Microarray Data Analysis Workflows

Optimizing analysis efficiency for low- and high-throughput environments

Introduction

Illumina's whole-genome genotyping BeadChips have dramatically grown in complexity, with the latest ones providing nearly 5M markers per sample, leading to a substantial increase in the amount of data being processed. Such large data sets can significantly increase the import and processing time required by the analysis pipeline. To help users optimize data processing efficiency, Illumina offers specific recommendations for analysis workflows based on the throughput volume of a given project.

Workflow Considerations

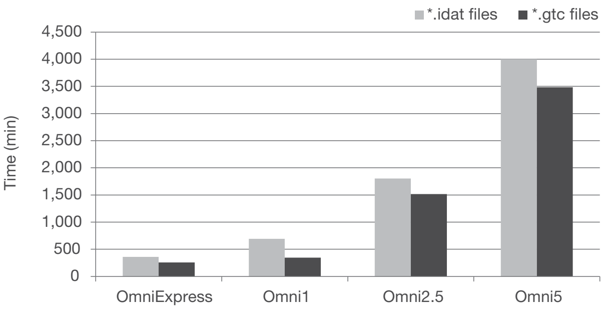
The data analysis pipeline (Figure 1) can be configured at multiple points to optimize processing time. Each of these points can be configured for a given project to provide the maximum efficiency for the pipeline. The following sections describe the key considerations for minimizing processing time.

Import File Type

After a BeadChip is scanned, the data needs to be imported into GenomeStudio® Software for analysis. Two different file types can be used for this process: Intensity Data files (*.idat) or Genotype Call files (*.gtc). GenomeStudio software processes *.gtc files roughly 20% faster (Figure 2), so it is always best to convert the *.idat files to *.gtc files before importing.

During *.gtc file generation, signal intensity data from the *idat files is combined with information about SNP/probe content on the array (*.bpm files) and the cluster reference information for each loci (*.egt files)— all of which are needed to make a genotype call. The *.gtc files should be generated using the AutoConvert function within the Instrument Control Software (iCS), or using AutoCall Software on the Illumina Laboratory Information Management (LIMS) server. Users should have ICS v3.2.45 or greater installed before proceeding.

Figure 2: Comparison of Time Requirements with Different File Types



Time required to generate sample and SNP statistics for 1,000 samples, using GenomeStudio software with two different import file types. The data were processed on a workstation with 8 GB RAM.

Data Pre-Filtering

Data set size has the greatest impact on the processing time of the analysis pipeline (Table 1). Using GenomeStudio to import and process large data sets can be slow, requiring multiple days, especially when using limited computational resources (< 16 GB RAM). To minimize the time needed for analysis, large data sets should be pre-filtered prior to import into GenomeStudio software.

Illumina's Beeline software pre-filters large data sets, reducing the overall data volume by eliminating any poorly performing loci (such as those with low signal intensity) and non-relevant variants across all samples prior to import into GenomeStudio. While Beeline and GenomeStudio software can perform similar tasks, Beeline was designed for optimal data import processing performance, while

Figure 1: Data Analysis Workflow Overview

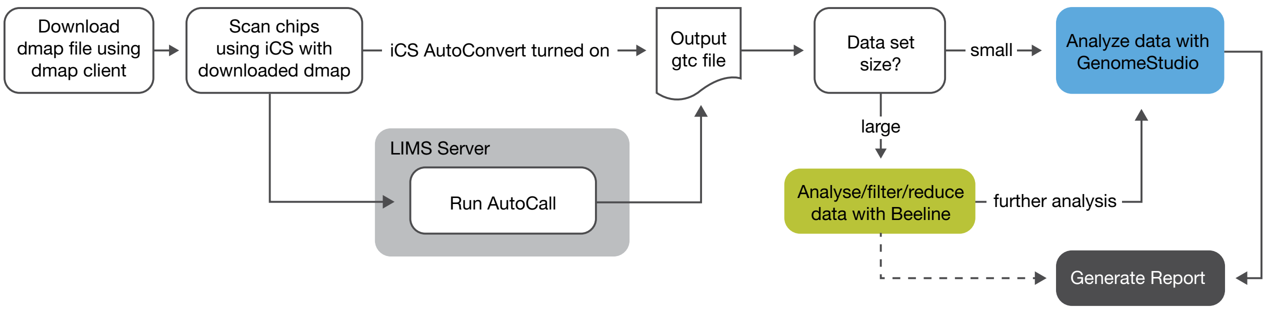

Download dmap file using dmap client
Scan chips using iCS with downloaded dmap
iCS AutoConvert turned on
Output gtc file
LIMS Server
Run AutoCall
Data set size?
small
large
Analyze data with GenomeStudio
Analyse/filter/reduce data with Beeline
further analysis
Generate Report

Table 1: Processing Time for Progressively Larger Data Volumes

OmniExpress	Omni1	Omni2.5	Omni5
257 min (4.28 hr)	346 min (5.77 hr)	1,517 min (25.28 hr)	3,484 min (58.07 hr)

Time to create a 1,000 sample project from *.gtc file, using GenomeStudio v2011.1 on a computer with 8GB of RAM. In this context, a large data set is one that takes more than 24 hours (1440 min) to create a GenomeStudio project from *.gtc files.

GenomeStudio was designed for optimal data analysis and reporting (Table 2). Using both programs in the analysis workflow greatly improves the overall time required to create a GenomeStudio project (Figure 3).

Hardware Considerations: System Memory

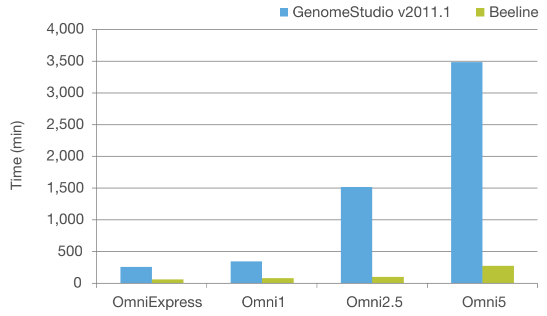
The amount of RAM available on a system will impact the data processing time, with more RAM resulting in shorter times for project creation (Figure 4). Internal tests of GenomeStudio v2011.1 on machines with 32 GB of RAM have shown that no more than 13 GB of RAM was consumed at maximum performance. While GenomeStudio software can be run with 16 GB systems, researchers should include sufficient additional RAM to effectively run all other programs.

Table 2: Comparison of Data Analysis Features for Beeline and GenomeStudio

	Beeline	GenomeStudio
Data import speed	✓✓	✓
Diverse output file formats	✓	✓✓
Can create GenomeStudio project	✓	✓
Thresholding and filtering	✓✓	✓
Clustering	✗	✓
GenTrain cluster generation	✗	✓
Controls dashboard	✗	✓
Reporting capabilities	✓	✓✓

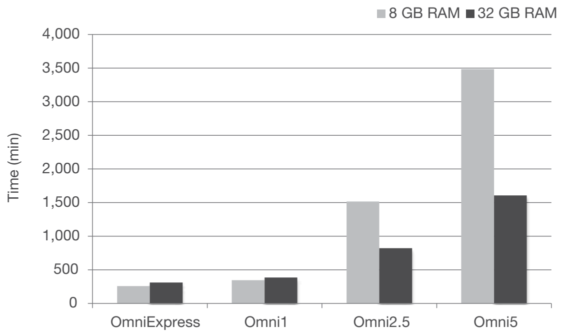
✓ = software includes feature
✓✓ = software is optimized for the feature
✗ = software does not include the feature

Figure 3: Beeline Offers Faster Performance Times than GenomeStudio



Comparison of time required to generate sample & SNP statistics for 1,000 samples using Beeline or GenomeStudio software. The data were processed on a workstation with 8 GB RAM.

Figure 4: Impact of a Greater Amount of RAM on Processing Time



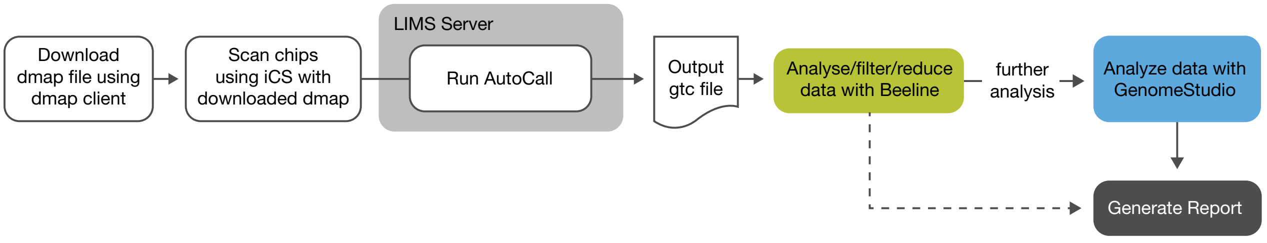
Comparison of time required to generate sample and SNP statistics for 1,000 samples using GenomeStudio on workstations with 8 GB and 32 GB RAM.

High-Throughput Recommendations

Labs processing data sets that require greater than 24 hours are high-throughput environments (Table 1). In this setting, instruments are often running at very high capacity (24 hours/day, 7 days/week). To minimize instrument downtime (i.e., the time required to convert *.idat to *.gtc files), high-throughput users should turn off AutoConvert on the instrument (Figure 5). The *.idat data should be directed to an offline server with AutoCall (LIMS users) or AutoConvert (non-LIMS users) for the *.gtc file conversion.

Import the *.gtc files from AutoCall/AutoConvert directly into Beeline prior to analysis in GenomeStudio software. Beeline filters the loci across all samples to include only the useful data, reducing the final number of loci to be analyzed and resulting in better overall performance time. Users also have the option use the speed of Beeline to rapidly import the data and run basic analysis and reporting without using GenomeStudio software.

Figure 5: High-Throughput Workflow


Download dmap file using dmap client
Scan chips using iCS with downloaded dmap
LIMS Server
Run AutoCall
Output gtc file
Analyse/filter/reduce data with Beeline
further analysis
Analyze data with GenomeStudio
Generate Report

Hardware Requirements

For high-throughput environments, workstations should contain at least 16 GB of RAM (Table 3). To achieve optimal performance, systems should be equipped with 32 GB of RAM. This will provide GenomeStudio and Beeline software sufficient memory to load and process data.

Table 3: High-Throughput System Requirements

Minimum Specifications	Recommended Specifications
2.0 GHz or greater	2.2 GHz or greater
2 or more cores	2 or more cores
16 GB RAM	32 GB RAM

Software Requirements

Due to the lengthy time requirement for GenomeStudio software to import and process a high volume of data, Beeline software should be used in conjunction with GenomeStudio software to pre-filter and reduce the data set (Table 4).

Table 4: High-Throughput Software

	Beeline	GenomeStudio
Data Filtering	required	required
No Data Filtering (not recommended)	not required	required

Low-Throughput Recommendations

For low-throughput environments, where instruments are not running at high capacity, the AutoConvert functionality within iCS should be used to convert *.idat files to *.gtc files (Figure 6). Running this procedure on the instrument reduces the expense of an offline AutoCall server.

With a smaller data volume, low-throughput users may not need to use Beeline to pre-filter the loci before analysis in GenomeStudio. The *.gtc files from AutoConvert can be imported directly into GenomeStudio software from iCS.

Hardware Requirements

For low-throughput environments, workstations must contain a minimum of 8 GB RAM (Table 5). While GenomeStudio software can be run using the minimum specifications, Illumina recommends using at least 16 GB RAM to optimize processing time.

Table 5: Low-Throughput System Requirements

Minimum Specifications	Recommended Specifications
2.0 GHz or greater	2.2 GHz or greater
2 or more cores	2 or more cores
8 GB RAM	16 GB RAM

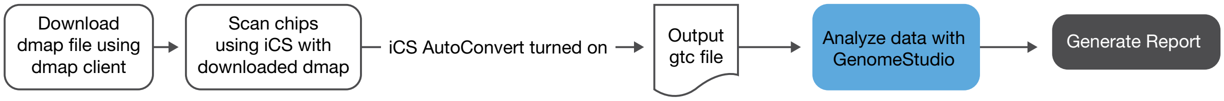
Software Requirements

Because data from iCS can be directly imported into GenomeStudio software, users working in a low-throughput environment might not need to pre-filter data with Beeline software (Table 6).

Table 6: Low-Throughput Software

	Beeline	GenomeStudio
Data Filtering	required	required
No Data Filtering	not required	required

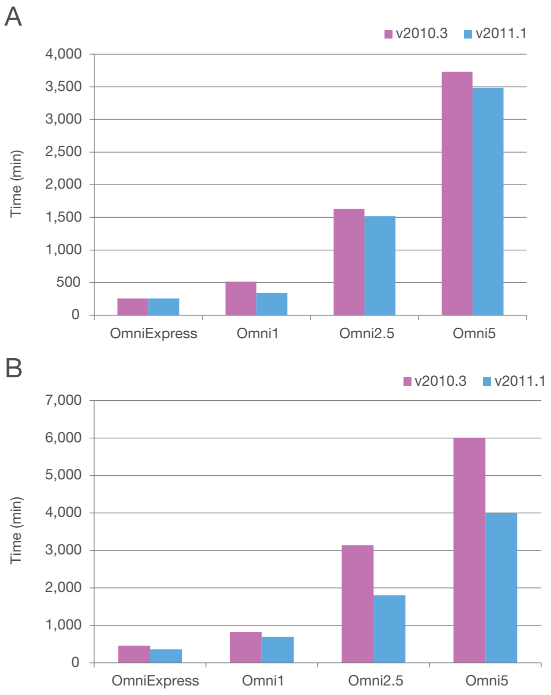
Figure 6: Low-Throughput Workflow


Download dmap file using dmap client
Scan chips using iCS with downloaded dmap
iCS AutoConvert turned on
Output gtc file
Analyze data with GenomeStudio
Generate Report

Improved Performance Using the Latest GenomeStudio Software Release

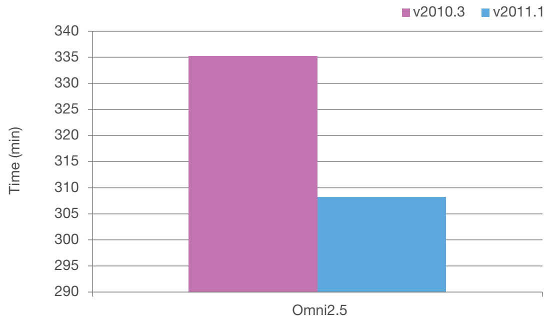
GenomeStudio software is continually optimized to deliver improved performance for the latest whole-genome arrays. Users in both high- and low-throughput labs should consider upgrading to the most recent GenomeStudio version for optimal performance. To demonstrate the improved performance time using the most recent version of the software, 1,000 samples were imported and processed using two versions of GenomeStudio software. As shown, the more recent version of GenomeStudio software (v2011.1) offers improved project creation times (Figure 7) and improved sample re-clustering times (Figure 8) over the preceding version.

Figure 7: Improved Project Creation Times with GenomeStudio v2011.1



Time required to create a project file from 1,000 samples scanned on four whole-genome arrays, processed using A) *.gtc files and B) *.idat files. The project creation time involves calculating sample and SNPs statistics. The data were processed using an 8 GB workstation with Windows 7 OS. For both file types, GenomeStudio v2011.1 delivered improved performance times across all arrays.

Figure 8: Improved Recluster Time with GenomeStudio v2011.1



Time to recluster 1000 samples on the Omni2.5 array using two recent versions of GenomeStudio. The data were process using 32 GB workstation with Windows 7 OS. GenomeStudio v2011.1 delivered significantly better performance time over the previous version of the software.

Summary

The microarray analysis pipeline can be configured at multiple points to optimize efficiency based on the volume of data being processed. By considering multiple factors such as the data file conversion, loci pre-filtering, and system hardware and software requirements, researchers can minimize the sample processing time for high- and low-throughput environments.

FOR RESEARCH USE ONLY

Pub. No. 970-2011-006 Current as of 07 July 2011