

The Cancer Genome Atlas Pilot Project
Broad Institute of MIT & Harvard
Raw Data Description

Platforms: Affymetrix SNP 6.0 Array & Affymetrix HTA Array

Raw Data

The CEL file stores the results of the intensity calculations on the pixel values of the DAT file. This includes an intensity value, standard deviation of the intensity, the number of pixels used to calculate the intensity value, a flag to indicate an outlier as calculated by the algorithm and a user defined flag indicating the feature should be excluded from future analysis. The file stores the previously stated data for each feature on the probe array.

The Cancer Genome Atlas Pilot Project
Broad Institute of MIT & Harvard
Normalized Data Description

Platforms: Affymetrix SNP 6.0 arrays & Affymetrix HTA arrays

Analyses: Copy Number Analysis & Loss of Heterozygosity

Normalized Data

Affymetrix SNP 6.0 – Copy number analysis (.snp File Format)

This is a tab-delimited file format that contains SNP array data. It contains one row for each SNP (or copy number probe) and two columns for each SNP array: the intensity value and the call value. It is organized as follows:

1. The first line contains a list of labels identifying the SNP arrays.
 - Line format: SNP (tab) Chromosome (tab) PhysicalPosition (tab) (array_1_name) (tab) (array_1_name) Call (tab) ... (array_N_name) (tab) (array_N_name) Call
 - For example: SNP (tab) Chromosome (tab) PhysicalPosition (tab) MYNAH_p_Affy_plate_9_Mapping250K_Sty_A01_49068 (tab) MYNAH_p_Affy_plate_9_Mapping250K_Sty_A01_49068 Call (tab) ... MYNAH_p_Affy_plate_9_Mapping250K_Sty_A01_49084 (tab) MYNAH_p_Affy_plate_9_Mapping250K_Sty_A01_49084 Call
2. The rest of the SNP file contains one row of data for each probe including the probe intensity value and the SNP calls generated by the SNP microarray scanning software (such as Affymetrix's GeneChip software).
 - Line format: (snp) (tab) (chromosome) (tab) (position) (tab) (array_1_intensity) (tab) (array_1_call) (tab) ... (array_N_intensity) (tab) (array_N_call)
 - For example: SNP_A-4249904 (tab) 17 (tab) 41420045 (tab) 676.42 (tab) AB (tab) ... 1145.411 (tab) AA
 - Copy number probes are named CN_xxxx and will have a blank in the allele columns.

Normalized Data

Affymetrix SNP 6.0 – Loss of Heterozygosity (.loh File Format)

This is a tab-delimited file format that contains the output results of the LOH module. The LOH module, a SNP analysis module, detects loss of heterozygosity (LOH). The LOH file format is organized as follows:

1. The first line contains a list of labels identifying the paired samples.
 - Line format: SNP (tab) Chromosome (tab) PhysicalPosition (tab) (pair_1_name) (tab) ... (pair_N_name)
 - For example: SNP (tab) Chromosome (tab) PhysicalPosition (tab) SM-12VZ (tab) SM-12W1

2. The rest of the SNP file contains one row of data for each probe.
 - Line format: (snp) (tab) (chromosome) (tab) (position) (tab) (pair_1_loh) (tab) ... (pair_N_loh)
 - For example: SNP_A-1855068 (tab) 17 (tab) 41089766 (tab) R (tab) R

LOH call values are:

- L (LOH): AB in normal and A or B in tumor
- R (Retention): AB in both normal and tumor or No Call in normal and AB in tumor
- C (Conflict): A or B in normal and AB in tumor
- N (Non-informative call): A or B in normal, or No Call in normal or tumor

Normalized Data

Affymetrix HTA arrays - Gene expression data [.res file Format (RMA expression level and Affymetrix Absent/Present call)]

The RES file format is a tab delimited file format that describes an expression dataset. The main difference between RES and GCT file formats is the RES file format contains labels for each gene's absent (A) versus present (P) calls as generated by Affymetrix's GeneChip software. The file is organized as follows:

of rows (probe sets) Scaling info

	A	B	C	D	E	F	G	H
1	Description	Accession	ALL_19759		ALL_23953		ALL_28373	
2		CH1999021515AA		CH1999021511AA/scale f	CH1999021507AA/sc	CH199902		
3		1000						
4	Semaphorin E	AB000220_at	36 A		39 A		39 A	
5	MNK1	AB000409_at	-299 A		-11 A		237 P	
6	VRK1	AB000449_at	57 A		274 P		311 P	
7	VRK2	AB000450_at	186 P		245 P		186 P	
8	mRNA, clone RES4-	AB000460_at	1647 P		2128 P		1608 P	
9	SH3 binding protein,	AB000462_at	137 A		-82 A		204 P	
10	mRNA, clone RES4-	AB000464_at	803 P		1489 P		322 P	
11	mRNA, clone RES4-	AB000466_at	-894 A		-969 A		-444 A	
12	mRNA, clone RES4-	AB000467_at	-632 A		-909 A		-254 P	
13	Zinc finger protein,	clAB000468_at	378 P		266 P		554 P	
14	Prostate differentiat	AB000584_at	-26 A		-181 A		16 A	
15	Cadherin FIB1, partia	AB000895_at	-691 A		-900 A		-58 A	
16	Cadherin FIB2, partia	AB000896_at	2 A		-237 A		-78 A	
17	Cadherin FIB3, partia	AB000897_at	-156 A		-156 A		-95 A	

Description: can be either column 1 or column 2)

Accession (probe set id) can be either column 1 or column 2. Must not have duplicate entries

Absent or present call info

2. The first line contains a list of labels identifying the samples associated with each of the columns in the remainder of the file. Two tabs (\t\t) separate the sample identifier labels because each sample contains two data values (an expression value and a present/marginal/absent call).
 - Line format: Description (tab) Accession (tab) (sample 1 name) (tab) (tab) (sample 2 name) (tab) (tab) ... (sample N name)
 - For example: Description Accession DLBC1_1 DLBC2_1 ... DLBC58_0
 3. The second line contains a list of sample descriptions. Currently, GenePattern ignores these descriptions.
 - Line format: (tab) (sample 1 description) (tab) (tab) (sample 2 description) (tab) (tab) ... (sample N description)
 - For example, our RES file creation tool places the sample data file name and scale factors in this row: MG2000062219AA
MG2000062256AA/scale factor=1.2172 ... MG2000062211AA/scale factor=1.1214
 4. The third line contains a number indicating the number of rows in the data table that is contained in the remainder of the file. Note that the name and description columns are not included in the number of data columns.
 - Line format: (# of data rows)
 - For example: 7129
 5. The rest of the data file contains data for each of the genes. There is one row for each gene and two columns for each of the samples. The first two fields in the row contain the description and name for each of the genes (names and descriptions can contain spaces since fields are separated by tabs). The description field is optional but the tab following it is not. Each sample has two pieces of data associated with it: an expression value and an associated Absent/Marginal/Present (A/M/P) call. The A/M/P calls are generated by microarray scanning software (such as Affymetrix's GeneChip software) and are an indication of the confidence in the measured expression value. Currently, GenePattern ignores the Absent/Marginal/Present call.
 - Line format: (gene description) (tab) (gene name) (tab) (sample 1 data) (tab) (sample 1 A/P call) (tab) (sample 2 data) (tab) (sample 2 A/P call) (tab) ... (sample N data) (tab) (sample N A/P call)
 - For example: AFFX-BioB-5_at (endogenous control) AFFX-BioB-5_at -104 A -152 A ... -44 A
-

The Cancer Genome Atlas Pilot Project
Broad Institute of MIT & Harvard
Segmented Data Description

Platforms: Affymetrix SNP 6.0 Array & Affymetrix HTA Array

Segmented Data

Copy number - Affymetrix SNP 6.0 Array

This is a tab-delimited file format that contains the output results of the GLAD module. The GLAD module, a SNP analysis module, runs the R package *Gain and Loss analysis of DNA* (GLAD) [Hupè et al., 2004], which detects segments of the genome which have altered copy numbers. The GLAD file format is organized as follows:

1. The first line contains a list of labels identifying the columns.
 - o Line format: Sample (tab) Chromosome (tab) Start.bp (tab) End.bp (tab) Num.SNPs (tab) Seg.CN
2. The rest of the file contains one row of data for each altered chromosomal region.
 - o Line format: (sample) (tab) (chromosome) (tab) (startPosition) (tab) (endPosition) (tab) (numberOfSNPs) (tab) (regionCN)
 - o For example: MYNAH_p_Affy_plate_9_Mapping250K_Sty_A02_49084 (tab) 17 (tab) 41419603 (tab) 36581538 (tab) 6427 (tab) 2.06

Segmented Data

Loss of Heterozygosity - Affymetrix SNP 6.0 Array (.seg-loh file)

This is a tab-delimited file format that contains the output results of a hidden Markov model to identify continuous segments of LOH. The file format is organized as follows:

1. The first line contains a list of labels identifying the columns.
 - o Line format: Sample (tab) Chromosome (tab) Start.bp (tab) End.bp (tab) Num.SNPs
2. The rest of the file contains one row of data for each region of LOH.
 - o Line format: (sample) (tab) (chromosome) (tab) (startPosition) (tab) (endPosition) (tab) (numberOfSNPs)