User Manual

MAMGED (Meta-Analysis of Microarray Gene Expression Data) is a tool written in R to perform meta-analysis of microarray gene expression data. The current version of this tool is able to handle data from three different platforms i,e Affymetrix, Codelink and Illumina. The process is to quantify the consistency of absolute expression calls (transcribed or not) made across experiments. The tool can be used to analyze the data from multiple studies across different types of microarray platforms to derive a consensus expression status of genes in a specific tissue or cell-type and a condition of interest. In addition, the tool extends the application of the method to provide a score for the consistency of differential expression pattern for each gene. The tool is able to handle both pre-processed and raw data. This document is divided into three sections; Section I deal with meta-analysis of Affymetrix data, Section II with Codelink, followed by Section III of Illumina data.

1. Section I:

Meta-analysis of microarray Affymetrix data is divided into three parts.

1.1 Meta-analysis of processed data:

The processed data from GEO is used to perform meta-analysis as per the methodology proposed by Acharya et al [1]. Scoring of an individual data file is based on the significance of p-value. A score of 2 is assigned for Present (p-value 0-0.05), -2 for Absent (p-value 0.065-1) and 0 for Marginal (p-value 0.051-0.0649) and then the cumulative score is calculated by combining the score of individual files. The cumulative scores are then sorted in descending order.

Four samples of pre-processed Affymetrix data are shown below. Data files can have probe sets, signal intensity, present/absent call and p-value present as show in Table 1. Files with probe sets, present/absent call and p-value or probe sets and p-values are also valid for meta-analysis, data samples of which are shown in Table 2 and Table 3. The data files can be in any order, sample shown in Table 4.

	A	B	C	D
1	ID_REF	VALUE	ABS_CALL	P-VALUE
2	10071_s_at	3473.6	P	0.000219
3	1053_at	643.2	P	0.000673
4	117_at	564	P	0.000322
5	1255_g_at	9.4	A	0.602006
6	1294_at	845.6	P	0.000468
7	1320_at	94.3	A	0.204022
8	1405_i_at	6546.2	P	0.000491
9	14312_at	54.1	P	0.003067
10	1438_at	461.3	P	0.000562
11	1552289_a_a	92.6639844	A	0.5
12	1552291_at	411.388556	P	0.003823
13	1552261_at	177.616892	M	0.054755
14	1552266_at	40.6103783	A	0.65
15	1552265_at	40.6103783	P	0.05101
16	1552266_at	40.6103783	P	0.001
17	1552266_at	40.6103783	P	0.001
18	1552263_at	435	P	0.002
19	1316_at	151.1	M	0.05447
20	117_at	564	P	0.000322
21	1053_at	643.2	P	0.000673
22	1438_at	461.3	P	0.000562
23	1552261_at	177.616892	M	0.005476
24	1255_g_at	9.4	P	0.002006
25	1320_at	94.3	A	0.05104

Table 1: Affymetrix processed sample file containing probe sets in first column, signal intensity in second column, present/absent calls in column third and finally p-value in column forth.

	A	B	C
1	ID_REF	VALUE	P-VALUE
2	10071_s_at	3473.6	0.000219
3	1053_at	643.2	0.000673
4	117_at	564	0.000322
5	1255_g_at	9.4	0.602006
6	1294_at	845.6	0.000468
7	1320_at	94.3	0.204022
8	1405_i_at	6546.2	0.000491
9	14312_at	54.1	0.003067
10	1438_at	461.3	0.000562
11	1552289_a_a	92.6639844	0.5
12	1552291_at	411.388556	0.003823
13	1552261_at	177.616892	0.054755
14	1552266_at	40.6103783	0.65
15	1552265_at	40.6103783	0.05101
16	1552266_at	40.6103783	0.001
17	1552266_at	40.6103783	0.001
18	1552263_at	435	0.002
19	1316_at	151.1	0.05447
20	117_at	564	0.000322
21	1053_at	643.2	0.000673
22	1438_at	461.3	0.000562
23	1552261_at	177.616892	0.005476
24	1255_g_at	9.4	0.002006
25	1320_at	94.3	0.05104

Table 2: Affymetrix processed sample file containing probe sets in first column, signal intensity in second column, and p-values in third column.

	A	B
1	ID_REF	P-VALUE
2	10071_s_at	0.000219
3	1053_at	0.000673
4	117_at	0.000322
5	1255_g_at	0.602006
6	1294_at	0.000468
7	1320_at	0.204022
8	1405_i_at	0.000491
9	14312_at	0.003067
10	1438_at	0.000562
11	1552289_a_at	0.500000024
12	1552291_at	0.003822926
13	1552261_at	0.054755469
14	1552266_at	0.65
15	1552265_at	0.05101
16	1552266_at	0.001
17	1552266_at	0.001
18	1552263_at	0.002
19	1316_at	0.05447
20	117_at	0.000322
21	1053_at	0.000673
22	1438_at	0.000562
23	1552261_at	0.005475547
24	1255_g_at	0.002006
25	1320_at	0.05104022

Table 3; Affymetrix processed data sample with probe sets in first column and p-values in second column

	A	B	C
1	ID_REF	VALUE	ABS_CALL
2	10071_s_at	3473.6	P
3	1053_at	643.2	P
4	117_at	564	P
5	1255_g_at	9.4	A
6	1294_at	845.6	P
7	1320_at	94.3	A
8	1405_i_at	6546.2	P
9	14312_at	54.1	P
10	1438_at	461.3	P
11	1552289_a_a	92.6639844	A
12	1552291_at	411.388556	P
13	1552261_at	177.616892	M
14	1552266_at	40.6103783	A
15	1552265_at	40.6103783	P
16	1552266_at	40.6103783	P
17	1552266_at	40.6103783	P
18	1552263_at	435	P
19	1316_at	151.1	M
20	117_at	564	P
21	1053_at	643.2	P
22	1438_at	461.3	P
23	1552261_at	177.616892	M
24	1255_g_at	9.4	P
25	1320_at	94.3	A

Table 4: Affymetrix processed data sample with probe sets in first column, signal intensity in second column and present/absent calls in third column

Simple five step process to perform meta-analysis of Affymetrix processed data.

1) Load processed text or CSV files (a minimum of 2 files are required)

2) Choose annotation file from the database (appropriate file recommended)

3) Choose an appropriate choice for uploading data (here Affymetrix need to be choosen)

4) Submit

5) Navigate through tab-set

The five-step process is shown below in the screen shots, Figure 1, Figure2 and Figure 3. After all the steps are followed, navigation need to be done through different tabs in order to check the source data, annotation data, summary and supplementary file. 'Source-data' is the data files loaded by the user (contents of only one file will be displayed), 'Annotation-data' shows the platform annotation file, 'Summary' is the meta-analysis information with cumulative score and finally 'Supplementary file' shows the full information of files with p-value and signal intensity or fold change in case of differential expression.

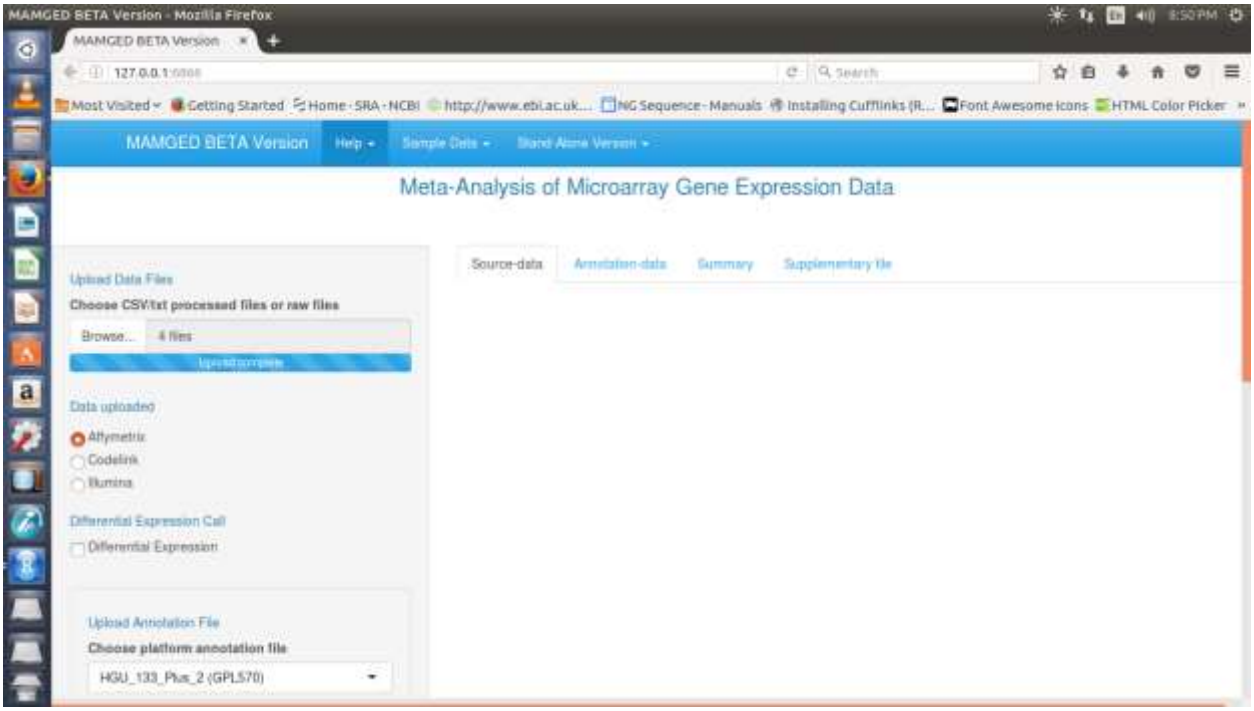

Figure 1: Parameter setting for absolute expression call for Affymetrix data

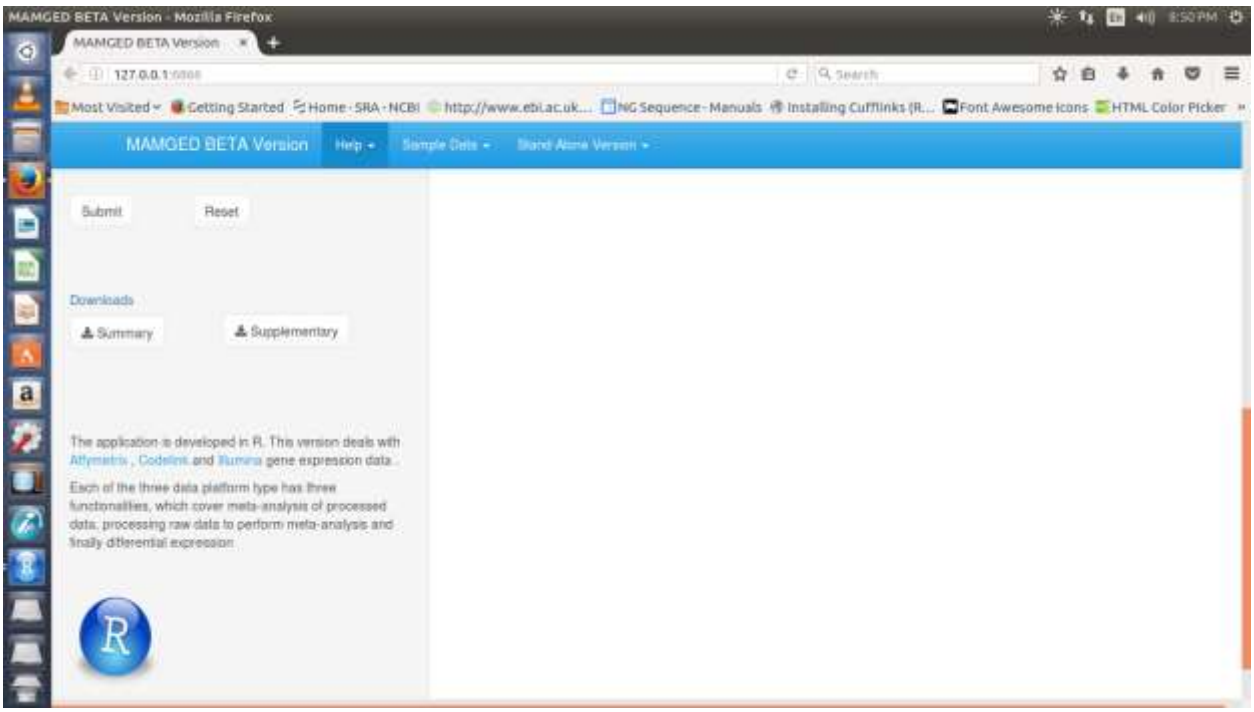

Figure 2: Parameter setting for absolute expression call for Affymetrix data

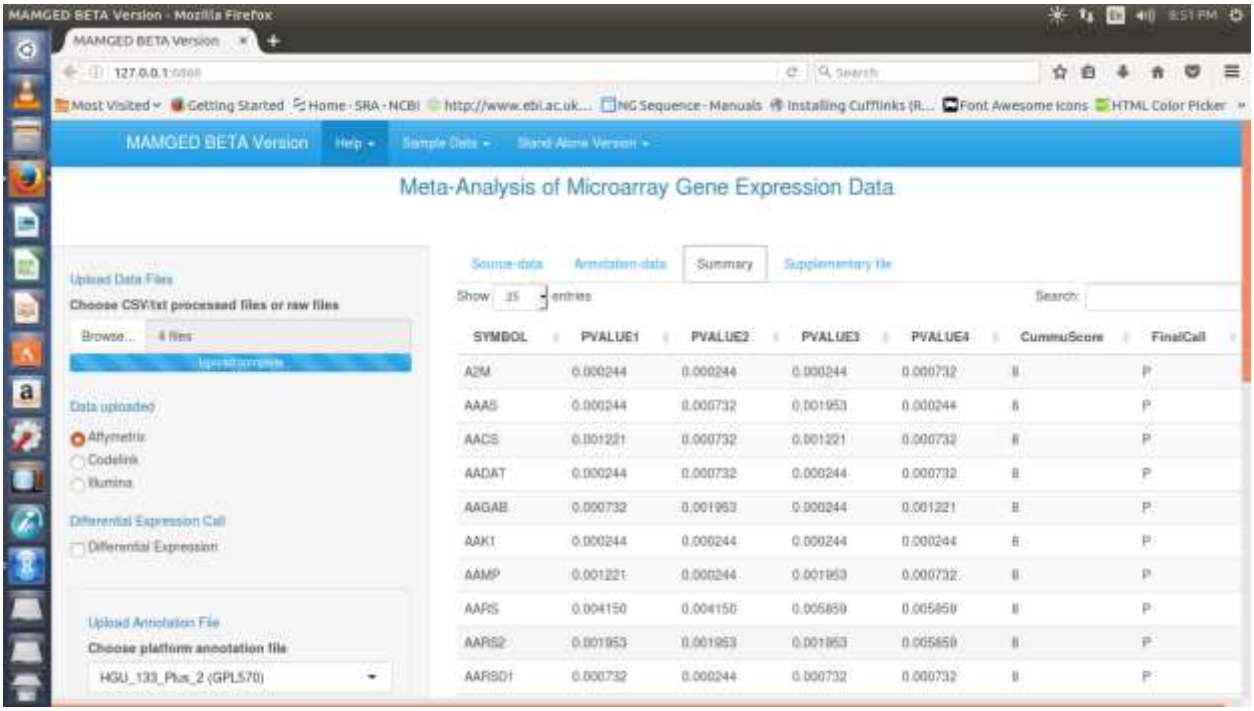

Figure 3: Absolute expression call summary generated. The column "SYMBOL" denotes the gene symbol, "PVALUE1" is the p-value of first sample, "PVALUE2" p-value of second sample and so on till "PVALUE4", "cummuScore" is the cumulative score based on four sample data files and "FinalCall" is the absolute expression call.

1.2 Meta-analysis of raw data:

In case of non-availability of processed data or user wants to use raw data, CEL files need to be supplied as input. The raw .CEL files are pre-processed and normalized before performing meta-analysis. All the process will go internally and user just need to follow the above five steps. Please note that .CEL file does not follow the pattern shown in the sample files above.

1.3 Differential expression Meta-analysis:

For differential expression two types of files need to be submitted: the data files and the target files. Data files are supposed to be CEL files, while target files in .txt format contain information of data files with corresponding conditions. At least two set of files containing at least two different conditions is supported. Sample target file which need to be prepared by the user is shown below, where the first column containing sample names and the second with treatment conditions. The **Target file 1** contains information about the files of study 1 and **Target file 2** containing information of study 2 of Affymetrix platform.

Filenames	Treatment
GSM248652.CEL	A
GSM248655.CEL	A
GSM248238.CEL	A
GSM248650.CEL	B
GSM248651.CEL	B
GSM248661.CEL	C
GSM248660.CEL	C

Target file 1

Filenames	Treatment
GSM248653.CEL	A
GSM248651.CEL	A
GSM248659.CEL	B
GSM248660.CEL	B
GSM248661.CEL	B

Target file 2

Steps to follow

1) Load .CEL and target files (more than 2 files recommended)

2) Choose annotation file

3) Choose data uploaded (Affymetrix)

4) Check box of differential expression need to be checked

5) Choose fold change (e.g., 1, 1.5 or so)

6) Choose p-value (0.01 to 0.5)

7) Make Contrasts. (see Make Contrasts below)

8) Navigate through tab-set

Make Contrasts option is to get comparison between different conditions. Let us consider the above two target files. In **Target file 1** three experimental conditions A, B, C are present, while in **Target file 2** conditions A and B are present. In `Make Contrasts` user can choose B-A of A-B as the comparison as shown in the following screen shot. All the process is captured in Figure 4, Figure 5, Figure 6 and Figure 7.

Note: Please make sure the contrasts are present in each set of files. User can't submit a case C-A, C-B as condition C is missing in **Target file 2**. So the possible combination is A-B or B-A.

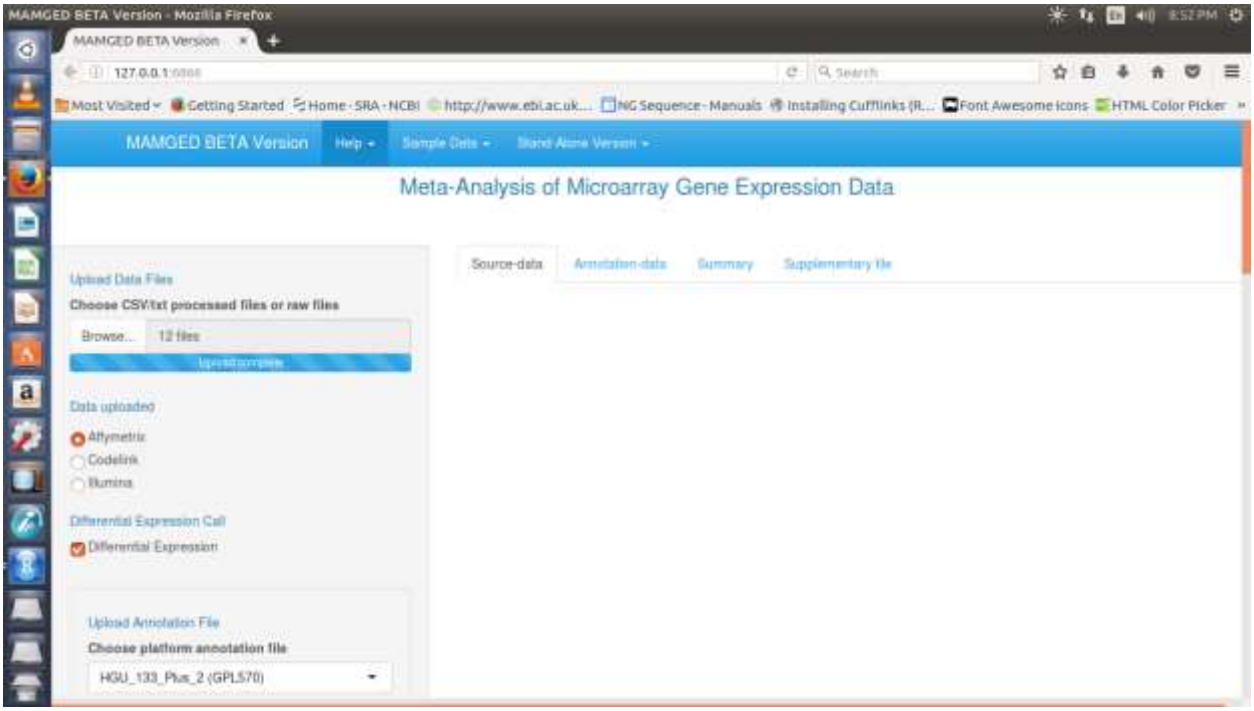

Figure 4: Parameter setting for differential expression call for Affymetrix data

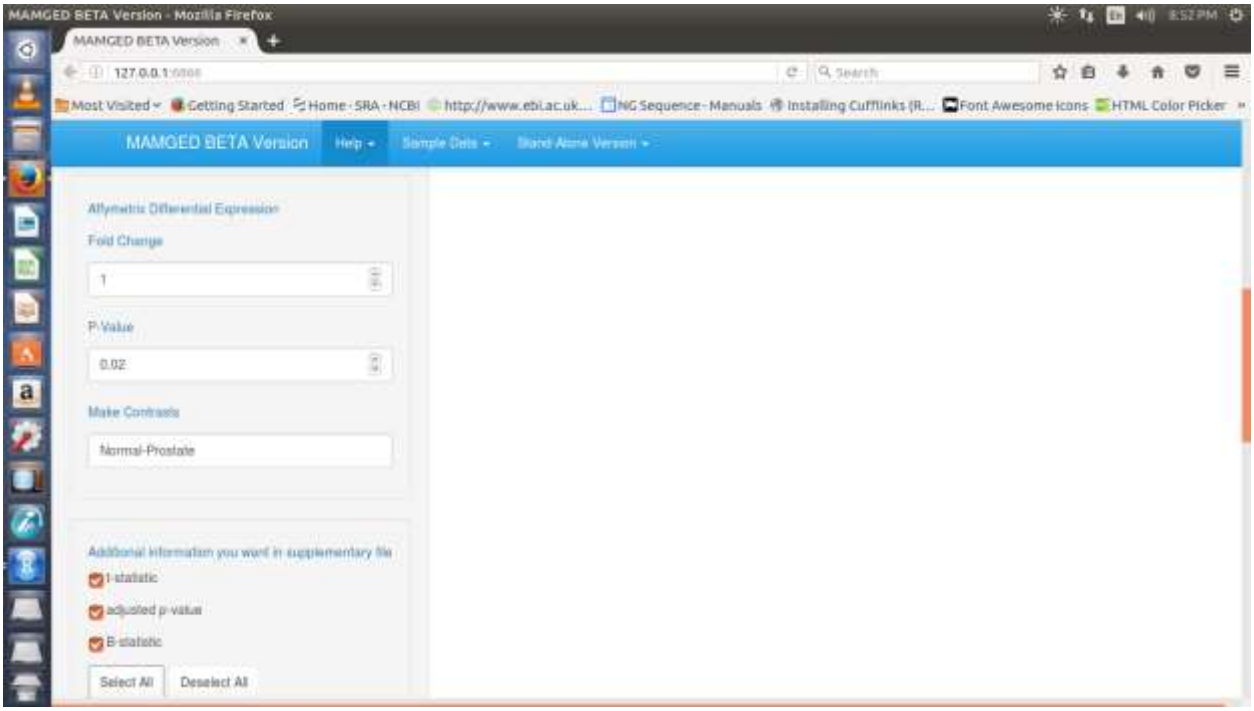

Figure5: Parameter setting for differential expression call for Affymetrix data

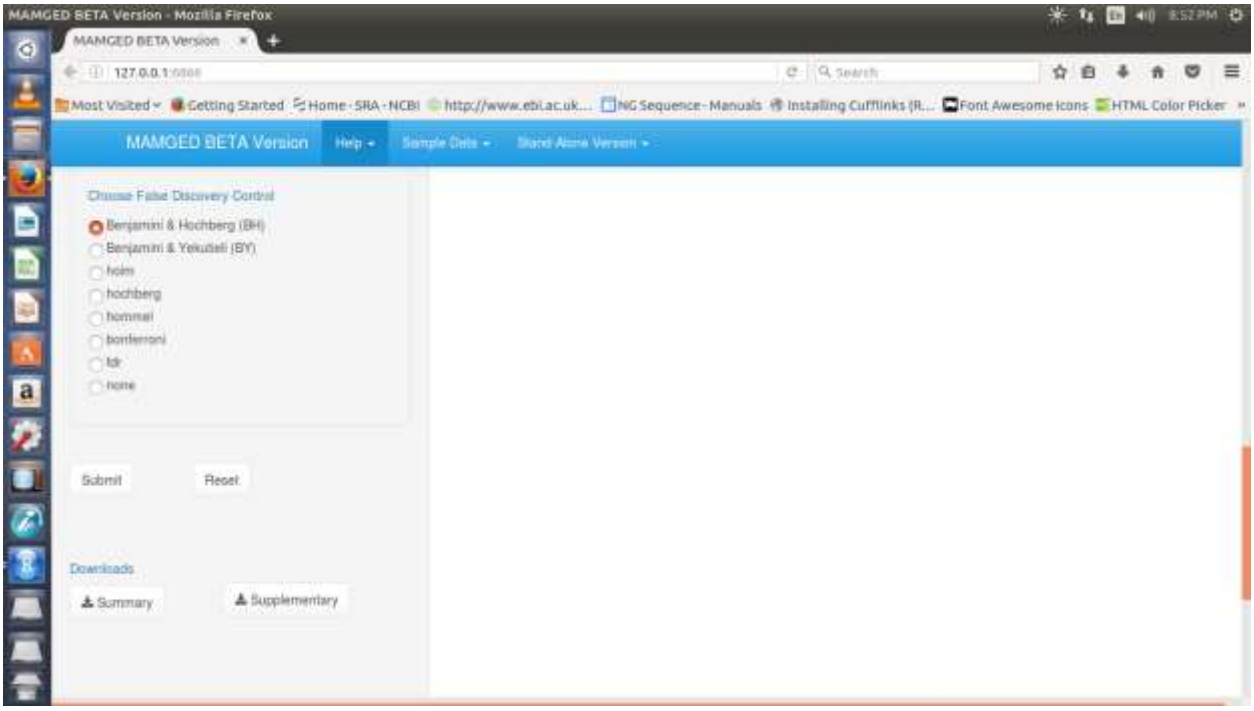

Figure6: Parameter setting for differential expression call for Affymetrix data

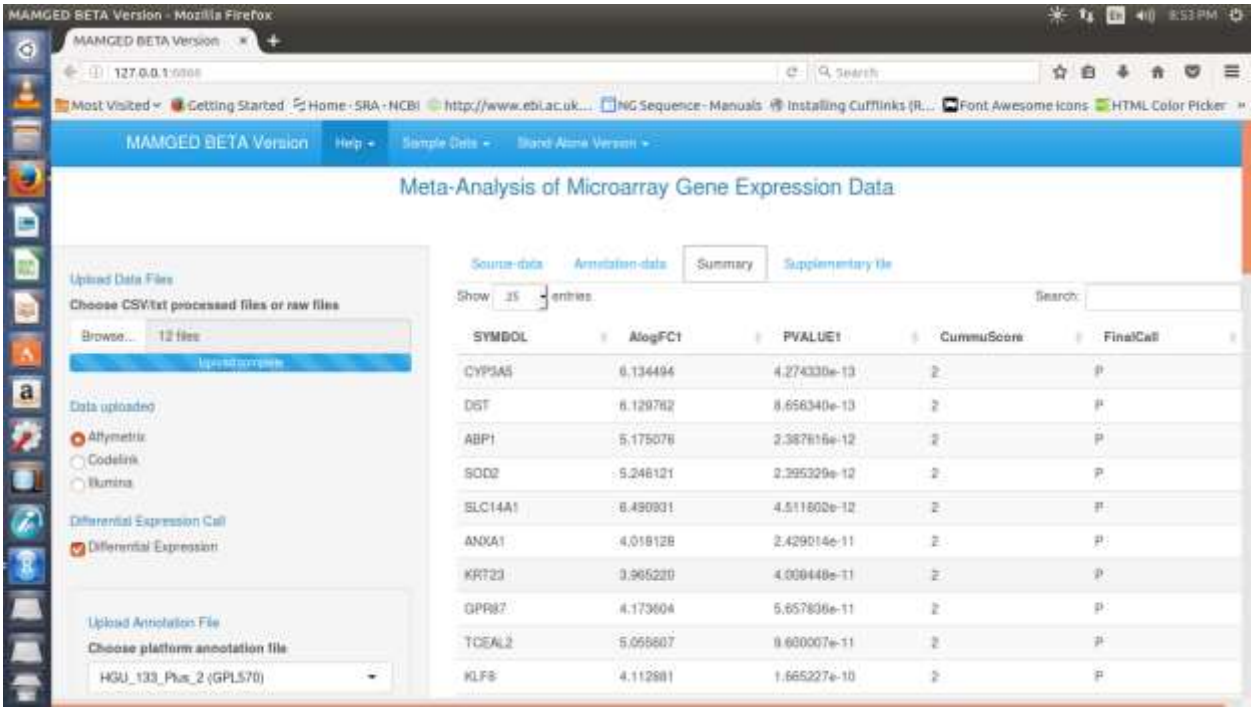

Figure 7: Differential expression summary. The colum “SYMBOL” denotes the gene symbol, “AlogFC1” is the fold change of the experimental conduction, “PVALUE1” is the p-value associated, “CummuScore” is the cumulative score and “FinalCall” is the differential expression call.

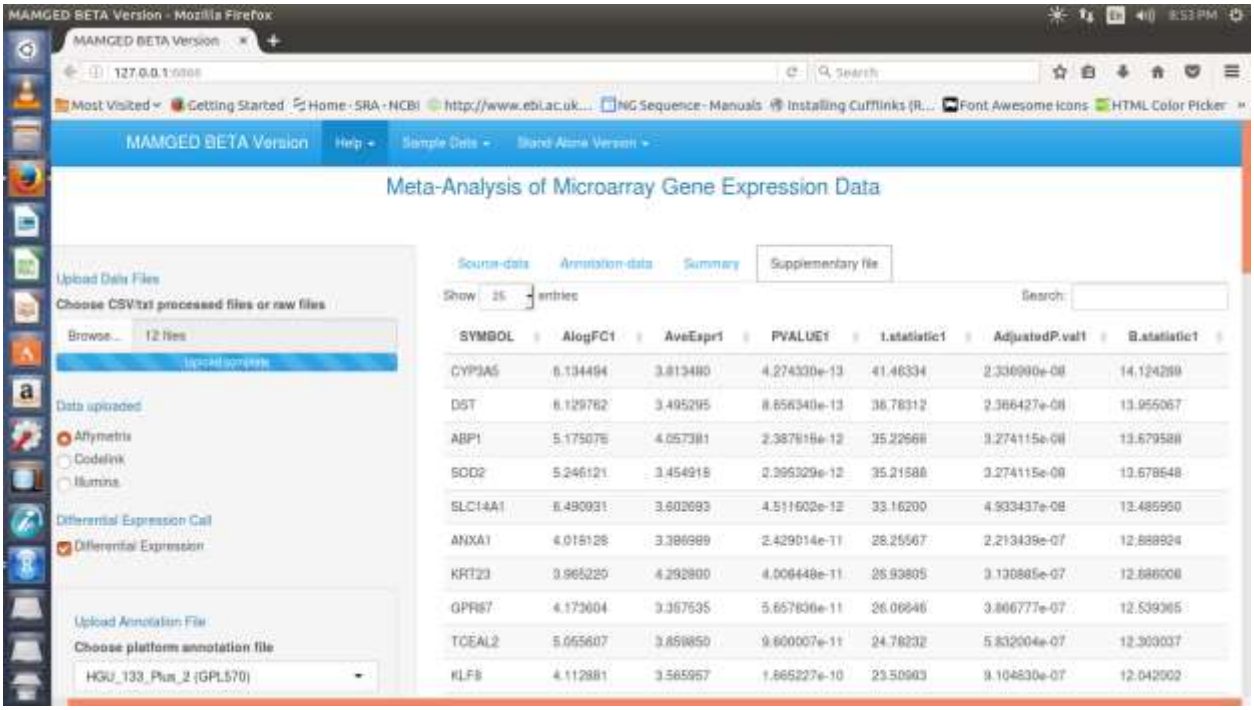

Figure 8: Shows additional information such as test statistic, adjusted p-value and B statistic for differential expression call

Section 2: Meta-analysis of Codelink data

2.1 Processed data: In order to perform meta-analysis of processed Codelink data, five steps as given in processed Affymetrix data section need to be followed. .CSV or .txt files are accepted as input. A score of +2 is given if Signal to Noise Ratio (SNR) is greater than 1 otherwise -2.

Steps to follow

1) Load processed text or CSV files (at least 2 files mandatory)

2) Choose annotation file from the database (Appropriate file recommended)

3) Choose an appropriate choice for data uploaded (here Codelink need to be checked)

4) Submit

5) Navigate through tab-set

Codelink processed sample files are displayed below. The data files must have either 'ID', Signalintensity','SNR' or 'ID' and 'SNR' information present to perform meta-analysis.

	A	B	C
1	ID	SignalIntensity	SNR
2	109	6.817373834	1.00358
3	110	11.54599254	11.9439
4	111	NA	NA
5	112	9.780420345	3.289488
6	113	9.871984222	4.017027
7	114	9.078185774	1.631052
8	116	8.228931216	1.640541
9	117	8.043536445	1.662541
10	118	6.179345832	0.848912
11	119	10.55817989	7.531226
12	120	10.47453298	7.283291
13	121	4.157294381	0.573258
14	123	NA	NA
15	124	6.045174275	0.814975
16	125	8.310021951	1.998102
17	126	7.455877066	1.247705
18	127	7.146474996	0.99608
19	128	11.42843864	10.34027
20	130	9.627379575	3.632827
21	131	9.68074173	2.55405
22	132	13.33649444	41.96874
23	133	10.43046719	6.203528

Table5: Codelink sample file with : probe set ID in first column, signal intensity in second column and p-values in third column

	A	B
1	ID	SNR
2	109	1.00358
3	110	11.9439
4	111	NA
5	112	3.289488
6	113	4.017027
7	114	1.631052
8	116	1.640541
9	117	1.662541
10	118	0.848912
11	119	7.531226
12	120	7.283291
13	121	0.573258
14	123	NA
15	124	0.814975
16	125	1.998102
17	126	1.247705
18	127	0.99608
19	128	10.34027
20	130	3.632827
21	131	2.55405
22	132	41.96874
23	133	6.203528

Table 6: Codelink sample file with probe set ID in first column, and p-values in third column

Screen shots of the above process are captured in Figure 9, Figure 10 and Figure 11.

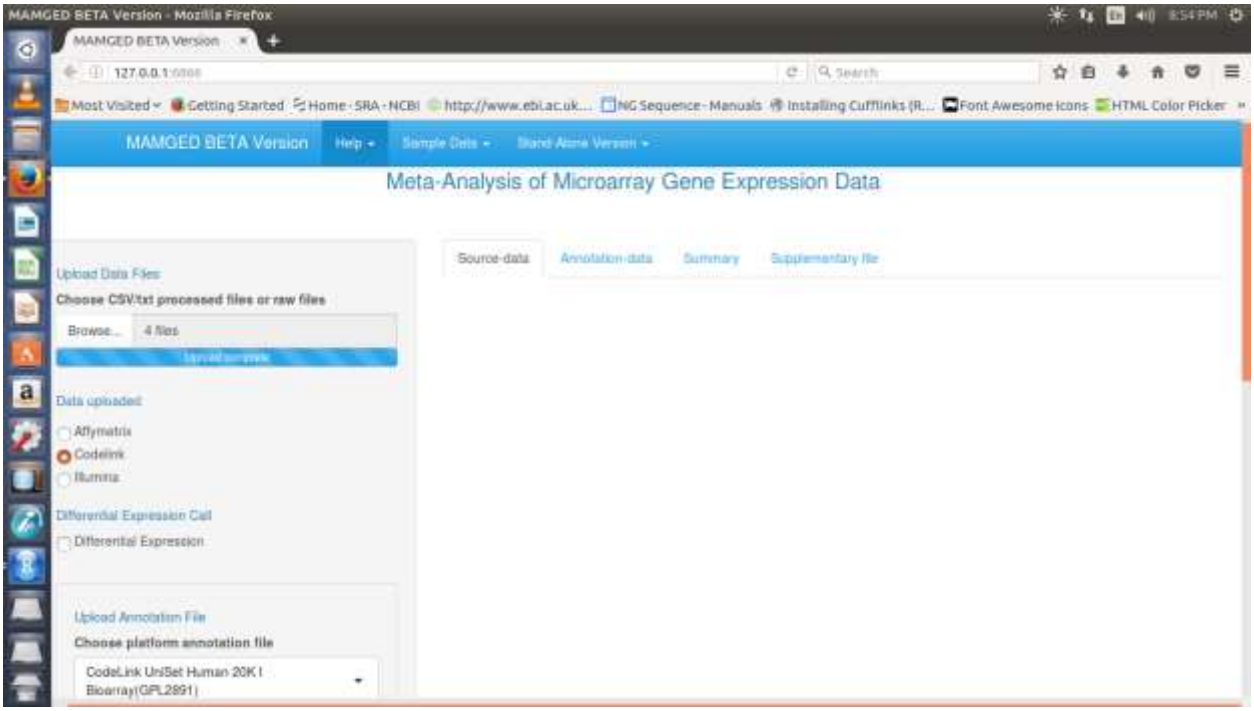

Figure 9: Parameter setting for absolute expression call for Codelink data

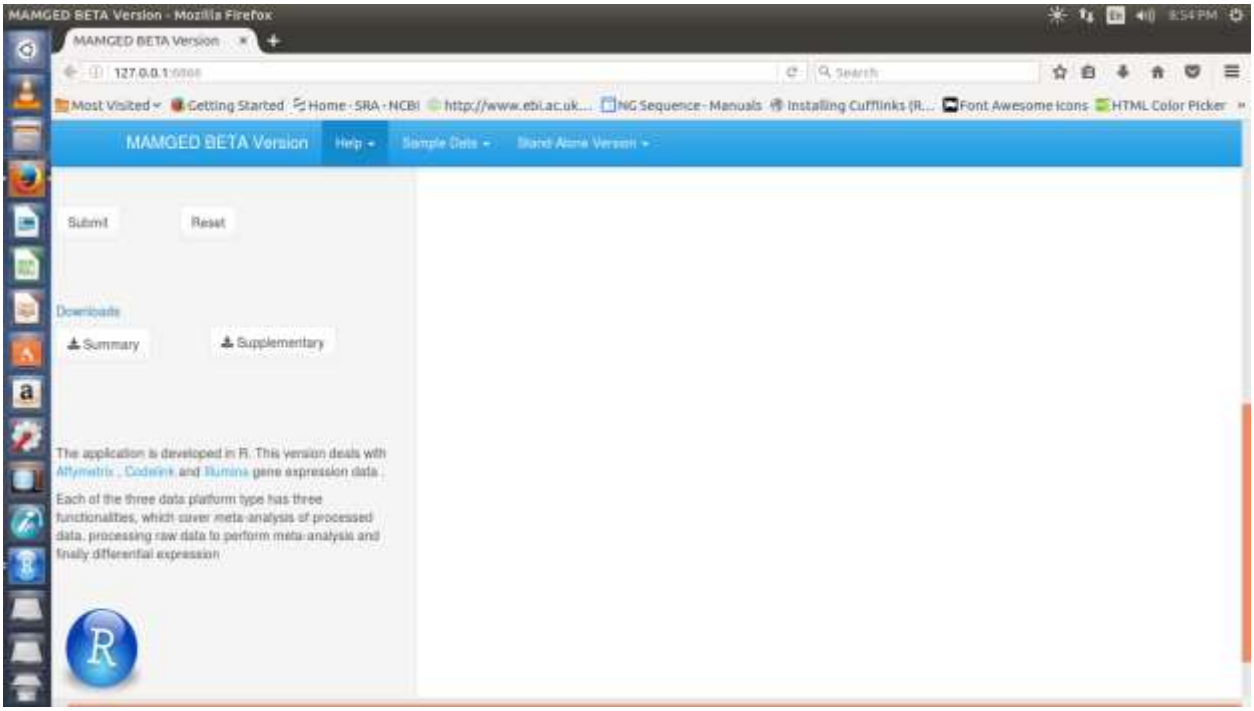

Figure 10: Parameter setting for absolute expression call for Codelink data

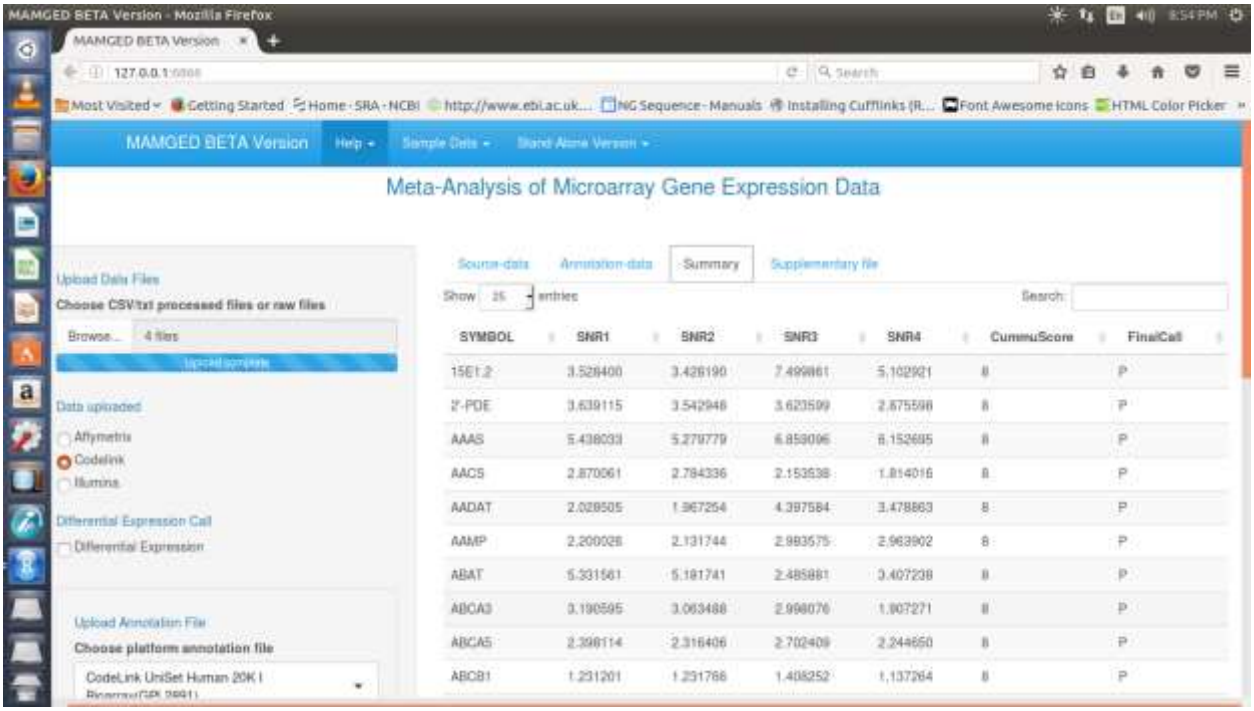

Figure 11: Shows the absolute expression call for codelink data. "SNR1", SNR2, SNR3 and SNR4 are the signal intensity values for first, second, third and fourth sample respectively.

2.2 Meta-analysis of raw data

If the user is interested to use raw data, valid .TXT files need to be supplied as input and follow the five-step process. File with missing background mean, spot mean etc is not valid Codelink file. For Codelink sample validity look at (GSE4797 VS GSE15524). The raw .TXT files are pre-processed and normalized before performing meta-analysis. Scoring method is same as for pre-processed Codelink data.

2.3 Meta-analysis of Differential expression:

To accomplish meta-analysis of Codelink differential expression, a set of files containing .TXT (raw files) and .txt (target files) need to be supplied. The procedure to prepare target file is same as that in affymetrix platform, already discussed in above section. As the Codelink data suffers from lot of inconsistencies, annotation of probes with GPL annotation file is not a good choice. To overcome this problem, Bioconductor annotation packages [5] [6] [7] are used to annotate the Codelink raw data. Please note, Bioconductor annotation packages are used with differential expression of Codelink data only.

Cumulative score is generated on the basis of log fold change. If fold change is negative score of -2 is assigned otherwise +2

Steps to follow

1) Load .TXT (data files) and .txt (target files)

2) Skip annotation file

3) Choose data uploaded (codelink)

4) Differential expression checkbox (checked)

5) Make contrasts

6) Choose annotation DB

7) Choose Fold change

8) Choose p-value

9) Submit

10) Navigate through tab-set

All the above steps are captured in Figure 12, Figure 13, Figure 14 and Figure 15.

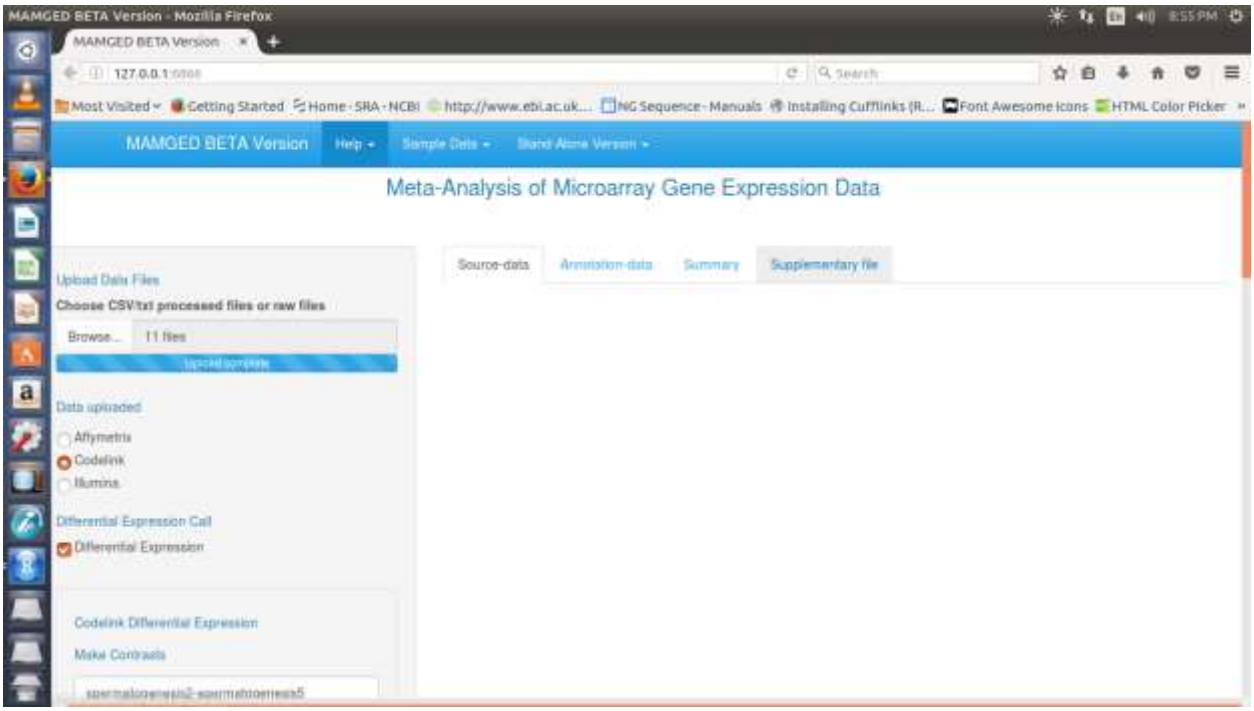

Figure 12: Parameters setting for differential expression call for Codelink data

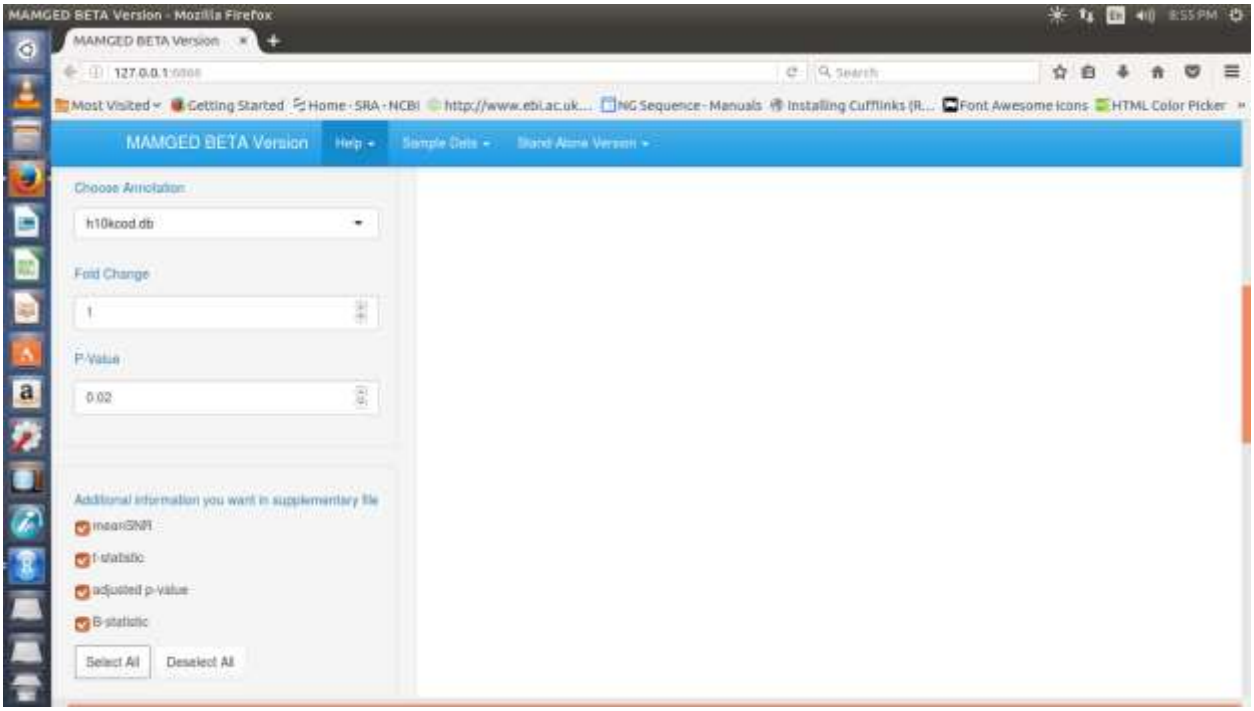

Figure 13: Parameters setting for differential expression call for Codelink data

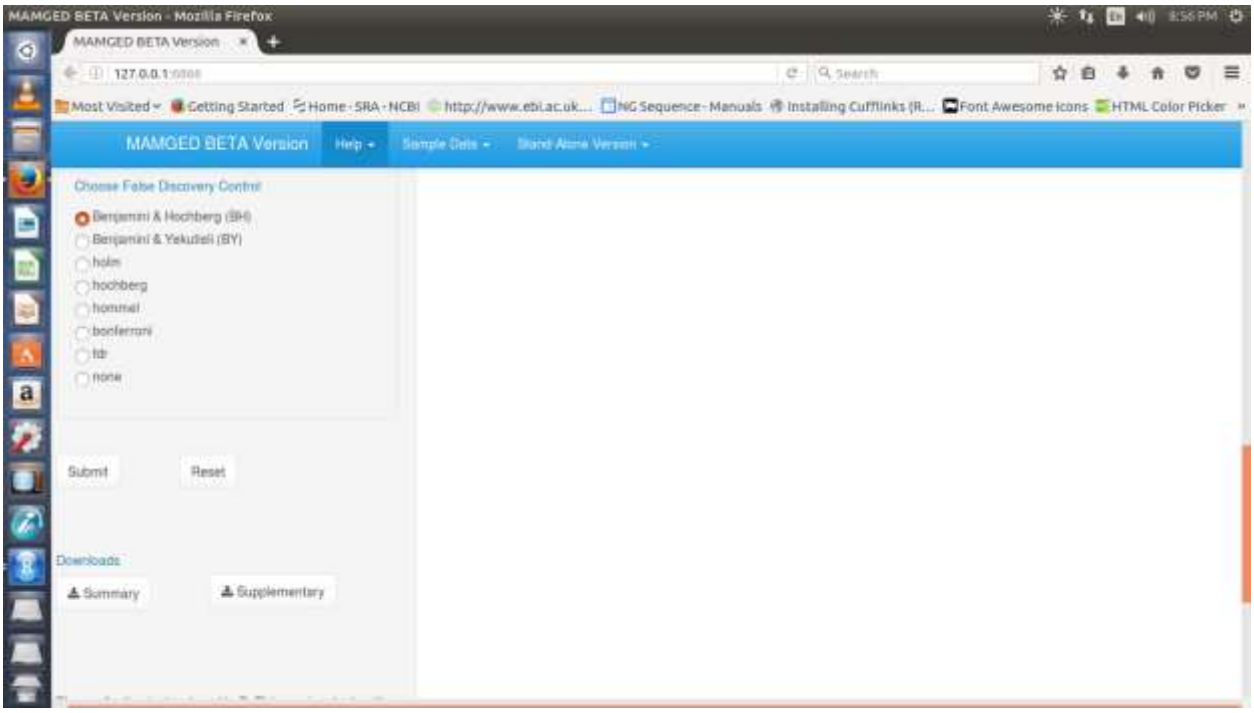

Figure 14: Parameters setting for differential expression call for Codelink data

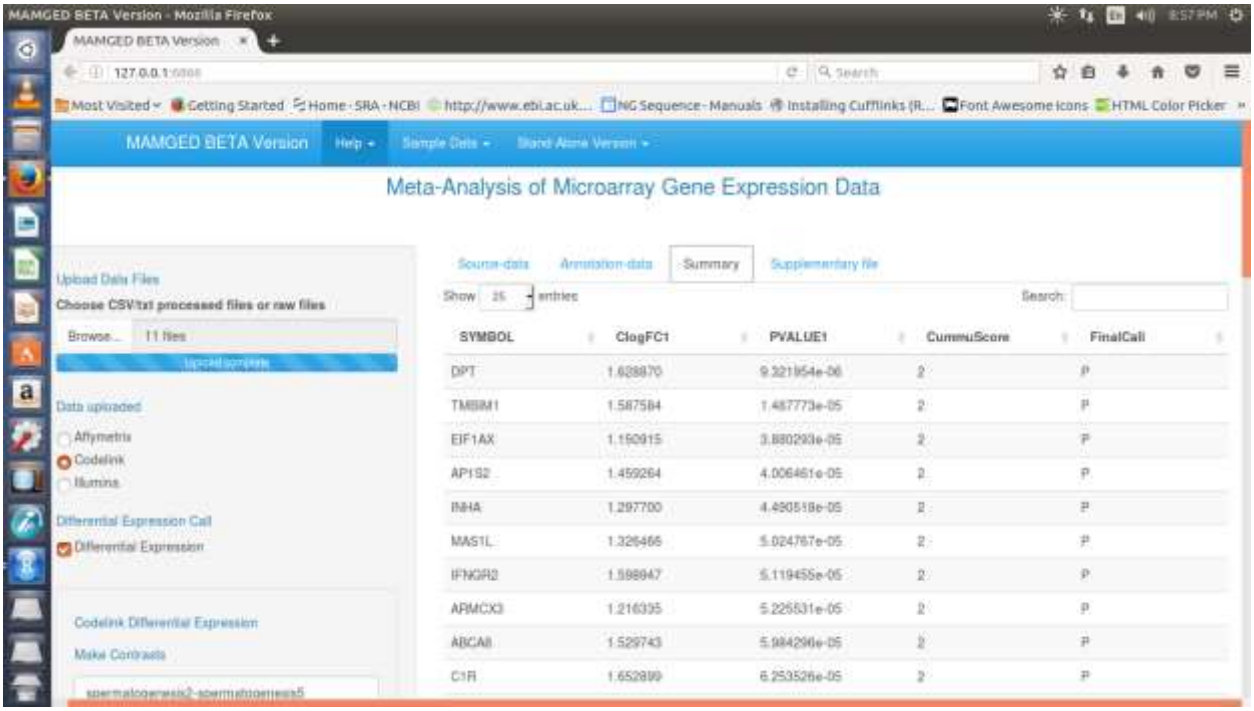

Figure 15: Differential expression call summary for Codelink data

Section 3:

3.1 Meta-analysis of processed Illumina data:

Illumina text or CSV files are accepted as input to perform meta-analysis. Like Affymetrix and Codelink discussed above, five step procedure need to be followed. Scoring is done on the basis of p-value either Present (2), Absent (-2) or Marginal (0). Processed Illumina sample files are displayed below. Screen shots captured in Figure 16, Figure 17 and Figure 18.

	A	B	C
1	ID	Intensity	P-value
2	ILMN_1681101	6.966361	0.27403
3	ILMN_2094942	7.003377	0.18961
4	ILMN_1703142	7.60047	0
5	ILMN_2271336	6.93546	0.37662
6	ILMN_2337789	7.064877	0.08312
7	ILMN_1669592	7.245969	0.00519
8	ILMN_1735038	7.089583	0.05325
9	ILMN_1699644	7.092096	0.05065
10	ILMN_1655796	7.151017	0.01558
11	ILMN_1789991	6.910043	0.45974
12	ILMN_2047430	7.735894	0
13	ILMN_1702764	7.274611	0
14	ILMN_1757106	8.700071	0
15	ILMN_1717337	10.04946	0
16	ILMN_1677318	7.07106	0.07792
17	ILMN_2257749	6.713215	0.97273
18	ILMN_2341626	7.044121	0.1026
19	ILMN_2336335	6.959712	0.29221
20	ILMN_1812607	6.779719	0.87143
21	ILMN_1748393	6.90147	0.47403
22	ILMN_3305472	6.827501	0.73247
23	ILMN_3228863	6.857359	0.62727
24	ILMN_3234892	7.168672	0.01299
25	ILMN_1671854	7.045214	0.1

Illumina sample file 1

	A	B
1	ID	PVALUE
2	ILMN_1681101	0.27403
3	ILMN_2094942	0.18961
4	ILMN_1703142	0
5	ILMN_2271336	0.37662
6	ILMN_2337789	0.08312
7	ILMN_1669592	0.00519
8	ILMN_1735038	0.05325
9	ILMN_1699644	0.05065
10	ILMN_1655796	0.01558
11	ILMN_1789991	0.45974
12	ILMN_2047430	0
13	ILMN_1702764	0
14	ILMN_1757106	0
15	ILMN_1717337	0
16	ILMN_1677318	0.07792
17	ILMN_2257749	0.97273
18	ILMN_2341626	0.1026
19	ILMN_2336335	0.29221
20	ILMN_1812607	0.87143
21	ILMN_1748393	0.47403
22	ILMN_3305472	0.73247
23	ILMN_3228863	0.62727
24	ILMN_3234892	0.01299
25	ILMN_1671854	0.1

Illumina sample file 2

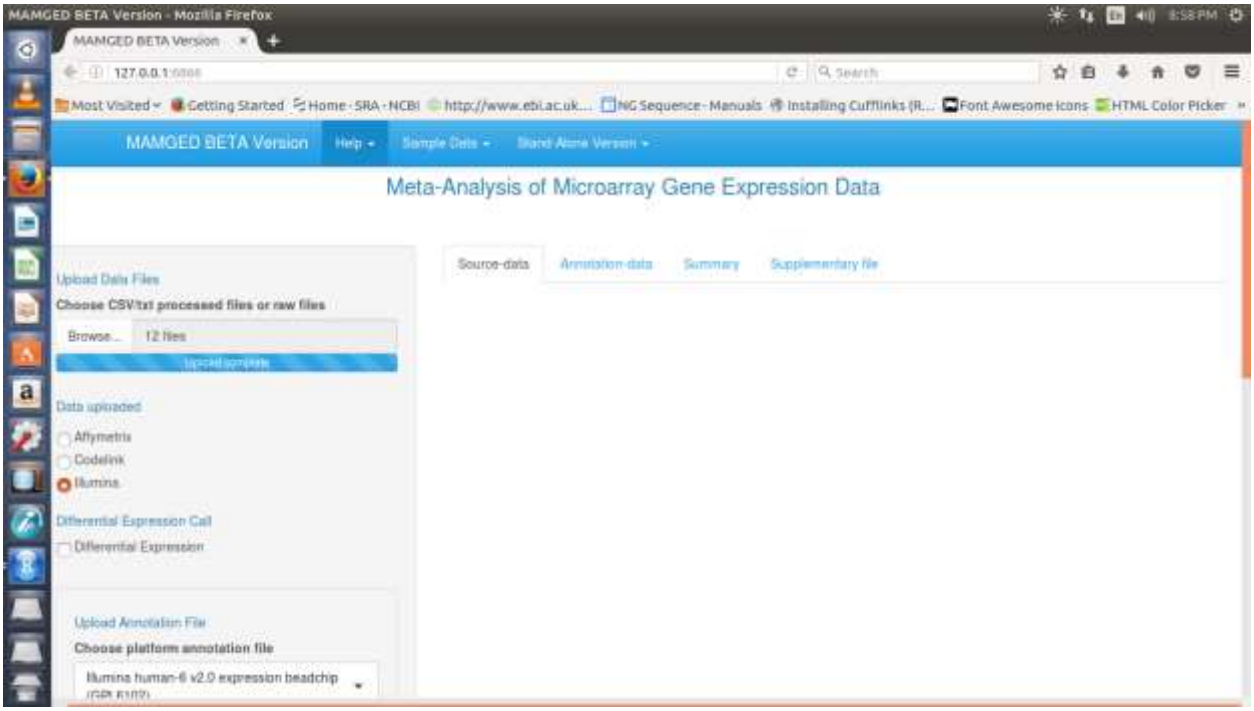

Figure 16: Parameter selection for absolute expression call for Illumina data

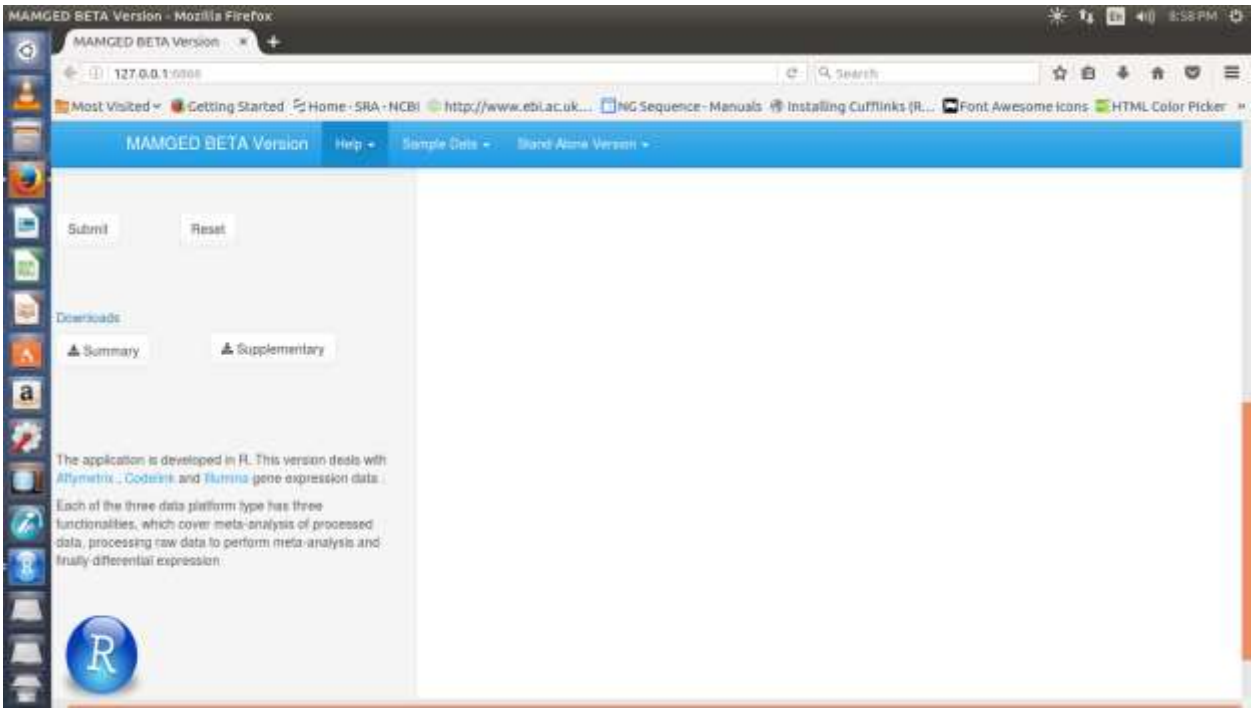

Figure 17: Parameter selection for absolute expression call for Illumina data

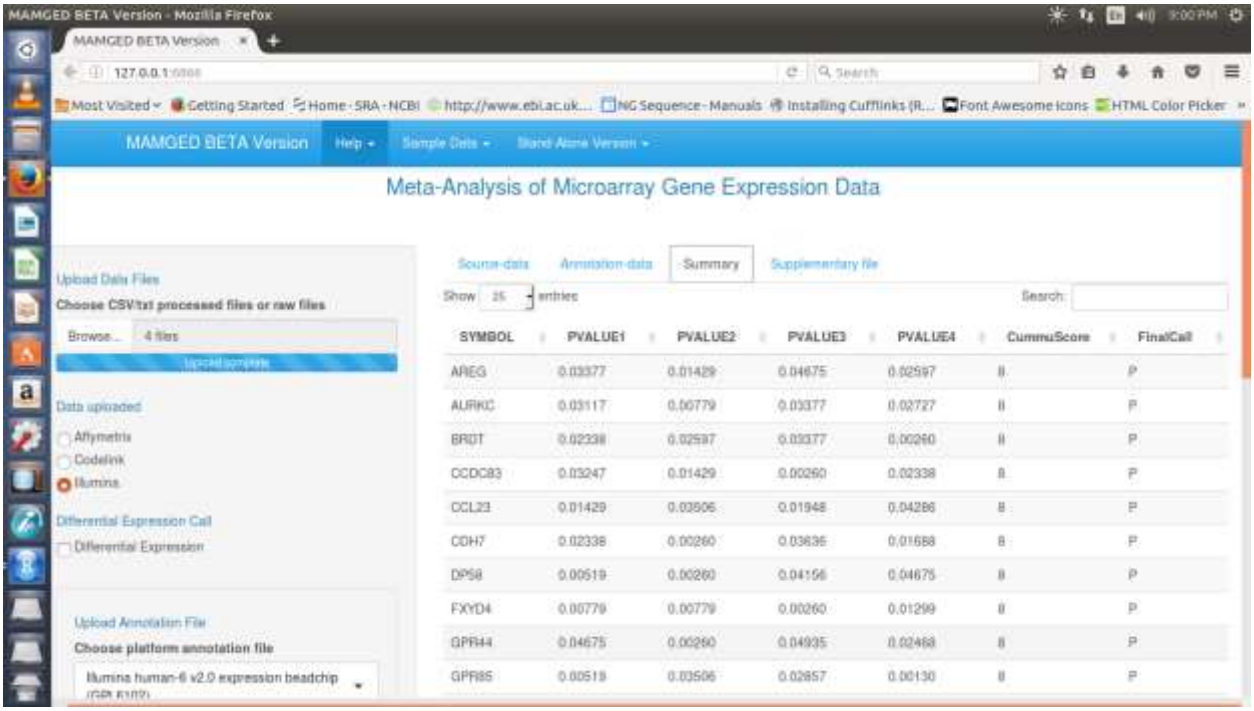

Figure 18: Absolute expression call summary for Illumina data

3.2 Meta-analysis of raw data:

Illumina arrays processed together are written on same file unlike Affymetrix and Codelink platform. Individual raw text (.txt) or many raw data files written on same text file can be given as input. More than two individual files are accepted. If the raw data files are written in same text file, then the file should contain the information of at least two files. To perform meta-analysis of Illumin raw data, same procedure need to be followed as discussed for Affymetrix raw data.

File with missing Average Signal, Average Normalization Bead information is not a valid Illumina file.

3.3 **Meta-analysis of Differential Expression:** Data and target files need to be in text (.txt) format. A sample target file is displayed below. The file has two columns, first referring to files and the second belong to experimental conditions.

Filenames	Targets
File1	C
File2	TT
File3	D
File4	DT
File5	C
File6	D
File7	TT
File8	DT
File9	C
File10	D
File11	TT
File12	DT

Illumina Target file sample

Steps to follow

1) Load .txt data and target files (more than 2 recommended)

2) Choose annotation file

3) Choose data uploaded (out of three options, here Illumina)

4) Check box of differential expression need to be checked

5) Choose fold change (e.g., 1, 1.5 or so)

6) Choose p-value (0.01 to 0.5)

7) Make Contrasts.

8) Navigate through tab-set.

References

[1] K. K. Acharya, D. S. Chandrashekar, N. Chitturi, H. Shah, V. Malhotra, K. S. Sreelakshmi, H. Deepti, A. Bajpai, S. Davuluri, P. Bora, and L. Rao, "A novel tissue-specific meta-analysis approach for gene expression predictions, initiated with a mammalian gene expression testis database.," *BMC Genomics*, vol. 11, p. 467, Jan. 2010.

[2] A. C. J. Miller, "Miller CJ. simpleaffy: Very simple high level analysis of Affymetrix data.R packag

eversion2.44.0,http://www.bioconductor.org,http://bioinformatics.picr.man.ac.uk/simpleaffy/.," 2015.

[3] G. K. Smyth, M. Ritchie, N. Thorne, and J. Wettenhall, "Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W and Smyth GK (2015). 'limma powers differential expression analyses for RNA-sequencing and microarray studies.' Nucleic Acids Research, 43(7), pp. e47.," no. April, 2010.

[4] D. Diez, "Diego Diez, Rebeca Alvarez and Ana Dopazo (2007). 'codelink: an R package for analysis of GE Healthcare Gene Expression Bioarrays.' Bioinformatics, 23(9), pp. 1168-9.http://dx.doi.org/10.1093/bioinformatics/btm072, http://dx.doi.org/10.1093/bioinformatic," no. 1, pp. 1–9, 2015.

[5] A. M. Manufacturer, A. N. Description, D. For, E. Gene, E. Gene, and D. With, "Diez D. h20kcod.db: Codelink UniSet Human 20k I Bioarray annotation data (chip h20kcod). R package version 3.2.0.," no. xx, 2015.

[6] A. M. Manufacturer, A. N. Description, D. For, E. Gene, E. Gene, and D. With, "Diez D. h10kcod.db: Codelink UniSet Human I Bioarray (~10 000 human genes) annotation data (chip h10kcod). R package version 3.2.0.," no. xx, 2015.

[7] A. M. Manufacturer, A. N. Description, D. For, E. Gene, E. Gene, and D. With, "Diez D. hwgcod.db: Codelink Human Whole Genome Bioarray (~55 000 human genes) annotation data (chip hwgcod). R package version 3.2.0.," no. xx, 2015.