Using BiSEp to nominate candidate Synthetic Lethal gene pairs

Mark Wappett

March 3, 2015

1 Introduction

BiSEp (Bimodality subsetting expression) is a set of computational tools that enable the user to nominate candidate synthetic lethal (SL) gene pairs. The toolkit is based on the assumption that a clear on / off gene expression profile is indicitive of tumour loss, and is detectable as bimodality or non-normality.

This vignette demonstrates how the toolkit can be used to nominate, assess and visualise candidate SL pairs nominated from gene expression and mutation datasets.

2 Importing gene expression data

Processed gene expression data from most platforms can be input. We recommend that values are all distributed above zero and are in the log2 scale. Example input data format is shown below:

```
> require(BiSEp)
> data(INPUT_data)
> INPUT_data[1:2,1:6]
```

```
        42MGBA     5637     639V     647V      769P    8305C
MICB 0.3340947 4.469222 3.877129 5.504680 0.2911058 2.806195
YAP1 4.2810073 4.213072 2.743619 3.611307 2.9417470 3.386272
```

All input data should be read in a gene by sample format. Our dataset is gene expression data from the Cancer Cell Line Encyclopedia (CCLE) [1], normalised using fRMA [2] and scaled.

3 Identifying bimodal genes in expression data

We next take the processed data matrix and run the bimodal detection tool across it. This generates a list object containing three matrices, the third of which is called DATA and is simply a capture of the input data matrix.

```
> BISEP_data <- BISEP(INPUT_data)
> biIndex <- BISEP_data$BI
> bisepIndex <- BISEP_data$BISEP
```

The output data frames called biIndex and bisepIndex are the output from the bimodal index function [3] and the novel BISEP function. The output is displayed below:

```
> biIndex[1:10,]
```

```
             mu1      mu2     sigma    delta         pi        BI
MICB   0.9338792 3.870837 0.7634495 3.846958 0.28285275 1.7326138
YAP1   0.3069453 3.189808 0.6000751 4.804169 0.11022879 1.5045444
BOK    1.3373621 3.743565 0.6750419 3.564524 0.18984581 1.3979316
MTAP   0.5245744 3.470745 0.6534969 4.508315 0.18604307 1.7543722
EPHB2  2.1878062 4.164659 0.7103492 2.782931 0.34627140 1.3240656
BRCA2  3.2406223 4.316125 0.7019131 1.532245 0.76532428 0.6493590
TUSC3  0.5724434 3.854679 0.6501654 5.048309 0.25574133 2.2024615
PHLDA3 1.9049445 4.051928 0.8062587 2.662897 0.37102715 1.2863916
MLH1   0.5957125 5.146278 0.4644327 9.798116 0.03846154 1.8842532
BRCA1  3.2844328 4.989221 0.5501418 3.098817 0.06754281 0.7776782
```

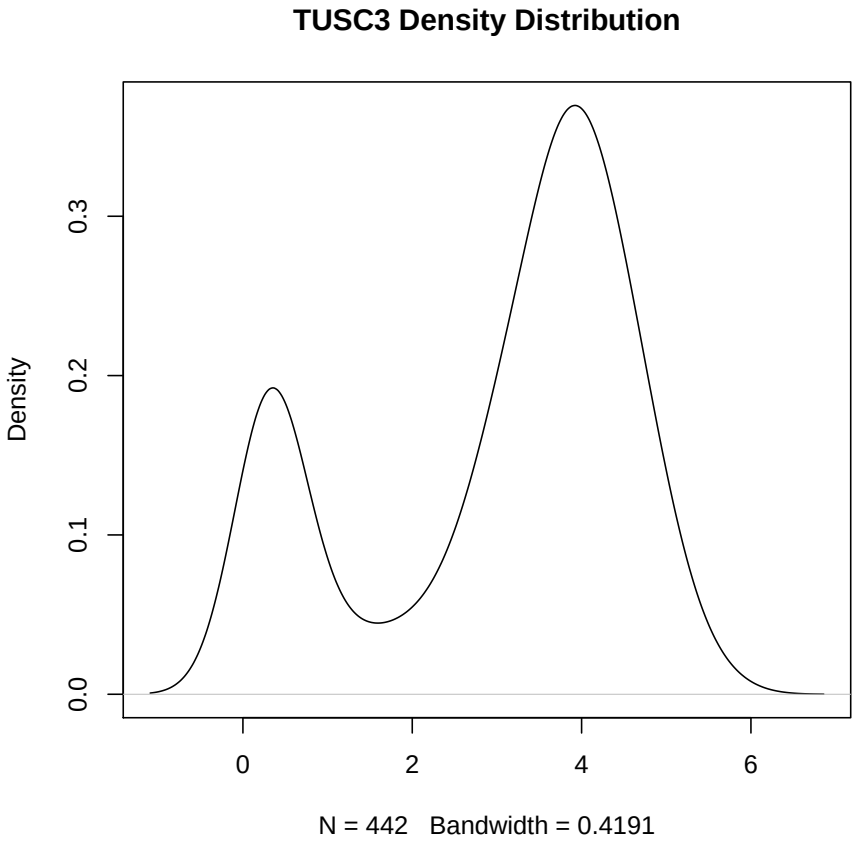
```
> bisepIndex[1:10,]
```

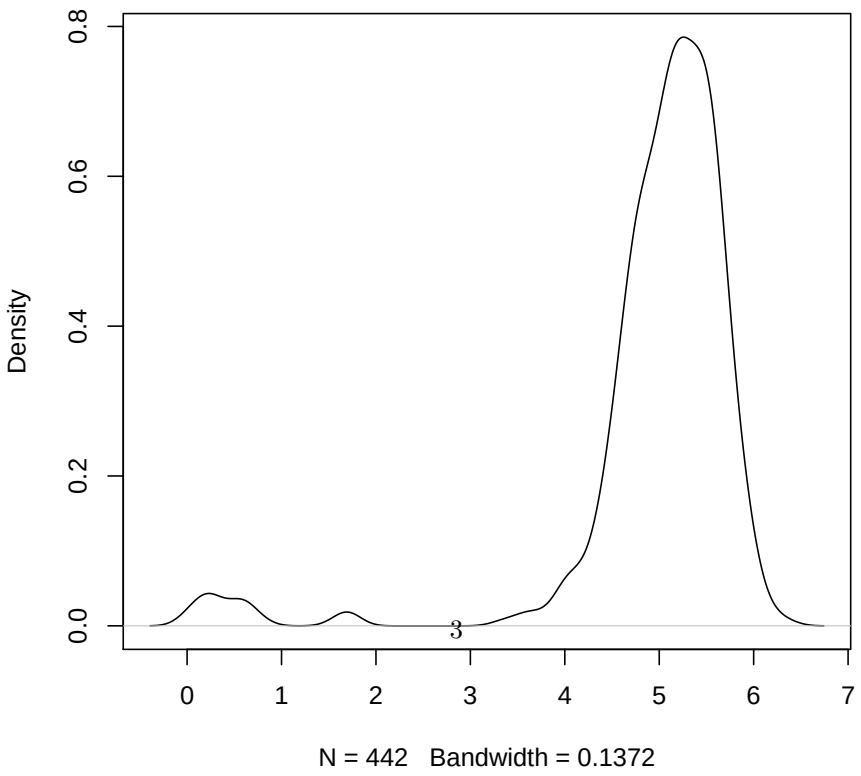
```
             V1         V2
MICB   1.448461 0.44843049
YAP1   1.052045 0.27472527
BOK    1.982416 0.64516129
MTAP   0.966708 0.14124294
EPHB2  2.173364 0.68965517
BRCA2  5.170457 0.95238095
TUSC3  2.363063 0.23752969
PHLDA3 5.379839 1.00000000
MLH1   2.543707 0.07092199
BRCA1  3.609235 0.71942446
```

The biIndex matrix contains all the bimodal scoring information provided to us by the bimodal index function. This includes the delta (distance between two expression modes), pi (proportion of samples in each expression mode) and BI. When combined, these give us an optimal assessment of bimodality in expression data. The bisepIndex function provides a p-value score for non-normality (column 2) and accurately pin-points the mid-point between the two expression modes for a gene (column 1).

TUSC3 scores the highest in the biIndex table - the density distribution below highlights this:

```
> plot(density(INPUT_data["TUSC3",]), main="TUSC3 Density Distribution")
```


TUSC3 Density Distribution
Density
0.0
0.1
0.2
0.3
0
2
4
6
N = 442 Bandwidth = 0.4191


MLH1 Density Distribution
Density
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
6
7
N = 442 Bandwidth = 0.1372

```
> plot(density(INPUT_data["MLH1",]), main="MLH1 Density Distribution")
```

By comparison, MLH1 does not score high for bimodality - but has the lowest p value for non-normality. The density plot demonstrates the unbalanced nature of, and distance between the two populations in a typical non-normal distribution.

4 BIGEE: Bimodality in Gene Expression Exclusivity

Here we take the bimodal / non-normal output from the BISEP tool, and use it as input to the first of the two candidate synthetic lethal detection tools. There are four sample input options to this tool based on the sample type and sample numbers **cell line, cell line low, patient and patient low**. When sample numbers are below 200 we recommend using the input parameters with the low suffix in order to prevent a high false positive rate.

```
> BIGEE_out <- BIGEE(BISEP_data, sampleType="cell_line")
```

```
[1] "Selected CELL LINE sample type"
[1] "Subsetting bimodal index"
[1] "Filtering"
[1] "Setting up synthetic lethal detection"
[1] "Running SL detection"
[1] "Number of bimodal genes: 12"

  |
  |                                                                      |   0%
  |
  |======                                                                |   8%
  |
  |============                                                          |  17%
  |
  |==================                                                    |  25%
  |
  |========================                                              |  33%
  |
  |==============================                                        |  42%
  |
  |====================================                                  |  50%
  |
  |==========================================                            |  58%
  |
  |================================================                      |  67%
  |
  |======================================================                |  75%
  |
  |============================================================          |  83%
  |
```

```
  |=================================================================      |  92%
  |
  |=======================================================================| 100%
[1] "Summarising..."
```

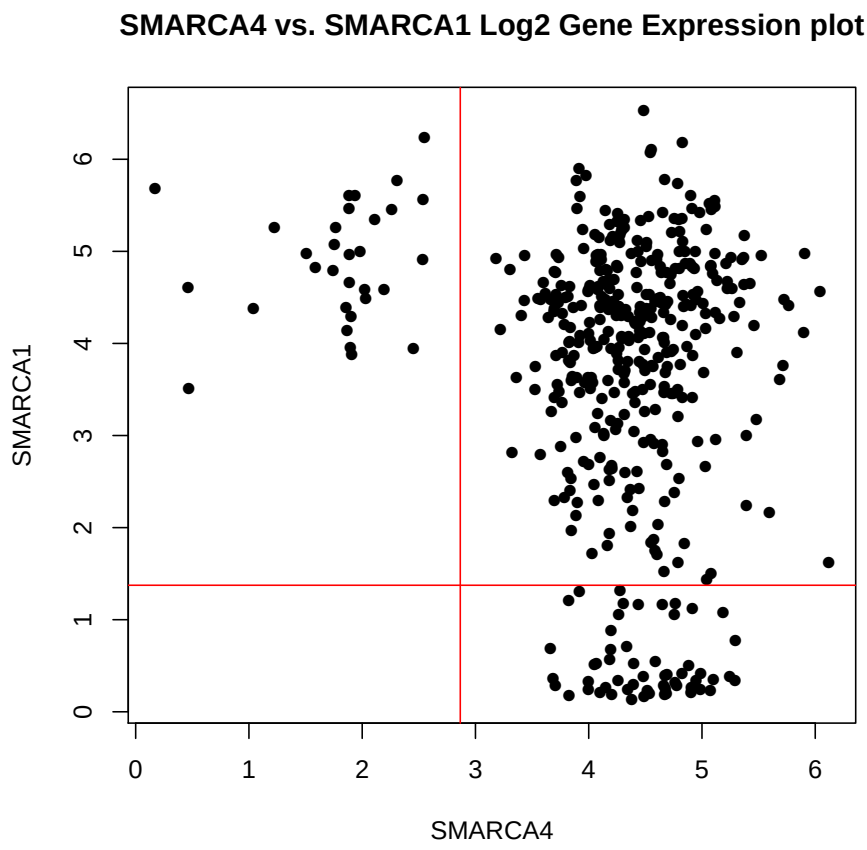
The percent completion graphic displays the progress of the SL detection component of the tool. This will typically take longer the larger the dataset is, and the more bimodal genes that there are. The output from this tool is a matrix containing gene pairs that look potentially synthetic lethal in the dataset, along with a score.

```
> BIGEE_out[1:4,]
```

```
    gene2    gene     score
2    MTAP    MLH1 11.686754
1    MLH1     BOK 10.013253
9    YAP1    MLH1  7.357821
7 SMARCA4 SMARCA1  6.822608
```

It is possible to visualise any candidate relationships using the expressionPlot function:

```
> expressionPlot(BISEP_data, gene1="SMARCA4", gene2="SMARCA1")
```



and look for those gene pairings that ideally are never expressed at low levels together - the signature that we propose could be indicative of synthetic lethality.

```
> expressionPlot(BISEP_data, gene1="MTAP", gene2="MLH1")
```

5 BEEM: Bimodal Expression Exclusive with Mutation

Here we take the bimodal / non-normal output from the BISEP tool, and use it as input to a tool that detects mutual exclusive loss between bimodally expressed genes and mutated genes. Again, there are four sample input options to this tool based on the sample type and sample numbers **cell line, cell line low, patient and patient low**. Additionally we also require a second input matrix containing discreet mutation call information. This matrix must be in the rownames = genes, colnames = samples format and there must be overlap between sample names in this mutation matrix, and sample names in the INPUT data matrix seen earlier. The calls in this matrix must be either WT or MUT as shown below:

```
> data(MUT_data)
> MUT_data[1:4,1:10]
```

```
        5637 42MGBA 639V 647V 769P 8305C 8505C 8MGBA A101D A2058
BRCA2     WT     WT   WT   WT   WT    WT    WT    WT    WT    WT
PBRM1     WT     WT   WT  MUT   WT    WT    WT   MUT    WT    WT
SCN2A     WT    MUT  MUT   WT   WT    WT    WT    WT    WT    WT
CACNA1D   WT     WT   WT  MUT   WT    WT    WT    WT    WT    WT
```

Now we can run the function by doing the following:

```
> BEEMout <- BEEM(BISEP_data, mutData=MUT_data, sampleType="cell_line",  minMut=40)
```

```
[1] "Minimum number of mutations considered for each gene is: 40"
[1] "Selected CELL LINE sample type"
[1] "Number of bimodal expression genes : 12"
[1] "Number of mutation genes wih frequency greater than 40 : 4"

  |
  |                                                                      |   0%
  |
  |======                                                                |   8%
  |
  |============                                                          |  17%
  |
  |==================                                                    |  25%
  |
  |========================                                              |  33%
  |
  |==============================                                        |  42%
  |
  |====================================                                  |  50%
  |
  |==========================================                            |  58%
  |
  |================================================                      |  67%
  |
  |====================================================                  |  75%
  |
  |============================================================          |  83%
  |
  |======================================================================|  92%
  |
  |======================================================================| 100%
[1] "Summarising..."
```

As with the BIGEE tool, the percent completion graphic displays the progress of the SL detection component of the tool. The output from the tool is a matrix containing the gene pairs that look potentially synthetic lethal, along with a number of other columns of metadata including size of high and low expression population, numbers of those populations that are mutant.

```
> BEEMout
```

	Gene1	Gene2	LowerExpressionMutationCount	HighExpressionMutationCount
2	MICB	PBRM1	0	42
9	BOK	BRCA2	0	55
20	EPHB2	CACNA1D	0	41
7	YAP1	SCN2A	0	59
4	MICB	CACNA1D	3	38
12	BOK	CACNA1D	2	39
5	YAP1	BRCA2	2	53
35	BRCA1	SCN2A	0	59
13	MTAP	BRCA2	5	50
18	EPHB2	PBRM1	3	39

	Fishers P Value	Percentage of lower samples mutated
2	0.0002127793	0
9	0.0002552670	0
20	0.0018895294	0
7	0.0019129823	0
4	0.0576643837	3.7037037037037
12	0.0973332166	3.17460317460317
5	0.0975156120	4.44444444444444
35	0.1490033473	0
13	0.1701101513	6.94444444444444
18	0.2438573394	4.76190476190476

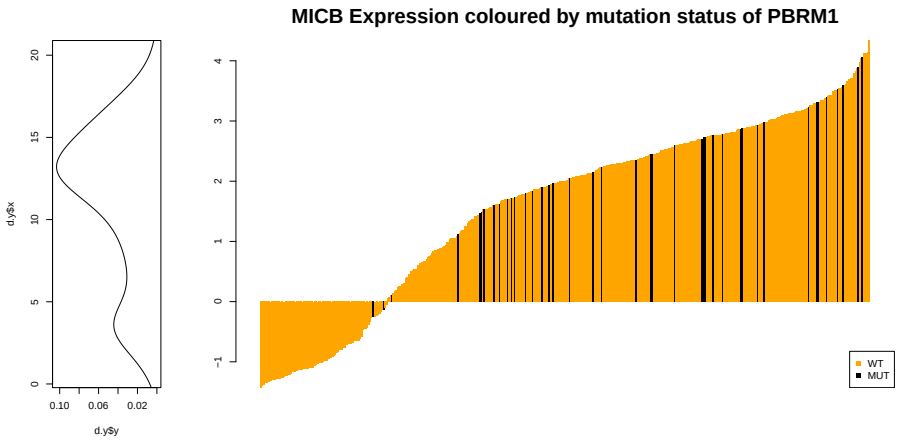
	Percenage of high samples mutated	Size of low expression population
2	11.6343490304709	81
9	14.5118733509235	63
20	10.8179419525066	63
7	14.8614609571788	45
4	10.5263157894737	81
12	10.2902374670185	63
5	13.3501259445844	45
35	13.9150943396226	18
13	13.5135135135135	72
18	10.2902374670185	63

	Size of high expression population	Enrichment Status
2	361	HighEnriched
9	379	HighEnriched
20	379	HighEnriched
7	397	HighEnriched
4	361	HighEnriched
12	379	HighEnriched
5	397	HighEnriched
35	424	HighEnriched
13	370	HighEnriched
18	379	HighEnriched

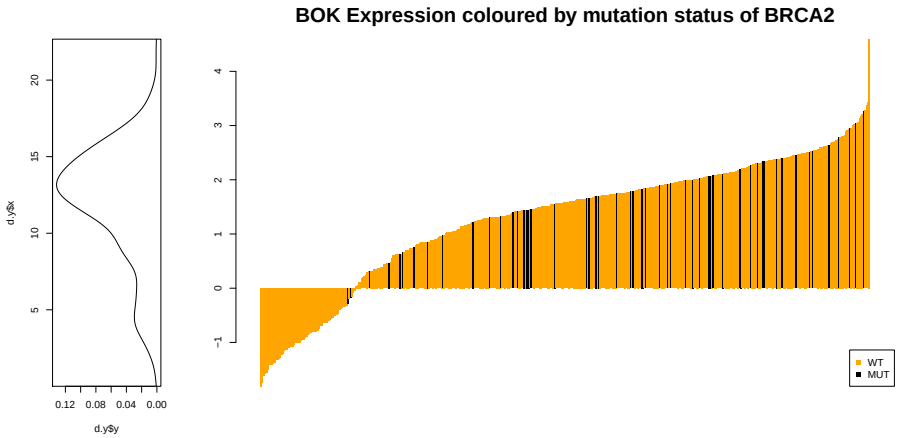
Gene pairs where the mutant gene2 is exclusively mutated, or significantly enriched for mutation in the high expression mode of expression gene1 are those that we propose as candidate SL pairs. It is another manifestation of the never-low-together relationship we were looking for in the expression data above.

We can visualise these gene pairs using the waterfall plotting function built into the package

```
> waterfallPlot(BISEP_data, MUT_data, expressionGene="MICB",
+ mutationGene="PBRM1")
```




```
> waterfallPlot(BISEP_data, MUT_data, expressionGene="BOK",
+ mutationGene="BRCA2")
```



The left panel is the density distribution of the bimodal / non-normal expression gene. The right hand panel is a bimodal-mid-point-centered barplot coloured by the mutation status of the mutation gene.

6 FURE: Functional redundancy between synthetic lethal genes

It is assumed that either gene in a synthetic lethal pair is able to functionally compensate for the loss of the other. We developed this tool to enable the user

to prioritise gene pairs that have some sort of biological redundancy and score these according to gene ontology[4,5].

The tool takes as input either the output from the BIGEE or the BEEM tools. The following example is run on the first couple of results from the BIGEE output

```
> fOut <- FURE(BIGEE_out[1,], inputType="BIGEE")
> frPairs <- fOut$funcRedundantPairs
> allPairs <- fOut$allPairs

> allPairs[1,]

  gene2 gene    score           redundant_ids   redundant_terms
2  MTAP MLH1 11.68675 GO:0005634 / GO:0005634 nucleus / nucleus
  MolecularFunctionScore BiologicalProcessScore CellularComponentScore
2                  0.158                  0.371                 0.364
```

7 References

1. Berretina J *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature, 283:603-607.

2. McCall MN, Bolstad BM, and Irizarry RA. (2010) Frozen Robust Multiarray Analysis. Biostatistics, 11(2):242-253.

3. Wang J *et al.* (2009) The Bimodality Index: A criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data. Cancer Informatics, 7:199-216.

4. Carlson, M. (2013) Org.Hs.eg.db: Genome wide annotation for human. R package version 2.8.0.

5. Guangchuang, Y *et al.* (2010) An R packahe for measuring semanic similarity among GO terms and gene products. Bioinformatics, 26(7), 976-978.