MBatch 03 Data for MBatch: Standardized Data
Tod Casasent
2017-11-02-0845

# 1 Introduction

These instructions are aimed at people familiar with R and familiar with TCGA/GDC platforms and data types. They are intended to introduce the reader to the format for Standardized Data, including the data feature formats (such as, gene symbols and probe ids).

Standardized Data is the results of taking TCGA data from the DCC or GDC and converting it into a format read-to-use by analysts--in other words a matrix. Particularly in the case of the DCC, converting the cryptic, oft undocumented, collections of files into an analyzable matrix can take a lot of time. The origin of MBatch was in analyzing TCGA data for batch effects, hence the format of Standardized Data became the file format for MBatch.

The website http://bioinformatics.mdanderson.org/TCGA/databrowser provides TCGA DCC Standardized Data. Within this data, the "matrix_data.tsv" files contain the actual data while the "batches.tsv" files contain the batch information.

# 2 Standardized Data

Standardized Data comes from one of two sources. DCC Standardized Data is from the now defunct DCC as-submitted TCGA data. Details on the TCGA project are available at https://cancergenome.nih.gov/. GDC Standardized Data is from the GDC Data Portal. Details on the GDC project are available at https://gdc.cancer.gov/. In both cases, we transformed the original data into the standard data matrix described below. Standardized Data also has batch information available, also described below.

## 2.1 Standardized Data "Data Matrix" Format

The Standardized Data "Data Matrix" format is a tab delimited file. The first line of the file begins with a tab and contains sample identifiers. For Standardized Data, the sample identifiers are TCGA bar codes. Each subsequent row begins with a Feature Identifier and is followed by numeric data. Feature Identifiers are specific to the platform and explained later, but can be values such as Hugo Gene ids, probe ids, or microRNA identifiers.

This extract from the Data Matrix format shows four sample ids and five feature ids. Note that the first blank cell indicates the starting tab for the sample identifiers line.

| | TCGA-OR-A5J2-01A-21-A39K-20 | TCGA-OR-A5J3-01A-21-A39K-20 | TCGA-OR-A5J6-01A-41-A39K-20 | TCGA-OR-A5J7-01A-21-A39K-20 |
|---|---|---|---|---|
| 14-3-3_beta-R-V | 0.211404 | -0.14778 | 0.220188 | -0.02738 |
| 14-3-3_epsilon-M-C | -0.03151 | -0.12861 | -0.0762 | -0.02275 |
| 14-3-3_zeta-R-V | -0.01203 | 0.032791 | -0.34541 | 0.136629 |
| 4E-BP1-R-V | 0.589134 | 0.365167 | 0.297887 | 7.34E-05 |
| 4E-BP1_pS65-R-V | -0.13521 | 0.182058 | -0.23654 | -0.0974 |

## 2.2 Standardized Data Batch File Format

The Standardized Data Batch File format is also a tab delimited file. The first line of the file contains the sample id column id and batch type identifiers, none of which should contain spaces. The first entry should be the "Sample" column, which contains sample ids. For TCGA data (from the DCC and the GDC), the other batch type identifiers are Type, BatchId, PlateId, ShipDate, and TSS.

| Sample | Type | BatchId | PlateId | ShipDate | TSS |
|---|---|---|---|---|---|
| TCGA-OR-A5J2-01A-21-A39K-20 | 1 | 304 | A39K | 5/7/2014 | OR - University of Michigan |
| TCGA-OR-A5J3-01A-21-A39K-20 | 1 | 304 | A39K | 5/7/2014 | OR - University of Michigan |
| TCGA-OR-A5J6-01A-41-A39K-20 | 1 | 304 | A39K | 5/7/2014 | OR - University of Michigan |
| TCGA-OR-A5J7-01A-21-A39K-20 | 1 | 304 | A39K | 5/7/2014 | OR - University of Michigan |

# 3 Data Features

DCC and GDC data features in matrix_data.tsv files vary. Feature components in <> are required. Components in {} are optional. If optional components are missing, no component separators (- or | or .) will exist.

## 3.1 DCC Features

For DCC Standardized Data, the following features are used for the following data types. A "mature mir id" is a mir Id cut off at the third dash, so that, for example, hsa-miR-299-3p becomes hsa-miR-299. Some datasets using <gene symbol>|<HGNC id> will have "?" for HGNC ids that lack a corresponding HUGO gene symbol.

| Data Type | Feature |
|---|---|
| bisulfiteseq/illuminahiseq_wgbs_percmethNOxy/Level_3 | <chromosome>-<start>-<end> |
| bisulfiteseq/illuminahiseq_wgbs_percmethWxy/Level_3 | <chromosome>-<start>-<end> |
| cna/cgh-1x1m_g4447a_nocnvNOxy/Level_3 | <gene symbol>\|{chromosome}\|{EnsemblId} |
| cna/hg-cgh-244a_hmsNOxy/Level_3 | <gene symbol>\|{chromosome}\|{EnsemblId} |
| cna/hg-cgh-244a_hmsWxy/Level_3 | <gene symbol>\|{chromosome}\|{EnsemblId} |
| cna/hg-cgh-244a_mskccNOxy/Level_3 | <gene symbol>\|{chromosome}\|{EnsemblId} |
| cna/hg-cgh-415k_g4124a_NOxy/Level_3 | <gene symbol>\|{chromosome}\|{EnsemblId} |
| cna/hg-cgh-415k_g4124a_Wxy/Level_3 | <gene symbol>\|{chromosome}\|{EnsemblId} |
| cna/illuminahiseq_dnaseqc_hg19NOxy/Level_3 | <gene symbol>\|{chromosome}\|{EnsemblId} |
| cna/illuminahiseq_dnaseqc_hg19Wxy/Level_3 | <gene symbol>\|{chromosome}\|{EnsemblId} |
| exon/huex-1_0-st-v2_gene/Level_3 | <gene symbol> |
| methylation/humanmethylation27_hg19NOxy/Level_3 | <gene symbol>-<probe id>-<chromosome>-<probe location> |
| methylation/humanmethylation27_hg19Wxy/Level_3 | <gene symbol>-<probe id>-<chromosome>-<probe location> |
| methylation/humanmethylation450_level3/Level_2 | <gene symbol>-<probe id>-<chromosome>-<probe location> |
| methylation/humanmethylation450_methNOxy/Level_3 | <gene symbol>-<probe id>-<chromosome>-<probe location> |
| methylation/humanmethylation450_methWxy/Level_3 | <gene symbol>-<probe id>-<chromosome>-<probe location> |
| mirna/h-mirna_8x15k_gene/Level_3 | <mir id> |
| mirna/h-mirna_8x15kv2_gene/Level_3 | <mir id> |
| mirnaseq/illuminaga_mirnaseq_isoform/Level_3 | <mature mir id>.<MIMAT id> |
| mirnaseq/illuminahiseq_mirnaseq_isoform/Level_3 | <mature mir id>.<MIMAT id> |
| mutations/illuminaga_dnaseq_automated_hgsc.bcm.edu/Level_2 | <gene symbol> |
| mutations/illuminaga_dnaseq_automated_ucsc.edu/Level_2 | <gene symbol> |
| mutations/illuminaga_dnaseq_curated_broad.mit.edu/Level_2 | <gene symbol> |
| mutations/illuminaga_dnaseq_curated_genome.wustl.edu/Level_2 | <gene symbol> |
| mutations/illuminaga_dnaseq_curated_hgsc.bcm.edu/Level_2 | <gene symbol> |
| mutations/illuminahiseq_dnaseq_automated_bcgsc.ca/Level_2 | <gene symbol> |
| mutations/illuminahiseq_dnaseq_automated_genome.wustl.edu/Level_2 | <gene symbol> |
| mutations/illuminahiseq_dnaseq_automated_sanger.ac.uk/Level_2 | <gene symbol> |
| mutations/mixed_dnaseq_automated_hgsc.bcm.edu/Level_2 | <gene symbol> |
| mutations/mixed_dnaseq_curated_hgsc.bcm.edu/Level_2 | <gene symbol> |
| mutations/solid_dnaseq_curated_hgsc.bcm.edu/Level_2 | <gene symbol> |
| protein_exp/mda_rppa_core_ProteinExpression/Level_3 | <antibodies> |
| rnaseq/illuminaga_rnaseq_bcgscGeneRPKM/Level_3 | <gene symbol>\|<HGNC id> |
| rnaseq/illuminaga_rnaseq_bcgscGeneRPKMv2/Level_3 | <gene symbol>\|<HGNC id> |
| rnaseq/illuminaga_rnaseq_uncGeneRPKM/Level_3 | <gene symbol>\|<HGNC id> |
| rnaseq/illuminahiseq_rnaseq_bcgscGeneRPKMhg18/Level_3 | <gene symbol>\|<HGNC id> |
| rnaseq/illuminahiseq_rnaseq_bcgscGeneRPKMhg19/Level_3 | <gene symbol>\|<HGNC id> |
| rnaseq/illuminahiseq_rnaseq_uncGeneRPKM/Level_3 | <gene symbol>\|<HGNC id> |
| rnaseqv2/illuminaga_rnaseqv2_gene/Level_3 | <gene symbol>\|<HGNC id> |
| rnaseqv2/illuminaga_rnaseqv2_isoform/Level_3 | <gene symbol>\|<HGNC id> |
| rnaseqv2/illuminahiseq_rnaseqv2_gene/Level_3 | <gene symbol>\|<HGNC id> |
| rnaseqv2/illuminahiseq_rnaseqv2_isoform/Level_3 | <gene symbol>\|<HGNC id> |
| snp/genome_wide_snp_6_hg18nocnvNOxy/Level_3 | <gene symbol>\|{chromosome}\|{EnsemblId} |
| snp/genome_wide_snp_6_hg18nocnvWxy/Level_3 | <gene symbol>\|{chromosome}\|{EnsemblId} |
| snp/genome_wide_snp_6_hg19nocnvNOxy/Level_3 | <gene symbol>\|{chromosome}\|{EnsemblId} |
| snp/genome_wide_snp_6_hg19nocnvWxy/Level_3 | <gene symbol>\|{chromosome}\|{EnsemblId} |
| snp/humanhap550_NOxy/Level_3 | <gene symbol>\|{chromosome}\|{EnsemblId} |
| snp/humanhap550_Wxy/Level_3 | <gene symbol>\|{chromosome}\|{EnsemblId} |

| Data Type | Feature |
|---|---|
| totalrnaseqv2/illuminahiseq_totalrnaseqv2_gene/Level_3 | <UCSC id> |
| totalrnaseqv2/illuminahiseq_totalrnaseqv2_isoform/Level_3 | <UCSC id> |
| transcriptome/agilentg4502a_07_1_gene/Level_3 | <gene symbol> |
| transcriptome/agilentg4502a_07_2_gene/Level_3 | <gene symbol> |
| transcriptome/agilentg4502a_07_3_gene/Level_3 | <gene symbol> |
| transcriptome/hg-u133_plus_2_gene/Level_3 | <gene symbol> |
| transcriptome/ht_hg-u133a_gene/Level_3 | <gene symbol> |
| transcriptome/illuminaga_mrna_dge_gene/Level_3 | <gene symbol> |

## 3.2 GDC Features

| Data Type | Feature |
|---|---|
| GeneExpressionQuantification/HTSeq-Counts | <gene symbol>\|<HGNC id> |
| GeneExpressionQuantification/HTSeq-FPKM | <gene symbol>\|<HGNC id> |
| GeneExpressionQuantification/HTSeq-FPKM-UQ | <gene symbol>\|<HGNC id> |
| IsoformExpressionQuantification/BCGSCmiRNAProfiling | <mir id>\|<MIMAT id or non-mature mir type> |
| MaskedSomaticMutation/MuSEVariantAggregationandMasking | <gene symbol> |
| MaskedSomaticMutation/MuTect2VariantAggregationandMasking | <gene symbol> |
| MaskedSomaticMutation/SomaticSniperVariantAggregationandMasking | <gene symbol> |
| MaskedSomaticMutation/VarScan2VariantAggregationandMasking | <gene symbol> |
| MethylationBetaValue/Liftover-IlluminaHumanMethylation27 | <probe id> |
| MethylationBetaValue/Liftover-IlluminaHumanMethylation450 | <probe id> |
| miRNAExpressionQuantification/BCGSCmiRNAProfiling | <mir id> |