Gene Survey: FAQ
Tod Casasent
2016-02-22-1245
DRAFT

1 What is this document?

This document is intended for use by internal and external users of the Gene Survey package, results, and output. This document provides guidance for when to use Gene Survey, the types of information available from the Gene Survey system, and comparisons with other publicly available data repositories.

2 Questions about Gene Survey

2.1 What is Gene Survey?

Gene Survey, as a system, is intended to make it easier for researchers interested in TCGA data to compare one or more genes accessed by gene symbols or gene equivalents (such as probes) across disease types. For example, if you wish to examine copy number estimates for TP53 across all TCGA tissues, this is easy to do.

Gene Survey consists of three components:

- GeneReports R package The R package is available as a tar.gz package file or as code via GutHub.
- Gene Survey Data This is the TCGA data used by Gene Survey. Gene Survey Data is TCGA data that has been "standardized" (as described below) and then converted to a format that enables more efficient gene-level access. This data is versioned and available for download.
- Gene Survey Results This is a selection of plots and diagrams for genes across disease types based on TCGA data. The selection of genes is based on either "interesting" genes or genes which are available across all platforms and disease types.

2.2 Where can I get Gene Survey Components?

Gene Survey Results are available, versioned, with the R package and Data used to generate the results internally at http://mdadqscfs01.mdanderson.edu/GSRESULTS/ and externally at a site to be announced.

The Gene Reports R package is available as a versioned tar.gz package with the corresponding Gene Survey Results or from GitHub http://mdadqscfs01.mdanderson.edu/GSRESULTS/ for the R package and https://github.com/GeneSurvey/TcgaGSData for the Java library.

The Gene Survey Data is available as a versioned ZIP archive with the corresponding Gene Survey Results.

2.3 What Data does Gene Survey Provide?

Gene Survey provides versioned access to TCGA level 3 data across all available disease types and to selected public clinical data. Currently, public clinical data includes days_to_last_followup, days_to_death, pathologic_grade, pathologic_stage, sex, age_at_diagnosis, and vital_status.

2.4 What is a version?

TCGA data is internally versioned by the institutions submitting the data. (Each submitted archive has its own version number.) No TCGA overall version number is available. Like other institutions using the data, MD Anderson uses a timestamp based on the download of the original data for use in creating Standardized Data.

Gene Survey has a version timestamp for when Standardized Data was prepared for Gene Survey package use.

Versioning allows other researchers to reproduce your results by downloading and processing the same data set you used.

2.5 What platforms are provided by Gene Survey?

Gene Survey provides the following platforms:

- humanmethylation27_hg19Wxy Human Methylation 27 for HG19 with sex chromosomes
- humanmethylation450_level3 Human Methylation 450 for HG19 with sex chromosomes (level 3 data calculated from level 2, to include all probes)
- illuminahiseq_rnaseqv2_gene IlluminaHiSeq RNASeq V2 for HG19 (gene)
- genome_wide_snp_6_hg19nocnvWxy Genome Wide SNP6 for HG19 without CNV (Copy Number Variant) and with sex chromosomes
- illuminahiseq_mirnaseq_isoform IlluminaHiSeq miRNASeq for HG 19 (isoform)
- mutations Mutation Data
- Selected Public Clinical Data entries. Currently:
    - days_to_last_followup
    - days_to_death
    - pathologic_grade
    - pathologic_stage
    - sex
    - age_at_diagnosis
    - vital_status

2.6 How are provenances indicated?

All data versions for Standardized Data give a date for when they were downloaded from the TCGA website. Generation of data for Gene Survey and Gene Survey results are timestamped when they are performed. Results timestamps are displayed in the website. Gene Survey data timestamps can be retrieved via a package function or from looking at the time.txt file in the combined directory within the ZIP archive.

2.7 How is data accessed?

Data is stored in tab-delimited files within a ZIP archive. The single ZIP archive provides a simple method to store and utilized a single version of Gene Survey data.

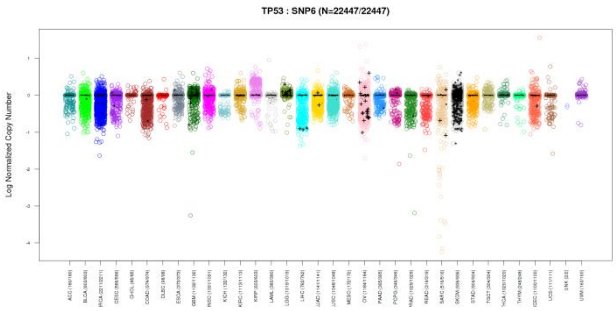
R users can access the data as a matrix by using the Gene Reports R package and specifying the gene equivalent(s) and platform in which they are interested. Java users can similarly use the Java classes provided as part of the R package to retrieve a matrix by specifying the gene equivalent(s) and platform in which they are interested. In either case, when appropriate, a gene symbol can be automatically expanded into probes and all data for all probes for a gene retrieved. Functions also exist for retrieving public clinical, location, and other meta-data for given gene equivalents and probes.

2.8 What format is the data in?

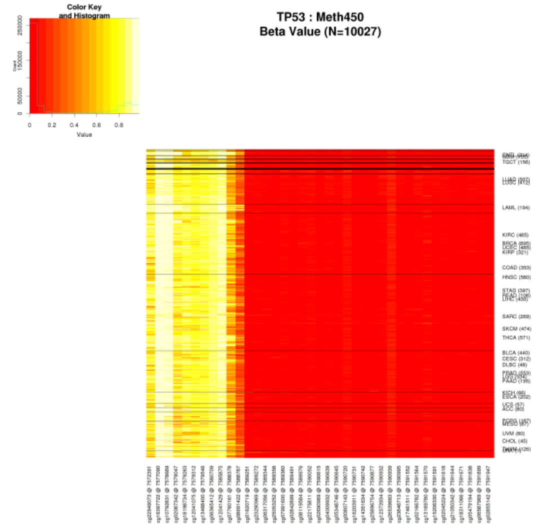
Direct file access is also available, with descriptions and formats available in other documentation. In general, data is stored in tab-delimited files within the ZIP archive.

2.9 What analyses are provided?

The Gene Survey Results website provides analysis for ~11,000 gene symbols. This includes stripcharts for gene symbols and probes, such as the one shown to the right.



The Gene Survey Results website also provides heatmaps for probes, such as the one below.



The R package creates these stripcharts and heatmaps, and has options for replicate diagrams. The R package also provides a flag for creating plots of the differences between tumor and normal samples.

3 Questions about Gene Survey and other public Data and Analysis Repositories

In this section, questions about data or analyses from Gene Survey (internally at http://mdadqscfs01.mdanderson.edu/GSRESULTS/ and externally at a site to be announced), cBio (http://www.cbioportal.org/), and FireBrowse (http://firebrowse.org/) will be addressed. These questions are in no particular order at present.

3.1 I am interested in versioned TCGA data, which system should I look at?

Gene Survey and FireBrowse both use timestamps for when data was downloaded from the TCGA website. Both provide access to older versions and are updated on a semi-regular basis (every 1-3 months). cBio does not provide versioning.

3.2 I am interested in using the original TCGA data, which system should I look at?

Gene Survey provides gene equivalent data across disease types on a gene equivalent or probe level. Also, MD Anderson TCGA Standardized Data (http://bioinformatics.mdanderson.org/TCGA/databrowser) provides access to level 3 TCGA data in a matrix format--this data format is simple to work with but is a transformation of the original data as described at that website.

FireBrowse provides tab delimited files, where multiple data files from the as-submitted files (often without barcodes in the file) have been rolled up into a single file. The standard matrix (samples by gene-equivalents) is not found in FireBrowse. If the user is most interested in the original data (such as, reads or probe methylation values) FireBrowse is a good place to get this data, without dealing with arcane SDRF formats.

cBio provides access to data, but that data is often compiled together across platforms based on the genetic component being evaluated, and summarized, such as providing TP53 copy number data.

3.3 I am interested in data from specific papers or studies, which system should I look at?

cBio provides data from certain studies and papers, including the main TCGA paper for each tissue type, and provides Pubmed links for those studies and papers.

3.4 I am interested in HG18 TCGA data, which system should I look at?

Gene Survey does not use HG18 data. MD Anderson TCGA Standardized Data (http://bioinformatics.mdanderson.org/TCGA/databrowser) provides access to TCGA data in a simple matrix format and the MD Anderson TCGA Batch Effects Website (http://bioinformatics.mdanderson.org/tcgambatch) provides batch effects and quality assurance oriented analysis of available TCGA HG18 data.

FireBrowse provides HG18 data for download and provides associated analyses for some tissue types. These files will require processing to convert them into matrices.

cBio data is not described as HG18 or HG19.

3.5 I am interested in HG19 TCGA data, which system should I look at?

Gene Survey uses HG19 data and provides associated analyses. MD Anderson TCGA Standardized Data (http://bioinformatics.mdanderson.org/TCGA/databrowser) provides access to TCGA data in a simple matrix format and the MD Anderson TCGA Batch Effects Website (http://bioinformatics.mdanderson.org/tcgambatch) provides batch effects and quality assurance oriented analysis of HG19 data.

FireBrowse provides HG19 data for download and provides associated analyses.

cBio data is not described as HG18 or HG19.

3.6 I am interested in accessing data using MATLAB, which system should I look at?

FireBrowse and Gene Survey do not directly support MATLAB, but provide for data download. The Gene Survey data, being a simple tab-delimited matrix is simpler to load into MATLAB. FireBrowse provides data in a variety of formats which generally require additional processing.

cBio provides a MATLAB API for leveraging their web service for access to their data.

3.7 I am interested in accessing data using R, which system should I look at?

Gene Survey provides an R package and a data download.

FireBrowse (https://github.com/mariodeng/FirebrowseR) and cBio (http://www.cbioportal.org/cgds_r.jsp) both provide R interfaces which leverage their web-based APIs to access the data.

3.8 I am interested in performing analysis of my own, TCGA, or modified TCGA data, which system should I look at?

Gene Survey's R package provides many routines MD Anderson's analysts are often called upon to perform on cross-disease type data. Most of the routines are oriented towards analysis across disease types.

FireBrowse and cBio do not provide analysis level routines for R or other languages. cBio does provide limited tools on their website for processing data.

Other R packages for processing TCGA data are available from CRAN, Bioconductor, and other sources. As time permits, these packages will be added to this evaluation.

3.9 What analyses are provided by each website?

Gene Survey provides pre-computed stripcharts for gene symbols or probes across tissue types and pre-computed heatmaps of probes for a particular gene across tissue types.

There are lots of pre-computed analyses provided by **FireBrowse**: reports, analysis profiles, and expression profiles.

Reports are the many analyses on the left side of the screen. There are lots of these listed in the Raw Notes section and include such varieties as different clustering reports, Gistic2 output on copy number (turning probes into genes), and comparisons between one data type (copy number) and another (mRNA). These reports cover one tissue type at a time.

Analysis Profiles are accessed through a text box taking the "cohort" (aka disease type) abbreviation and the button "View Analysis Profile" button at the upper right of the screen. This is a one page overview of a disease type with some interactivity. Each data type is represented, along with the samples, as either an annotation bar (clinical age, vital status, RNASeq, Methylation) or a heatmap (mutations). Users can sort in different ways and search for samples.

Expression Profiles are accessed from the top-left via a gene name text entry and a "View Expression Profile" button. This provides a cross-disease boxplot of a gene with options to switch between RSEM and RPKM data, and some simple sorting and options to turn portions of the boxplots on or off.

cBio provides a large number of interactive analyses on their website. This seems to be the greatest draw for cBio, in that a non-statistician or non-programmer can go to their site and look at informative graphics about data from certain studies. Plots can be exported as SVG or PDF.

One drawback is that for cross-study queries, the results are often either split between the studies or intermingled. That is, the data is either kept split, with the analysis separate for each study. Or, the data is intermingled and treated as a whole, without an easy way to compare one study against another since you cannot distinguish which data points are from which study.

Analyses include OncoPrint (visual mutation data), mutual exclusivity (table only), scatterplot and boxplot (type of plot is linked to type of data), Mutation Mapper lollipop plots, Co-Expression plots, Enrichments (tables looking for changes in mutation, copy-number, mRNA, or protein), Survival Plots, Network interaction diagram (similar to Cytoscape, with interactions and drug targeting), and IGV (Integrative Genomics Viewer) though this last seems non-functional.