

Sage Bionetworks Curated Data Packet Description

Introduction

All datasets released through the Sage Bionetworks Repository have been aggregated into a simple, uniform format, making them more widely accessible than in their source files. These data have undergone a series of integrity checks to identify and correct common problems, but are otherwise preserved in the rawest form available (for more information, please access the Curation Guidelines through the Repository website, <http://sagebase.org/commons/repository.php>). Herein we describe the directory structure and data files that define a “curated” data packet.

At its core, a data packet consists of a set of files containing data that have been gathered during a scientific experiment or clinical trial conducted on a group of participants. In addition to data files, a packet contains metadata files including licenses, data dictionaries, readme files, citations and scientific publications. The files in a packet are organized in a controlled directory structure designed to convey information about each file’s contents and its relationship to constituent files.

Directory Structure

The directory structure for each data packet is determined by the content of the

Figure 1: Directory listing for a representative data packet

```
+--DAT_015__pomp_breast_cancer
+-release_01_2011-01-24
+-pomp_breast_cancer
+-readme.txt
+-sage_bionetworks_user_agreement.pdf
+-cc_license.txt
+-ap_license.txt
+-version.txt
+-individuals.txt
|
+-expression
| +-illumina
|   +-gene_expression_min_signal.txt
|   +-gene_expression_avg_signal.txt
|   +-gene_expression_max_signal.txt
|   +-gene_expression_narrays.txt
|   +-gene_expression_array_stdev.txt
|   +-gene_expression_bead_stdev.txt
|   +-gene_expression_detection.txt
|   +-reporter.txt
|
+-genotype
| +-marker.txt
| +-genotype_call.txt
|
+-phenotype
| +-phenotype.txt
| +-description.txt
|
+-cnv
+-nimblegen
+-cnv_PM_532nm.txt
+-cnv_PM_635nm.txt
+-probe.txt
```

dataset. A separate directory is created for each data layer (genotype, phenotype, etc). Within each data layer, data is collected by source (e.g., microarray technology). All data files and data dictionaries for each source are presented together. Supplementary files including the abstract, citation, readme, and terms of use are provided in the top level of the dictionary. Information about the technologies used to collect the data and any notes from the curation team are located in the readme. An example of a simple directory structure from an actual data packet is included Figure 1.

Data Packet File Types

Data Files

Data files are tab-delimited text files, containing a rectangular data matrix, that use row and column headers to uniquely identify each cell in the matrix. Data files are organized such that each column contains data for a single individual and each row contains a single type of data, whether it is a phenotype, genotype or expression value. The intersecting cell shows the value of the data for the given individual. The first row and first column of a data file contains the column and row headers. These are text string identifiers that uniquely identify each row and columns. Missing values are indicated by “NA”.

Dictionary Files

Dictionary files are tab-delimited text files listing the identifiers used in a data file and providing their meanings. There is a 1:1 relationship between the identifiers used in the data file and the identifiers listed in the dictionary file. There is dictionary file for each data-type in a data packet. The data-type identifiers are unique for a given data-type but at this time we have no guaranty that they are unique across multiple data-files or across multiple data packets.

In contrast, there is a single, study-wide “individuals” dictionary listing all participant identifiers used throughout the packet. Each participant maps to exactly one identifier and the same participant identifiers are used in every data file in a given packet. For time-series study, the same participants will have multiple Participant-ID, one for each time point. Individual identifiers must be unique for a given study or data packet. However they may not be unique across multiple data packets.

Supplementary Files

Additional information relevant to the data in a packet that are not contained in the dictionary files is included in supplementary files.

These supplementary files include:

- Abstract: High level description of the study
- Citations: Selected scientific publications.
- Readme files: Description of the data packet content and notes from the curator regarding the data packet (optional).
- Terms of Use: We provide specific terms of use for each data packet that users must agree to prior to be allowed to download the data. The Terms of Use document highlights specific conditions to data sharing from the Licenses and some expectation of conduct in handling the data.
- Licenses:
 - Full text of the creative commons license (CCBY3.0) under which we share the data
 - Full text of the Apache license under which we share codes and scripts.
- Acknowledgements: Specific text to use in presentations and publications to recognize the contributions of data contributors.
- Version file: indicates the various versions of data release and changes to the data packet.

Data Layers

A “data layer” refers to a specific type of data contained in a data packet. Examples of common layers are Gene Expression, Genotype and Phenotype data.). In data packets, controlled vocabulary is used to refer to a data layer type (**Error! Reference source not found.**1 and **Error! Reference source not found.**). It is common for data layers to require multiple data values to fully specify the results for each individual. Each data value for a layer is contained in a separate data file, each of which is required to have the identical dimensions and contain the same row and column identifiers. There is no guarantee that the rows and columns for each file will be in the same order. File descriptions for the most common data types are listed below and described in detail in Tables 1 and 2. Screenshots of data and dictionary files for an example data packet are provided in Figures 2-5.

Genotype

The genotype data layer contains SNP genotypes or sequencing results and should contain results summarized into genotype calls and/or raw signals from microarrays. Genotype data is typically provided through a single data file, entitled, “genotype_call.txt” (Table 2).

Gene Expression

The gene expression data layer contains data for mRNA expression. Because expression

Table 1: Description of the expression data layer. Data dictionaries for agilent and illumina data contain reporter_ids (reporter.txt files) and for Affymetrix data contain feature_ids (feature.txt file).

Data layer	Platform	Required data values	Data file name	Data description
expression	agilent	g_intensity	expression_g_intensity.txt	green intensity
		r_intensity	expression_r_intensity.txt	red intensity
		mlavg	expression_mlavg.txt	mean log average intensity between channels across a flour-reversed pair
		mlratio	expression_mlratio.txt	mean log expression ratio across a flour-reversed pair
		pvals	expression_pvals.txt	pvalue associated with mlratio ratio
	affymetrix	pm	expression_pm.txt	perfect match intensity value
		mm	expression_mm.txt	mismatch intensity value
		sd	expression_sd.txt	standard deviation
	illumina	detection	expression_detection.txt	detection
		avg_signal	expression_avg_signal.txt	average signal
		min_signal	expression_min_signal.txt	minimum signal
		max_signal	expression_max_signal.txt	maximum signal
		array_stdev	expression_array_stdev.txt	array standard deviation
bead_stdev		expression_bead_stdev.txt	bead standard deviation	
narrays		expression_narrays.txt	number of arrays	

data requires multiple data parameters to be fully described, this data layer will typically include multiple data files. Descriptions of data files for the most common platforms are described in Table 1. For example, expression data collected using an Affymetrix platform is fully described by providing a perfect match intensity value, mismatch intensity value and standard deviation intensity value for each feature. This data would be provided in three text files, labeled, “expression_pm.txt, expression_mm.txt, and expression_sd.txt”. These files would be located in a folder entitled, “Affymetrix” within the “expression” layer of the directory structure.

Phenotype

The phenotype data layer contains all clinical traits and covariates in a single file entitled, “phenotype.txt”. Sage Bionetworks requires that all data are stripped of potential identifiers per HIPAA rules prior to contribution to the Sage Bionetworks Repository.

Table 1: Description of genotype, cnv, mirna and phenotype data layers. Data dictionaries for genotype files are called “marker.txt”, for CNV data are called “probe.txt”, for miRNA data are called “reporter.txt”, and for phenotype data are called “description.txt”.

Data layer	Platform	Required data values	Data file name	Data description
genotype	Affymetrix	call	genotype_call.txt	genotype call
	perlegen	call	genotype_call.txt	genotype call
	nimblegen	call	genotype_call.txt	genotype call
cnv	agilent	rawrat	cnv_rawrat.txt	raw ratio
		log2rat	cnv_log2rat.txt	log base 2 transformed ratio
		log2stddev	cnv_log2stddev.txt	standard deviation of the log base 2 ratio
		nreplic	cnv_nreplic.txt	number of replicates
		bad_p	cnv_bad_p.txt	bat probe flag
	nimblegen	PM_532nm	cnv_PM_532nm.txt	perfect match intensity at 532nm
		PM_635nm	cnv_PM_635nm.txt	perfect match intensity at 635nm
mirna	agilent	glsGeneDetected	mirna_glsGeneDetected.txt	flag indicating whether the gene was detected
		gTotalGeneSignal	mirna_gTotalGeneSignal.txt	total gene signal
		gMeanSignal	mirna_gMeanSignal.txt	mean signal
		gMedianSignal	mirna_gMedianSignal.txt	median signal
		glsSaturated	mirna_glsSaturated.txt	flag indicating where reporter was saturated
phenotype	other	NA	phenotype.txt	clinical phenotypes and covariates

Figure 2: Screen shot of a phenotype dictionary file

trait_id	trait_name	timepoint	transform	adjustment	trait_description	units
rep	rep	NA	NA	NA	NA	NA
diet	diet	NA	NA	NA	diet of the mouse. Either "High" for high fat diet or "Low" for control diet. HFD=45% of total calories from fat, 20% from protein, and 35% from carbohydrates. Control=10% of total calories from fat, 20% from protein, and 70% from carbohydrates.	NA
met	met	end	normal	None	number of pulmonary metastasis counted at sacrifice	NA
met_resid	met_resid	end	normal	None	NA	NA
ave_met_density	ave_met_density	end	normal	None	average metastatic density	NA
amd_resid	amd_resid	end	normal	None	average metastatic density resid	NA
ttw	ttw	end	normal	None	total weight of all tumors	grams
tumor_ax	tumor_ax	end	normal	None	total weight of the axillary tumors	grams
tumor_ing	tumor_ing	end	normal	None	total weight of the inguinal tumors	grams
weight_at_sacrifice	weight_at_sacrifice	end	normal	None	body weight at sacrifice (~11 weeks for females, ~14 weeks for males)	grams
percent_fat	percent_fat	end	normal	None	percent body fat at sacrifice	NA
tumor_count	tumor_count	end	normal	None	count of tumors at sacrifice	NA
ave_met_area	ave_met_area	end	normal	None	average metastatic area	NA
ave_mets	ave_mets	end	normal	None	average number of metastatic tumors	NA
ave_lung_area	ave_lung_area	end	normal	None	average lung area	NA
pre_tumor_percent_fat	pre_tumor_percent_fat	W7	normal	None	pre tumor percent fat	NA
tumor_onset_in_days	tumor_onset_in_days	NA	normal	None	days until first tumor was observed in mammary glands.	days
lean_mass_pre_tumor	lean_mass_pre_tumor	W7	normal	None	lean mass pre tumor	grams
lean_mass_at_sac	lean_mass_at_sac	end	normal	None	lean mass at sacrifice	grams
change_in_lean	change_in_lean	end	normal	None	change in lean mass	grams
fat_in_grams	fat_in_grams	end	normal	None	fat pad weight	grams
liver_wt	liver_wt	end	normal	None	weight of liver	grams
liver	liver	end	normal	None	weight of liver adjusted for body weight	grams
weight_3w	weight_3w	W3	normal	None	body weight at 3 weeks	grams
weight_6w	weight_6w	W6	normal	None	body weight at 6 weeks	grams
weight_9w	weight_9w	W9	normal	None	body weight at 9 weeks	grams

Figure 3: Screen shot of a CNV data file

probe_id	B	C	D	E	F	G	H	I	J
208	306	608-3	708	808	1508	3005-8	3405	3509	
CHR01FS003001832	2051.67	2222.44	1623.67	2639.11	1455.56	2111.78	2535	1905.56	2528.56
CHR01FS003018759	2358.78	2534.33	1983.78	2793.89	1896	2317.33	2797.89	1923.67	2877.78
CHR01FS003036253	1974.89	2110.22	1605.78	2834.67	1431.56	1576.44	2058.89	1549.22	2715.11
CHR01FS003041992	2514.44	3570.11	3012.89	4498.44	2731	3562	2894.89	3384.22	4504.22
CHR01FS003053606	3601.78	3164.78	2303.22	3807.78	2805.56	3420.22	2935.67	1891.11	3906.56
CHR01FS003065156	2285.56	2270.67	1598.78	2895.33	1748.56	1985.44	2438.56	1897.78	3128.44
CHR01FS003076536	3410.44	3664.56	2881.33	4104	2787.22	4061.56	3066.56	2716.22	3455.78
CHR01FS003087994	5185.89	4048	3393.89	3983	3741.56	5023.44	2956.67	4398.33	2554.44
CHR01FS003093673	5488.56	5167.11	4089.33	6820.67	4618.56	7096	5071.67	6626.44	4900.56
CHR01FS003105423	7739.22	6941.11	5710.78	8645.56	5414.56	7634.56	5721.78	5962.11	6765.11
CHR01FS003110919	2610	2574.33	1820.44	3299.22	1855.33	2876.78	2475.33	2035.44	3071.22
CHR01FS003116559	2561	3017.56	2278	3135.33	2018.22	2780.11	3403	2608.44	3674
CHR01FS003122452	2679.89	2978.89	2031.33	3317.44	2267.22	3220.11	3152.33	2588.11	3466.22
CHR01FS003134224	2094.78	2169.56	2135	3617.56	2316.78	1854.56	1689.67	2127.33	3005.11
CHR01FS003156928	2963.22	3808.67	2229.89	4726.67	2323.44	3479.22	3089.78	2940.56	3726.56
CHR01FS003162762	2981.11	4104.67	2898	4755.44	2277.22	3132.44	2885.78	3085.44	4691.56
CHR01FS003174487	1700.44	2028.33	1387.33	2401.33	1479.67	1763.56	1684.33	1339	3236.78
CHR01FS003180251	2045.89	2506	1585.67	3124	2140.78	2092.56	1813	1459.89	2543.44
CHR01FS003185980	2095.44	2906	2078.56	3340.11	1881.11	2472.11	2762.33	1658.22	3460.22
CHR01FS003191525	2578.89	2880.44	1732.22	3024.22	2101.78	2470.56	3251.78	1897.11	2851.33
CHR01FS003197138	3503.11	3528.56	2558.56	4453.89	3107.89	3620.33	2740.67	2428.11	4290.78
CHR01FS003203195	2465.67	3019.56	1593.89	2946	1749.22	2221.33	3155.56	2016.33	2882.44
CHR01FS003208636	4215.89	4776.22	2406.89	5563.22	3278.56	3573.89	3873.56	3017.67	4379.78
CHR01FS003214407	5055.56	4491.22	3624.22	7297.56	3647.44	4251	3456.56	2898	6260.89
CHR01FS003220004	1550.44	2245.56	1561.56	2947.78	1172.44	1642.78	2031.44	1527	2957.11
CHR01FS003237597	5313.33	5769.56	5752.89	8602.44	6573.78	6394.22	4090.78	4721.56	7261.44
CHR01FS003243214	2450	2856.89	1984	2604.44	1874	2951.22	2872.22	2417.67	3206.33
CHR01FS003254921	3830.67	4506.44	2523.67	4956.67	2794.11	3392.89	3112.56	3066.11	4835.11
CHR01FS003260419	4099.56	5091	4497.89	7546.67	3890.22	5622.78	4565.11	4534.44	5387.56
CHR01FS003277959	2672.33	2988.44	2570.89	3788.44	1992.33	3252.78	3272.78	2654.11	3175.89
CHR01FS003283516	7081.11	8614.33	6570.22	9861.56	6153.67	9314.89	8655.67	9122.11	8425.56
CHR01FS003289244	3677	3008.56	3226.56	5246.67	2828	3725.67	2322.33	3180.89	4173.67
CHR01FS003300897	2479.44	2745.78	1602.78	3358.67	2301.11	2109.44	3369.11	1812.44	2762
CHR01FS003312249	3450.22	3145.22	2748	4889.56	2200.89	3609.33	2912.44	2764	4653.44
CHR01FS003323571	3951.56	3831.56	2271.22	3995.89	2402.56	3349.56	3063.33	2970.11	3577.78
CHR01FS003329326	3927.44	3929	5821.44	5701.22	4369.89	5909.33	3169.11	4874.89	3696.33

Figure 4: Screen shot of an individuals dictionary

The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G
1	individual_id	cnv_cy3_sample_name	cnv_cy5_sample_name	gene_expression	genotype	phenotype	cnv
2	105	NA	NA	yes	yes	yes	no
3	106	106a	106t	yes	yes	yes	yes
4	1104	1104a	1104t	yes	yes	yes	yes
5	1107	NA	NA	yes	yes	yes	no
6	1204	1204a	1204t	yes	yes	yes	yes
7	1206	1206a	1206t	yes	yes	yes	yes
8	1308	NA	NA	yes	yes	yes	no
9	1409	NA	NA	yes	yes	yes	no
10	1507	NA	NA	yes	yes	yes	no
11	1508	1508-6	1508-26	yes	yes	yes	yes
12	1609	NA	NA	yes	yes	yes	no
13	1804	NA	NA	yes	yes	yes	no
14	1808	NA	NA	yes	yes	yes	no
15	1905	1905a	1905t	yes	yes	yes	yes
16	208	208-1	208-21	yes	yes	yes	yes
17	2206	NA	NA	yes	yes	yes	no
18	2209	NA	NA	yes	yes	yes	no
19	2303	NA	NA	yes	yes	yes	no
20	2304	NA	NA	yes	yes	yes	no
21	2406	NA	NA	yes	yes	yes	no

