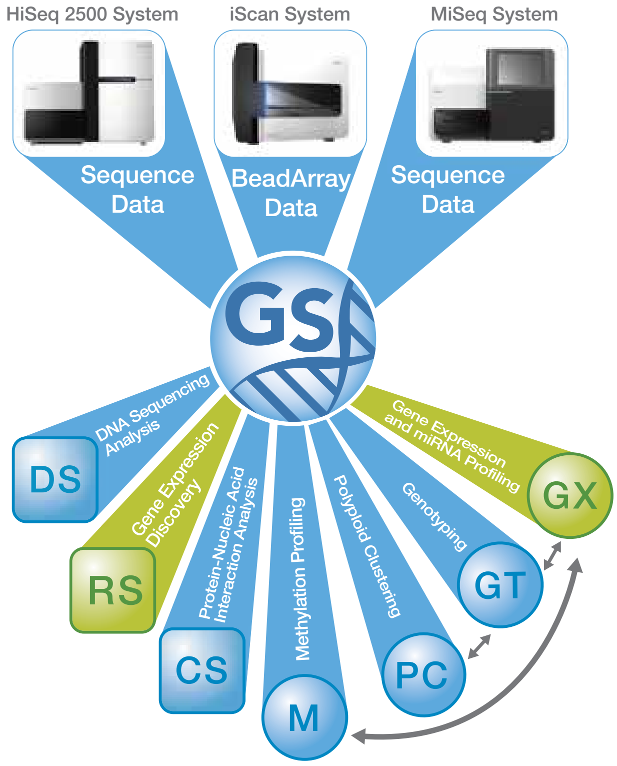
GenomeStudio software consists of several assay-specific modules, integrated into a single platform. The common framework provides a set of intuitive graphical user interface (GUI) and data visualization features for the control and display of results generated by individual modules. Thorough understanding of the massive amounts of data generated by Illumina assays may require multi-modal examination and integration of information gleaned from a 10,000-foot view all the way down to a fine-grained single-feature view. GenomeStudio framework displays results at all scales to enable researchers to effectively examine high-resolution genome-wide data.

Project Creation Wizard

GenomeStudio software makes it easy to create new projects and identify input data locations with an intuitive wizard interface. Once a project is created, users can easily visualize and display all data associated with each experiment type.

Users with the optional Infinium® LIMS, sample tracking, and robotic automation control can take advantage of GenomeStudio software integration for increased efficiency in overall project management for all Infinium genotyping assays. These systems, custom designed for Illumina workflows, allow labs to maximize their throughput.

Figure 1: GenomeStudio Software for Integrated Analysis of Data from All Illumina Platforms



Consists of several modules (bottom) that support the analysis of data generated using any Illumina platforms (top). Data from separate modules can be combined into a single project (grey arrows).

Data Visualization

Genome-wide orientation and broad trends can be quickly seen when data are displayed in the Illumina Genome Viewer (IGV). Chromosome- or region-level trends of sequence or array data are visualized in the Illumina Chromosome Browser (ICB). The ICB can be used to identify structural aberrants, gene expression levels, protein binding sites, or methylation of CpG islands in promoter regions. For higher resolution analysis, particularly with sequencing data, researchers can zoom in to

Table 1: GenomeStudio Display Options

Global Visualization
Illumina Genome Viewer (IGV)
Illumina Chromosome Browser (ICB)
Illumina Sequencing Viewer (ISV)
Graphs
Dendrograms and Clustering Analysis
Heat Maps
Scatter Plots
Histograms
Line Graphs
Box Plots
Frequency Plots
Pie Charts
Tables
Samples Table
Sequence or Lane Tables
SNP Table
Alleles Table
Probe Table
Gene, Exon, or Junction Tables
Plus other assay results tables

see single base calls in the Illumina Sequence Viewer (ISV) to precisely identify individual SNPs, CpG loci, splice junctions, or cSNPs.

To find trends across samples, markers, or different assays, the GenomeStudio framework provides a wide range of graphical plotting and display tools (Table 1). Researchers can choose to display data as line graphs, histograms, scatterplots, pie charts, dendrograms, box plots, frequency plots, or heat maps. These tools are used to easily compare samples from different experimental conditions in order to identify differential expression levels, protein expression, or methylation levels.

When trends or interesting regions are identified with graphical analysis tools, looking at individual data points becomes essential. GenomeStudio software supports this single-site level of analysis of individual SNP genotypes, splice junctions, gene or exon expression levels, CpG loci methylation status, or protein binding site occupation levels with table displays. Table views are customizable for sorting and to show or hide various data categories. Table data can also be exported in formats compatible with other downstream analysis tools.

Controls Dashboard

Illumina array-based assays, including Infinium, GoldenGate® Genotyping or Methylation Profiling, Direct Hyb, or DASL® assays contain internal sample-dependent and sample-independent controls so researchers have confidence that they are producing the highest quality data. The performance of all controls can be easily monitored with the GenomeStudio software integrated controls dashboard.

Application-Specific Analysis Modules

The modular nature of GenomeStudio software enables powerful assay-specific analysis and allows individual applications to be updated or added as necessary. GenomeStudio modules cover the spectrum of Illumina applications, including microarray-based genotyping, gene expression, methylation, and immunoprotein assay analysis, as well as DNA sequencing, chromatin immunoprecipitation sequencing (ChIP-Seq), and mRNA sequencing.

Array-Based Applications

Genotyping (GT) Module

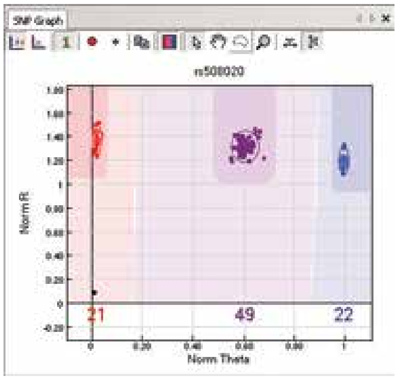
Genotyping data generated using the GoldenGate or Infinium assays on the iScan System are analyzed in the GenomeStudio Genotyping (GT) Module. This module uses algorithms to perform primary data analyses, such as raw data normalization, clustering, and genotype calling. Data quality is rapidly confirmed with internal controls and other QC functions. Individual SNPs can be viewed as GenoPlots and edited if necessary (Figure 2). Genotype summary statistics and results are automatically reported and exportable for use in third-party downstream analysis software.

Structural variation is identified using the same markers as genotyping and intensity-only probes with algorithms to calculate loss of heterozygosity (LOH) and abnormal copy numbers (CNVs). Identified structural variants can be bookmarked (with auto-bookmarking features) and viewed in the context of the entire chromosomes with the ICB or IGV. In addition, GenomeStudio software provides data plots displaying CNV values, log R ratios, B-allele frequencies, and bookmarks for one or more samples within the IGV.

Polyploid Clustering (PC) Module

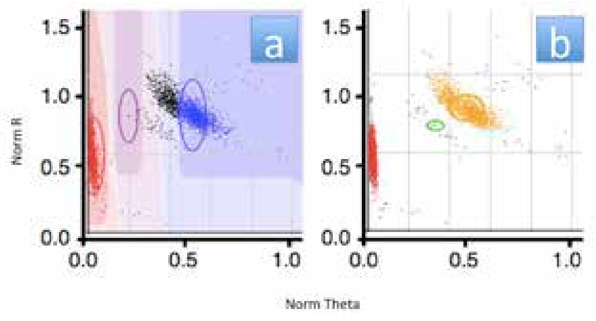
GenomeStudio now has the ability to analyze data obtained from genotyping of polyploid organisms. The Polyploid Clustering (PC) Module implements two well-known classic density-clustering algorithms, OPTICS and DBSCAN, to call as many clusters as desired.

Figure 2: Genotyping Module GenoPlot



The graphical display of results in GenomeStudio GT module is a GenoPlot with data points color coded for the call (red = AA, purple = AB, blue = BB). Genotypes are called for each sample (dots) by their signal intensity (Norm R, y-axis) and Allele Frequency (Norm Theta , x-axis) relative to canonical cluster positions (dark shading) for a given SNP marker.

Figure 3: Polyploid Clustering Module



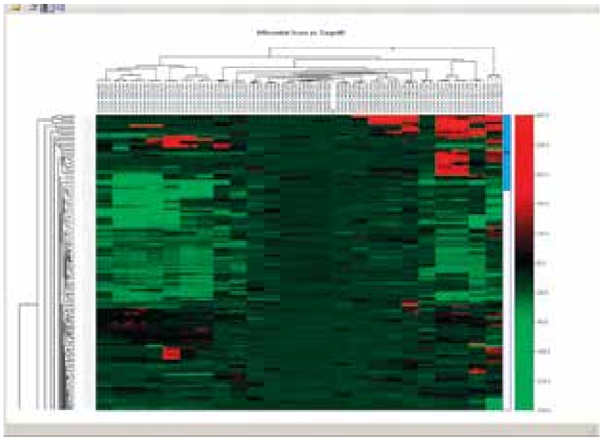
A locus analyzed within the Genotyping Module (a), which assumes diploidy, is compared to the same locus with samples assigned to cluster membership within the Polyploid Clustering Module (b). Poly call rate for this locus is higher in (b) than in (a) due to the higher number of samples with cluster assignment. The Polyploid Clustering Module does not have an *a priori* assumption of the shape of clusters, allowing for the detection of differences in allele dosing as well as hybridization efficiency. For this reason, the Polyploid Clustering Module does not call genotypes, providing researchers with the flexibility to determine genotype assignment based on the known biology of the organism.

The module (Figure 3) intentionally does not call polyploid genotypes. Instead, it allows the user to factor in experimental design and sample biology, and combine the population-level cluster analysis to intelligently call genotypes in a workflow outside of GenomeStudio. The flexibility built into the module allows for clustering of one, several, or all SNPs simultaneously. Once parameters are selected, they can be saved for automated clustering of new sample sets.

Gene Expression (GX) Module

Data from Direct Hyb, DASL, and Whole-Genome DASL gene expression profiling assays generated using the iScan System are analyzed using the Gene Expression (GX) Module. The results generated using this module provide meaningful conclusions from the continuous expression data on gene-level statistical analysis tools. Differential expression analysis can be visualized as line plots, histograms, dendrograms, box plots, heat maps, scatter plots, frequency plots, pie charts, samples tables, and gene clustering diagrams (Figure 4). Simplified data management tools include hierarchical organization of samples, groups, group sets, and all associated project analysis.

Figure 4: Gene Expression Module Heat Map



Using the heat map function in the GX Module allows easy visualization and analysis of large amounts of data. This heat map dendrogram clusters rows (Target ID) and columns (Differential Scores).

Methylation (M) Module

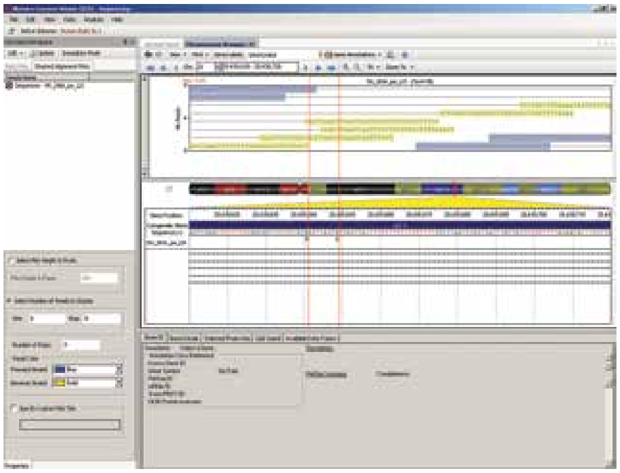
DNA methylation data from scanned microarray images collected from the iScan System are analyzed with the Methylation (M) Module. This module calculates methylation levels (beta values) and analyzes differences between experimental groups. CpG island methylation status is visualized across the genome with the IGV and ICB. Results from single-site resolution data are visualized as line plots, bar graphs, scatter plots, frequency plots, pie charts, histograms, dendrograms, box plots, or heat maps. Methylation data can also be combined with gene expression profiling experiments within the same GenomeStudio project to study any correlation between levels of methylated sites (beta values) and differential gene expression levels (p-values).

Sequencing Applications

DNA Sequencing (DS) Module

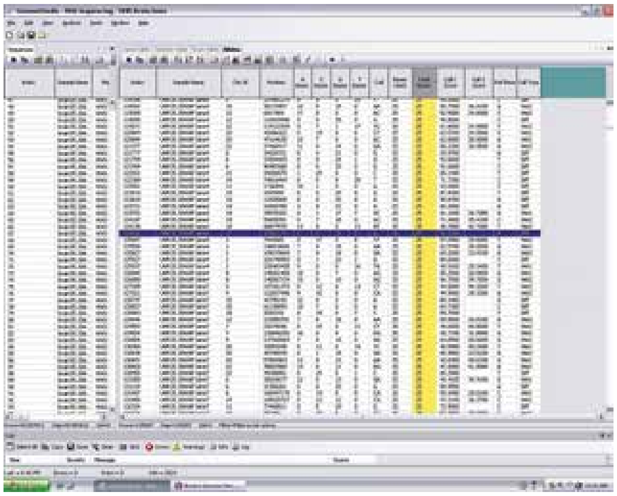
DNA sequencing data generated using the Genome Analyzer or HiSeq instruments and software tools can be analyzed to discover and confirm SNPs and chromosomal breakpoint regions in the DNA Sequencing (DS) Module. Visualization tools display consensus reads in the reassembled genome and indicate SNPs with colored letters (Figure 5). Newly discovered SNPs can be exported to use in customized iSelect® genotyping array designs.

Figure 5: SNPs Identified From Aligned Reads Displayed in DNA Sequencing Module



Aligned sequencing reads (yellow and purple blocks) are stacked on a reference genome in the ICB. SNPs are identified with red characters and in the called SNPs data track. Two SNPs are highlighted with a ruler indicating the position of the called SNPs in the aligned reads relative to the reference genome.

Figure 6: RNA Sequencing Module Sequencing and Alleles Table



Sequence information for more than 130 million reads from a flow cell is accessible and viewable in real-time in the Sequences Table of the RNA Sequencing Module (left). All called cSNPs can be individually explored in depth with the Alleles Table (right).

ChIP Sequencing (CS) Module

Data from whole-genome chromatin immunoprecipitation sequencing experiments performed using the Genome Analyzer or HiSeq instruments and software can be parsed to the GenomeStudio ChIP Sequencing (CS) Module to create global binding site maps of DNA-associated proteins. Differential binding levels between experimental groups can be identified by comparing sequences, regions, and peaks in table or chromosome views.

RNA Sequencing (RS) Module

Data generated from mRNA sequencing experiments using the Genome Analyzer or HiSeq instruments and software tools are displayed in the RNA Sequencing (RS) Module as expression levels and variants discovered. By aggregating data from the software, the RS Module is able to count the abundance of reads falling within specific exons, genes, and splice junctions. Data are then graphically displayed as tables or plots within GenomeStudio software (Figure 6). Genome views display consensus reads in the transcriptome by aligning reads to known abundant sequences and splice junctions. Coding SNPs and splice variants are identified and confirmed visually with single-base resolution in the ICB.

Integrated Analysis From Multiple Applications

The research community is taking advantage of the multiple types of assays and platforms offered by Illumina to perform a variety of genetic variation analysis studies. GenomeStudio software supports these powerful integrative studies with the ability to combine data sets from different assay types in a single project requiring minimal data handling and preparation by researchers. For example, data from methylation and gene expression assays can be analyzed together in a single GenomeStudio GX module project table where combined statistics are shown and integrated plots generated.

Sophisticated Analysis Algorithms

GenomeStudio software is part of the overall molecular biology informatics platform supporting all Illumina assays. Primary data analysis involves several algorithms that are either integrated with GenomeStudio software or in the upstream Pipeline software. Primary data analysis functions for genotype calling, CNV identification, sequence read alignment, exon and splice junction counting, and SNP calling tools are provided by algorithms such as GenCall, GenTrain, cnvPartition, ELAND, and CASAVA.

GenCall

By comparing the two-color signal intensities produced by a BeadChip marker to canonical genotype clusters, genotypes can be called. The millions of calls resulting from Infinium BeadChip assays are made quickly and reproducibly for display in GenomeStudio software with the integrated GenCall algorithm. Cluster position identification, when necessary, is performed by the GenTrain algorithm[1].

cnvPartition

cnvPartition uses Illumina BeadChip genotyping array data (signal intensities and genotype calls) to identify regions of unexpected copy number and calculate the copy numbers of those regions with confidence scores. The copy number values are then used to create CNV regions and bookmarks in GenomeStudio software for visualization of aberrant chromosomal regions across the genome.

Table 2: Minimum GenomeStudio System Recommendations

Parameter	Sequencing data analysis	Microarray data analysis	Microarray and sequencing data analysis
CPU Speed	2.0 GHz or greater	2.0 GHz or greater	2.0 GHz or greater
Processor	64-bit	64-bit	64-bit
Memory	8 GB or more	8 GB or more	8 GB or more
Hard Drive	250 GB or larger	250 GB or larger	250 GB or larger
Operating System	Windows XP, Vista, or 7	Windows Vista or 7	Windows XP, Vista, or 7

ELAND and CASAVA

Software outputs processed sequence data that GenomeStudio modules display graphically. Single or paired-end sequence read alignments to a reference sequence are performed by ELAND.

The CASAVA software package performs post-sequencing analysis (including SNP allele calls and counts of exons, genes, and splice junctions from RNA samples) of data from reads aligned to the reference genome.

Open Architecture

GenomeStudio software offers a flexible and open architecture for easy integration with third-party applications and tools. Available application programming interfaces (API) ensure that GenomeStudio software serves as a robust core of any analysis workflow. GenomeStudio software offers an API for each module that permits users to create report plug-ins (dlls) for parsing data from GenomeStudio software to downstream analysis tools. The illumina•Connect third-party partnership program encourages informatics software vendors and the open source community to leverage this open architecture. This program has led to several custom report plug-ins created and supported by various illumina•Connect partners[2].

Summary

GenomeStudio software provides a diverse and integrated platform for data analysis of Illumina assays. Researchers doing sequencing or array experiments use the same powerful software package. The graphical display of results generated from primary data analysis with assay-specific modules supports high-level and in-depth views of whole-genome variation. Integrated analysis is directly supported by combining data from different modules into a single project.

Ordering Information

Access to appropriate GenomeStudio modules is included with instrument purchase. Licenses for additional users and applications may be purchased separately. Learn more about this felxible informatics solution and third-party software tools at www.illumina.com/genomestudio.

References

1. www.illumina.com/documents/products/technotes/technote_gentrain2.pdf
2. www.illumina.com/illuminaconnect

Ordering Information

Product	Seat License	Catalog No.
GenomeStudio DNA Sequencing Module	Single Seat	SW-600-1001
	Five Seat	SW-600-5001
GenomeStudio ChIP-Seq Module	Single Seat	SW-500-1001
	Five Seat	SW-500-5001
GenomeStudio RNA Sequencing Module	Single Seat	SW-700-1001
	Five Seat	SW-700-5001
GenomeStudio Genotyping Module	Single Seat	SW-100-1001
	Five Seat	SW-100-5001
GenomeStudio Gene Expression Module	Single Seat	SW-200-1001
	Five Seat	SW-200-5001
GenomeStudio Methylation Module	Single Seat	SW-300-1001
	Five Seat	SW-300-5001
GenomeStudio Sequencing Bundle (Includes DS, RS, CS Modules)	Single Seat	SW-820-1001
	Five Seat	SW-820-5001
	Enterprise	SW-820-2001
GenomeStudio Microarray Bundle (Includes GT, GX, M Modules)	Single Seat	SW-810-1001
	Five Seat	SW-810-5001
	Enterprise	SW-810-2001
GenomeStudio Software Integrated System Bundle (All Modules)	Five Seat	SW-800-5001
	Enterprise	SW-800-2001

Illumina • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

FOR RESEARCH USE ONLY

Pub. No. 970-2008-038 Current as of 21 October 2013