QTL Analysis Software Tools for Illumina Data

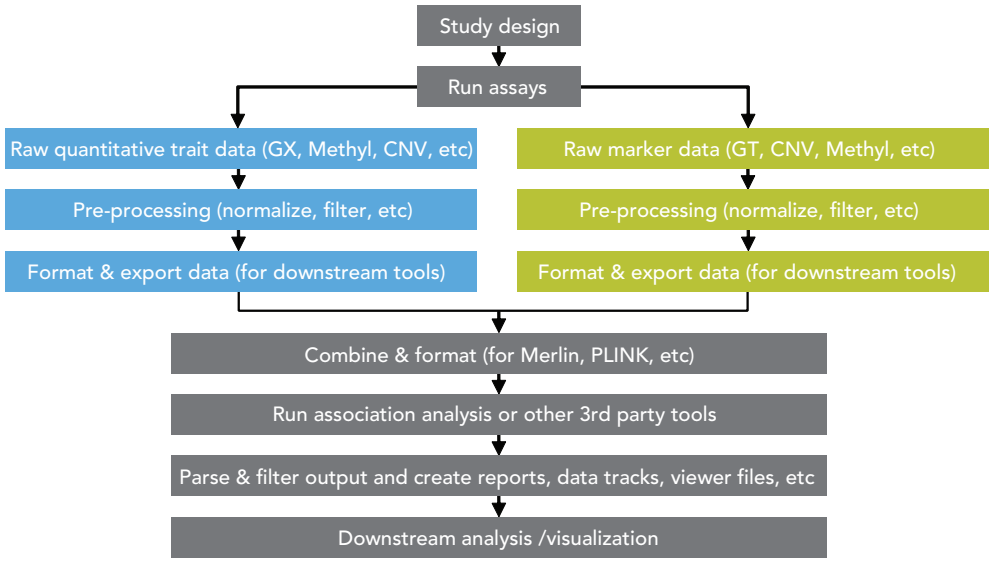
Combining the results from the wide range of Illumina genetic analysis assays is a powerful integrated approach for complex questions. Illumina provides data analysis software that supports data integration, such as genotyping with gene expression for eQTL analysis.

INTRODUCTION

In the past few years, researchers have used Illumina DNA Analysis BeadChips in genome-wide association studies (GWAS) to make numerous discoveries uncovering the genetic basis of common diseases. As geneticists turn toward elucidating the causes of more complex diseases, it is expected that associated variants will be rarer and pleiotropic phenotypes will be more common. Given these complicating factors, a typical GWAS experimental design may not be ideal. Several recent studies have suggested that an approach integrating expression profiling with genotyping—referred to as eQTL (expression quantitative trait locus) analysis—may be more powerful. eQTL study feasibility has improved recently, due to decreasing cost and enhanced analysis methods and tools. Illumina assay technologies have been used in several such studies[1–3], demonstrating their utility along with available software packages for gathering, formatting, processing, and analyzing eQTL data.

Not surprisingly, the multi-factorial nature of eQTL studies requires relatively complicated analyses to uncover significant associations. Several methods and software tools are shown to be successful in the analyses published to date. Although there is no single best method that works for every study, Illumina has continued to create flexible data analysis software tools to

FIGURE 1: GENERAL EQTL DATA WORKFLOW



This flowchart shows a general data workflow used for integrated analysis of gene expression data as quantitative trait data in an eQTL study.

ensure that customers have access to a range of available tools supporting successful start-to-finish eQTL studies using Illumina genotyping and gene expression arrays. While this document focuses on eQTL, the same principles apply to similar study designs, such as mQTL (methylation quantitative trait loci).

EQTL DATA WORKFLOW

The first step in performing an eQTL study is to run all samples and collect genotyping and gene expression data *(Figure 1)*. For each subject, there will be a pair of primary data: quantitative trait data in the form of a gene expression profile, and marker typing data in the form of genotypes or structural variants. Each of these streams of primary data is processed separately using standard tools to generate data sets ready for downstream analysis.

At this point, the data analysis workflow for an integrated study (e.g., eQTL) diverges from that of a single modality (e.g., traditional GWAS). Each data set must be formatted, exported, and combined into the correct file type(s) specific to the downstream software that will be used for the association analysis. Interpretation and visualization of results are also more complicated in such an integrated study, and must be handled after association analysis. There are a variety of options for parsing eQTL association analysis to generate reports and graphics, some of which are described below.

EQTL SOFTWARE TOOLS

Illumina GenomeStudio™ data analysis software is useful for performing the data collection and pre-processing in eQTL studies in the same manner as in non-integrated (single modality) studies. A pair of GenomeStudio plug-ins facilitate the downstream use of PLINK, a common and readily available software tool, for the association analysis steps.

Obtaining and Installing Plug-ins for eQTL Analysis

The Illumina-supplied plug-ins for eQTL and similar types of analysis using PLINK or Merlin are:

- Illumina GenomeStudio GX Custom Output Report
- Illumina GenomeStudio PLINK Input Report
- Illumina GenomeStudio Merlin Input Report
- Illumina QTL Viewer Demo

Two files are associated with each plug-in: a setup program and a user guide. The plug-ins and viewer for eQTL analysis are available for download from three locations:

- iConnect website (http://www.illumina.com/pagesnrn.ilmn?ID=229)
- GenomeStudio Portal (within GenomeStudio application)
- iCom (http://icom.illumina.com)

The Illumina GenomeStudio GX Custom Output Report can also be customized to generate input files for downstream analysis tools other than PLINK. As the field advances and generates new tools and methods for integrated genetic analyses, Illumina will continue to release other eQTL and data integration tools to support leading-edge research with a wide variety of downstream applications.

Extracting Quantitative Trait Data from the Gene Expression Module

Gene expression data from Illumina arrays can be used as quantitative trait data in an association study. The particular format required for quantitative trait data varies depending on the software tool used for the association analysis. By default, the Illumina GX Custom Output Report extracts average intensity data from the GenomeStudio Gene Expression Module in tab-delimited format, with sample IDs appearing in rows and probe IDs in columns. This export file can be used directly with the PLINK Input Report plug-in.

Extracting Quantitative Trait Data from Other Sources

Gene expression, methylation, or other text-based quantitative trait data from sources other than GenomeStudio software can also be used as input to the PLINK Input Report. The data input file should meet the following specifications:

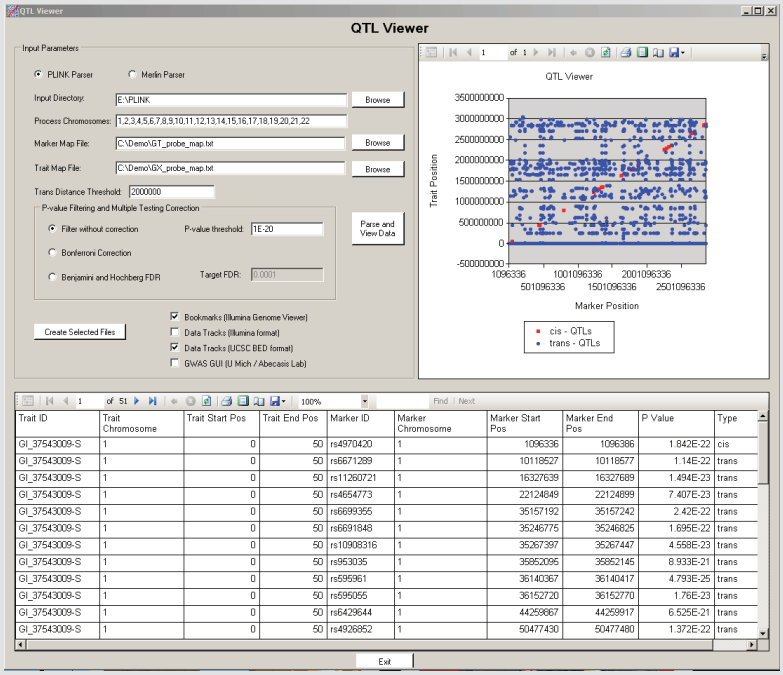
- File format is tab-delimited
- Trait sample IDs match marker sample IDs
- Sample IDs are listed in rows
- Probe IDs are listed in columns
- Quantitative traits are represented with floating point values (NaN is used to represent missing data)

ASSOCIATION ANALYSIS APPLICATIONS

Creating Formatted Input Data for Association Analysis Applications

Many different software applications can be used for eQTL analysis. PLINK is a freely available software package designed for QTL analysis that has been applied in several eQTL studies[4]. The PLINK Input Report plug-in can be used to create input data files for the association analysis feature of PLINK. This plug-in extracts the required genotyping data from an input file and offers the

FIGURE 2: ILLUMINA QTL VIEWER



The QTL Viewer demo application provides visualization tools for integrated analysis, including table and scatterplot displays.

option of combining it with a quantitative trait file. The output of the PLINK Input Report plug-in is five data files to be used when running PLINK.

- *.ped file: lists family ID, sample ID, father ID, mother ID, gender, affected status, and genotypes
- *.map file: lists chromosome, identifier, genetic position, and chromosomal position for each marker
- *.phenotype file: lists family ID, sample ID, and quantitative trait values
- *.script file: lists configuration options to be used in running PLINK
- *.bat file: provides batch file example to assist with running PLINK in Windows

Running the eQTL Association Analysis Application

After the PLINK input files have been created, they can be copied to a PLINK run folder. The location of the folder is chosen by the customer. PLINK uses these input files for the association analysis. PLINK output should be sent to a drive with several gigabytes of free disk space due to the potentially large output file sizes. PLINK package files and associated documentation are available for download at http://pngu.mgh.harvard.edu/~purcell/plink/.

Interpreting and Visualizing Results from Association Analysis Applications

Illumina provides a QTL Viewer demo application to support the interpretation of association analysis application output *(Figure 2)*. This standalone Windows application can read data from either PLINK or Merlin. The program shows a table listing all the QTL hits and a corresponding scatter plot showing the cis band in red. Users can filter data based on p-value and perform Bonferroni or FDR (false discovery rate) p-value adjustments. The QTL hit table can be exported as a *.xls file for use in Microsoft Excel.

Using Third-Party Software for eQTL Analysis

Given the study diversity and lack of clear standards in this evolving field, the workflow using GenomeStudio software and PLINK recommended above might not be optimal for every researcher. Another option for eQTL analyses is to use third-party software for more of the file format manipulation, data pre-processing, output file parsing, and visualization steps. These options include at least the following:

- Rosetta Resolver and/or Syllego
  - Resolver is useful for the filtering and QA of gene expression data
  - Syllego has many useful features, including data management, workflow management, PLINK integration, and an eQTL viewer
- Statistical software environment (e.g., R, MatLab, or SAS/JMP)

Performing an mQTL Analysis

mQTL (methylation quantitative trait loci) analysis can be performed using most of the same tools provided by Illumina with a strategy similar to an eQTL analysis. All steps are the same, except that methylation data are used for primary input instead of gene expression data. Customers using the GenomeStudio Methylation Module can use the Methylation Custom Output Report plug-in, which produces a data file in the same format as that produced by the GX Custom Output Report plug-in. This file can then be used with either the PLINK Input Report or Merlin Input Report as the quantitative trait file, as described above.

SUMMARY

By combining various Illumina assays, researchers are now able to perform powerful integrated analyses that have the promise of further advancing the study of complex diseases. For example, eQTL studies take advantage of the wealth of quantitative trait data generated in gene expression profiling assays that can be associated with genetic variants uncovered using whole-genome genotyping arrays. Illumina also provides a set of software tools to support this type of integrated analysis, and will continue to update this portfolio of tools as the field evolves. Powerful assay products combined with flexible analysis software options provide the fastest path to discovery and publication.

REFERENCES

(1) Goring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, et al. (2007) Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nat Genet 39: 1208–1216.

(2) Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, et al. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315: 848–853.

(3) Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, et al. (2007) Population genomics of human gene expression. Nat Genet 39: 1217–1224.

(4) Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: a tool-set for whole-genome association and population-based linkage analysis. AJHG, 81.

ADDITIONAL INFORMATION

Visit our website or contact us at the address below to learn more about Illumina genetic analysis products 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

FOR RESEARCH USE ONLY

Pub. No. 970-2008-022 Current as of 15 December 2008

illumina®