illumina®

Microarray Data Analysis Workflows

Optimizing analysis efficiency for low- and high-throughput workflows.

Introduction

The content and capacity of Infinium® whole-genome genotyping BeadChips continue to grow dramatically, leading to substantial increases in the amount of data being processed. Such large data sets can increase the import and processing time required by the analysis pipeline significantly. To help users optimize data processing efficiency, this technical note describes specific recommendations for analysis workflows based on the throughput volume of a given project.

Import File Type

Any array-based genotyping assay begins with data acquisition. After sample DNA is hybridized to a BeadChip, it is loaded onto an iScan® System and scanned. DMAP files, which can be downloaded from the Illumina support site, enable identification of bead locations on the BeadChip and quanitification of the signal associated with each bead. The resulting raw Intensity Data files (*.idat) can be converted to Genotype Call files (*.gtc). Both file types are compatible with GenomeStudio® Software. However, given that *.gtc files can be processed ~ 20% faster than *.idat files (Figure 2), it is best practice to convert *.idat files to *.gtc files before importing into GenomeStudio Software for analysis.

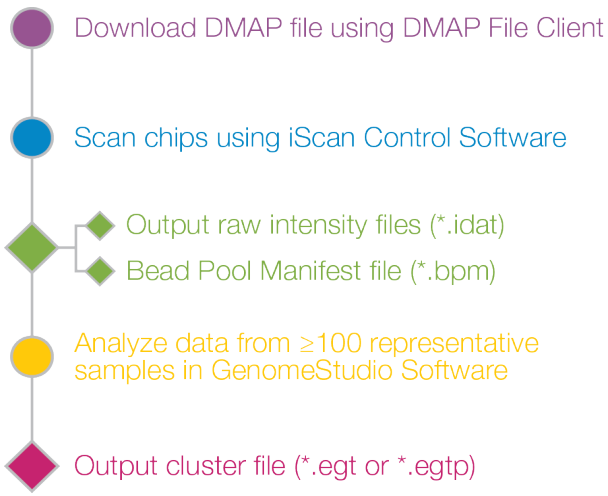

Figure 1: Cluster File Generation—Custom iSelect Arrays require generation of clusters files for each project. After a BeadChip is scanned, the resulting data files (*.idat) and manifest files (*.bpm) are used to analyze at least 100 representative samples in GenomeStudio Software to create cluster files.

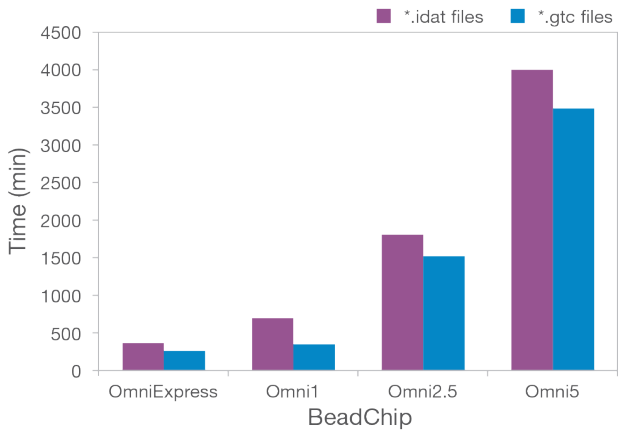

Figure 2: Comparison of Analysis Times with Different File Types—Time required to generate sample and SNP statistics for 1000 samples using GenomeStudio Software with 2 different import file types. Data were processed on a computer with 8 GB RAM.

During *.gtc file generation, raw signal intensity data from the *.idat files is combined with Bead Pool Manifest files (*.bpm) that contain information about single-nucleotide polymorphism (SNP)/probe content on the BeadChip and cluster files (*.egt or *.egtp), which have reference information for each interrogated locus. All these files are needed for Illumina software to call genotypes successfully.

iScan Control Software includes an AutoConvert feature, which automatically converts *.idat files to *.gtc files during scanning on a per chip basis. Alternatively, the Illumina Laboratory Information Management System (LIMS) has its own built-in automated *.idat file conversion feature, called AutoCall. As discussed in subsequent sections, the volume of a given project will determine which program to use for this file conversion.

Cluster File Generation

While pregenerated cluster files (*.egt for diploid organisms or *.egtp for polyploid organisms) are provided for Illumina commercial genotyping array products, cluster files are not provided for custom iSelect® Arrays. To generate cluster files for custom arrays, GenomeStudio Software uses a training data set of at least 100 representative samples (Figure 1). After a cluster file is created for a given experiment, the same cluster file can typically be reused throughout the study. If experimental conditions change enough to cause the quality of genotype calling to drift from the initial results, a new cluster file should be created using a new training data set.

Low-Throughput Workflow

For low-throughput environments where instruments are not running at high capacity, use the AutoConvert functionality within iScan Control Software to convert *.idat files to *.gtc files (Figure 3). Running this procedure on the instrument reduces the expense of an offline LIMS AutoCall server. However, the iScan System cannot be used to scan another BeadChip until after the AutoConvert process is complete.

With a smaller data volume, low-throughput users are unlikely to need to use Beeline™ Software to prefilter the loci before performing further analysis. The *.gtc files from AutoConvert can be imported directly into GenomeStudio Software from iScan Control Software for analysis. Report files (*.txt) generated by GenomeStudio Software can be analyzed further downstream using third-party software (Figure 3).

Hardware Requirements

For low-throughput environments, computer workstations must contain a minimum of 8 GB RAM (Table 1). While GenomeStudio Software can be run using the minimum specifications, at least 16 GB RAM is recommended to optimize processing time.

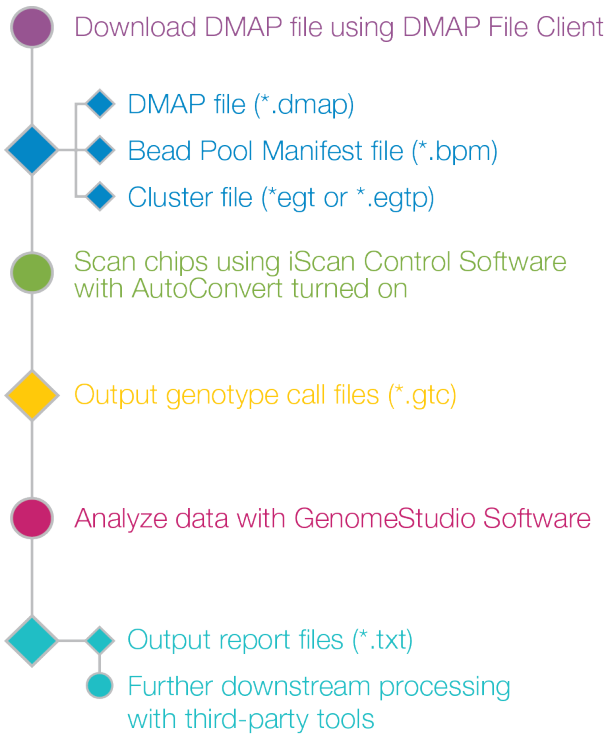

Figure 3: Low-Throughput Workflow—For low-throughput labs, it is recommended that users convert *.idat files into *.gtc files with the iScan Control Software AutoConvert feature before importing into GenomeStudio Software for analysis.

Table 1: Low-Throughput System Requirements

Parameter	Minimum Specifications	Recommended Specifications
CPU Speed	2.0 GHz or greater	2.2 GHz or greater
Number of CPU Cores	2 or more cores	2 or more cores
Memory	8 GB RAM	16 GB RAM

Software Requirements

Because data from iScan Control Software can be directly imported into GenomeStudio Software, users working in a low-throughput environment might not need to prefilter data with Beeline Software (Table 2).

Table 2: Low-Throughput Software Usage

Feature	Beeline Software	GenomeStudio Software
Data Filtering	Required	Required
No Data Filtering	Not Required	Required

High-Throughput Workflow

High-throughput labs often process data sets at high capacities (eg, 24 hours/day, 7 days/week). Because the iScan System must be offline to perform file conversion, turn off the AutoConvert feature during BeadChip scanning. Direct the *.idat files to an offline server with AutoCall (LIMS users) or AutoConvert (non-LIMS users) for the *.gtc file conversion (Figure 4).

The *.gtc files generated by AutoCall/AutoConvert are imported directly into Beeline Software, where data is filtered to exclude poorly performing loci across all samples (Figure 4). Including only useful data reduces the final number of loci to be analyzed, resulting in better overall performance time. Users can import data rapidly into Beeline Software to run basic analysis and reporting without using GenomeStudio Software, other than to generate cluster files. Users can create GenomeStudio projects from Beeline Software to investigate their data more deeply using the robust tools found in GenomeStudio Software. Alternatively, users can employ third-party software to perform further analysis (Figure 4).

Hardware Requirements

For high-throughput environments, computer 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 Beeline and GenomeStudio Software with sufficient memory to load and process data.

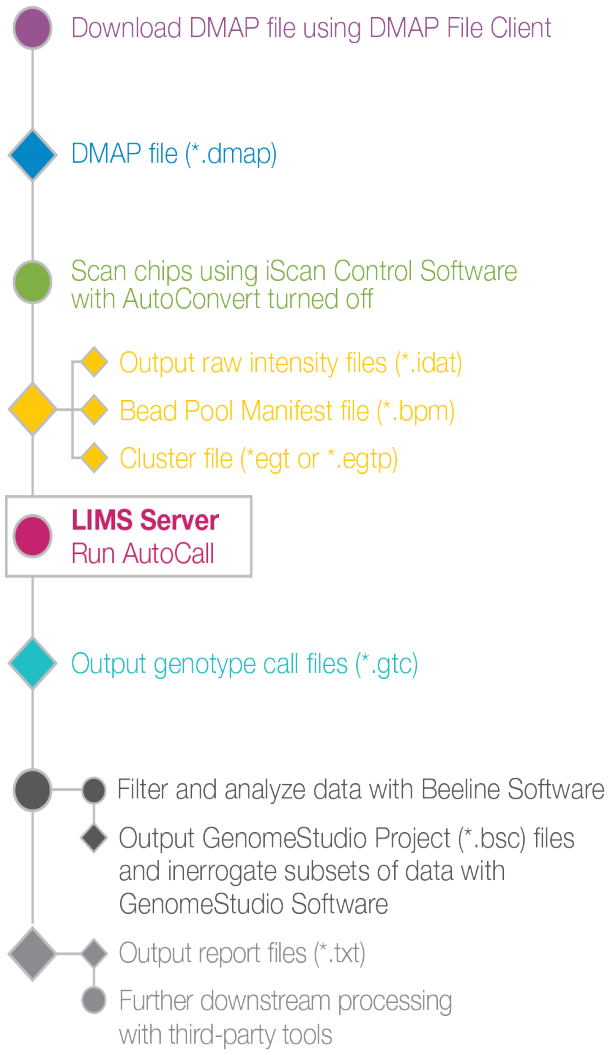

Figure 4: High-Throughput Workflow—For high-throughput labs, convert *.idat files to *.gtc files on an offline server using LIMS AutoCall to minimize instrument downtime. After conversion, Beeline Software filters and analyzes *.gtc files. Use GenomeStudio Software or third-party software for further analysis.

Table 3: High-Throughput System Requirements

Parameter	Minimum Specifications	Recommended Specifications
CPU Speed	2.0 GHz or greater	2.2 GHz or greater
Number of CPU Cores	2 or more cores	2 or more cores
Memory	16 GB RAM	32 GB RAM

Software Requirements

Because prefiltering data is important in high-throughput environments to improve the performance time of analysis, use of Beeline Software is recommended. GenomeStudio Software, while not strictly required for most analysis and reporting, is still needed to generate genotyping cluster files (Table 4).

Table 4: High-Throughput Software Usage

Feature	Beeline Software	GenomeStudio Software
Data Filtering	Required	Required (for cluster file generation)
No Data Filtering (not recommended)	Not Required	Required

Open-Source Tooling

While Beeline and GenomeStudio Softwares offer robust and flexible solutions for filtering, analyzing, and reporting genotyping data, some users prefer a fully automated approach that does not require interacting with any graphical user interfaces. To this end, an open-source library for parsing genotype call data encoded in *.gtc files generated by either AutoCall or AutoConvert is available. Experienced users can access this library to develop their own methods for filtering, analyzing, and reporting genotyping data that can be executed in a fully automated pipeline (Figure 5). GenomeStudio Software is still required to generate new genotyping cluster files (Table 5). For more information and to access the open-source library for parsing *.gtc files, visit the GitHub repository: github.com/Illumina/BeadArrayFiles.

Table 5: Open-Source Workflow Software Usage

Feature	Beeline Software	GenomeStudio Software
Data Filtering	Not Required	Required (for cluster file generation)
No Data Filtering	Not Required	Required (for cluster file generation)

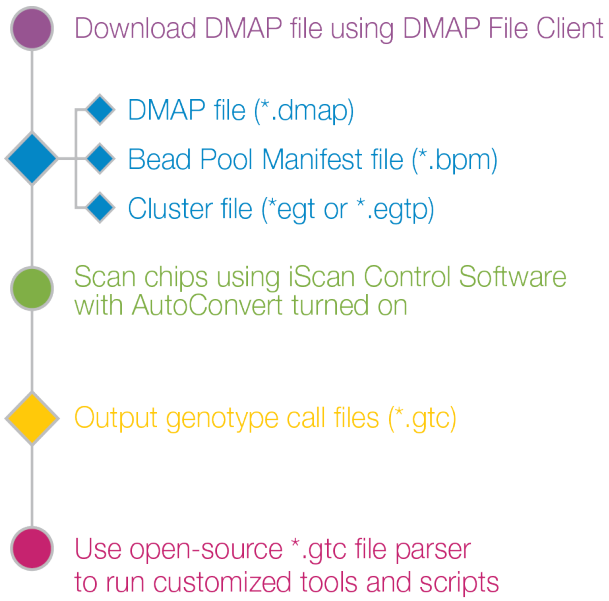

Figure 5: Open-Source Tooling Workflow—Experienced users can automate genotyping analysis and reporting from *.gtc files by writing customized scripts that use the Illumina open-source library for parsing *.gtc files.

Summary

The genotyping 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 loci prefiltering and system hardware and software requirements, researchers can minimize the sample processing time for low- and high-throughput environments.

Learn More

To learn more about microarray data analysis, visit www.illumina.com/techniques/microarrays/array-data-analysis-experimental-design.html

Illumina, Inc. • 1.800.809.4566 toll-free (US) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

For Research Use Only. Not for use in diagnostic procedures.

 Pub. No. 970-2011-006. Current as of 18 October 2016

illumina®