

cummeRbund: Visualization and Exploration of Cufflinks High-throughput Sequencing Data

Loyal A. Goff, Cole Trapnell

1 April, 2011

Contents

1	Introduction	1
2	CummeRbund Classes	2
2.1	CuffSet Class	2
2.2	CuffData Class	2
2.3	CuffFeatureSet Class	2
2.4	CuffFeature Class	3
3	Reading cuffdiff output	4
3.1	Adding additional feature annotation	4
4	Global statistics	5
5	Accessing Data	9
5.1	Writing your own SQL accessors	11
6	Creating Gene Sets	12
6.1	Geneset level plots	12
7	Individual Genes	18
7.1	Gene-level plots	19
8	Miscellaneous	19
9	Session info	22

1 Introduction

cummeRbund is a visualization package for Cufflinks high-throughput sequencing data. The base class, *cuffSet* is a 'pointer' to cufflinks data that are stored out-of-memory in a sqlite database.

2 CummeRbund Classes

2.1 CuffSet Class

A pointer class to control access to the sqlite tables holding the Cufflinks data. The primary slot is @DB which contains the RSQLite connection object. The additional slots (genes, isoforms, TSS, and CDS) are each instances of the *CuffData* class and are pointers to sets of tables for each data subtype. This is the default class created by *readCufflinks*. By default, *CuffData* accessor methods applied to a *CuffSet* class will operate on the 'genes' slot.

2.2 CuffData Class

The *CuffData* class is also a pointer class to the SQL backend, but each instance is specific for a data subtype (genes, isoforms, TSS, CDS). Again, there is an @DB slot that contains the RSQLite connection object. There are several accessor, setter, and plotting methods that allow for global analysis of all features within a *CuffData* class. Subsetting is currently being re-written, however, it is primarily done through the 'gene_id' field. Available slots for the CuffData class are:

- DB: RSQLite connection object
- tables: A *list* of tables in the SQLite DB that contain the cufflinks data.
- filters: A *list* of filters for subsetting (not implemented yet).
- type: A *character* field describing the data (ie. 'genes', 'isoforms', 'TSS', 'CDS', 'other')
- idField: The name of the identifying index field for this object (eg. 'gene_id' for type='gene', or 'isoform_id' for type='isoform')

Making the best use of either the CuffSet or CuffData classes will enable you to keep the entire dataset out of memory and significantly improve performance for large cufflinks datasets.

2.3 CuffFeatureSet Class

The *CuffFeatureSet* class is a data-storage container that holds all available data for a pre-determined list of features. Slots for FPKM data, differential regulation data, and feature-level annotation are all available. Unlike the previous classes, this class contains no connection information to the SQL database, but rather contains several slots with *data.frame* objects storing multiple-features worth of information. There are available accessors, and plotting methods that are designed to present multiple-features worth of information (eg. heatmaps, scatterplots, etc) Available slots for a *CuffFeatureSet* object include:

- annotation: Holds all feature-level annotation information for all features in object.

- fpkm: A data frame of FPKM data across all samples, for all features in object.
- diff: A data frame of differential expression/regulation data for all features in object.

A specialized sub-class of *CuffFeatureSet* is the *CuffGeneSet* class. This subclass adds additional slots to contain all isoforms, TSS, and CDS information for a given set of gene_ids. The *CuffGeneSet* class is designed to aggregate all relevant information for a set of genes into one object for easy analysis and/or manipulation. The *CuffGeneSet* object adds the following slots:

- ids: A 'character' list of all gene_ids used in object.
- isoforms: A *CuffFeatureSet* object for all isoforms of genes in object.
- TSS: A *CuffFeatureSet* object for all TSS of genes in object.
- CDS: A *CuffFeatureSet* object for all CDS of genes in object.

2.4 CuffFeature Class

The *CuffFeature* class is designed for single-feature-level data analysis and plotting. The methods available for this object are designed to analyze or visualize information about a specific feature. This is a 'data' object, as opposed to a 'pointer' object to the database backend. There is a validity requirement that a *CuffFeature* object only point to data from a single feature. Available slots for a *CuffFeature* object include:

- annotation: Holds feature-level annotation information for a given feature.
- fpkm: A data frame of FPKM data across all samples for a given feature.
- diff: A data frame of differential expression/regulation data for a given feature.

A specialized sub-class of *CuffFeature* is the *CuffGene* class. This subclass adds additional slots to contain all isoform, TSS, and CDS information for a given gene. The *CuffGene* object adds the following slots:

- id: The common 'gene_id' for all data in object
- isoforms: A *CuffFeature* object for all isoforms of a given gene.
- TSS: A *CuffFeature* object for all TSS of a given gene.
- CDS: A *CuffFeature* object for all CDS of a given gene.

Note: Future versions of cummeRbund may try to collapse the redundant functionality of the *CuffFeature* and *CuffFeatureSet* classes.

3 Reading cuffdiff output

One of the principle benefits of using cummeRbund is that data are stored in a SQLite database. This allows for out-of-memory analysis of data, quick retrieval, and only a one-time cost to setup the tables. By default, cummeRbund assumes that all output files from cuffdiff are in the current working directory. To read these files, populate the 'cuffData.db' database backend, and return the *CuffSet* pointer object, you can do the following.

```
> library(cummeRbund)

> cuff <- readCufflinks()
> cuff
```

CuffSet instance with:

```
3 samples
400 genes
1203 isoforms
575 TSS
545 CDS
960 promoters
1725 splicing
696 relCDS
```

Again, by default *dir* is assumed to be the current working directory and `cuff<-readCufflinks()` should work if all appropriate files are in the current working directory. Should you need to rebuild the SQLite backend for any reason, you can add the option `rebuild=T` to `readCufflinks`. Once the database is created, `readCufflinks` will default to using the SQL backend and should not need to rebuild this database. Each R session should begin with a call to `readCufflinks` so as to initialize the database connection and create an object with the appropriate RSQLite connection information.

3.1 Adding additional feature annotation

Gene- or feature-level annotation can be permanently added to the database tables for future querying. If you have a data.frame where the first column contains the 'tracking_id' (eg. 'gene_id' for genes, 'isoform_id' for isoforms, etc). You can easily add feature level annotation using the `addFeatures()` function:

By default, features added to a *CuffSet* object are assumed to be gene-level annotations, but the level can be selected using the argument *level*. Features added to a *CuffData* object are assumed to be of the same type as the object@type value (e.g. gene-level features for 'genes', isoform-level features for isoforms, etc.)

4 Global statistics

Several plotting methods are available that allow for quality-control or global analysis of cufflinks data. For example, to assess the distributions of FPKM scores across samples, you can use the *csDensity* plot (Figure 1).

```
> dens <- csDensity(cuff@genes)
> dens
```

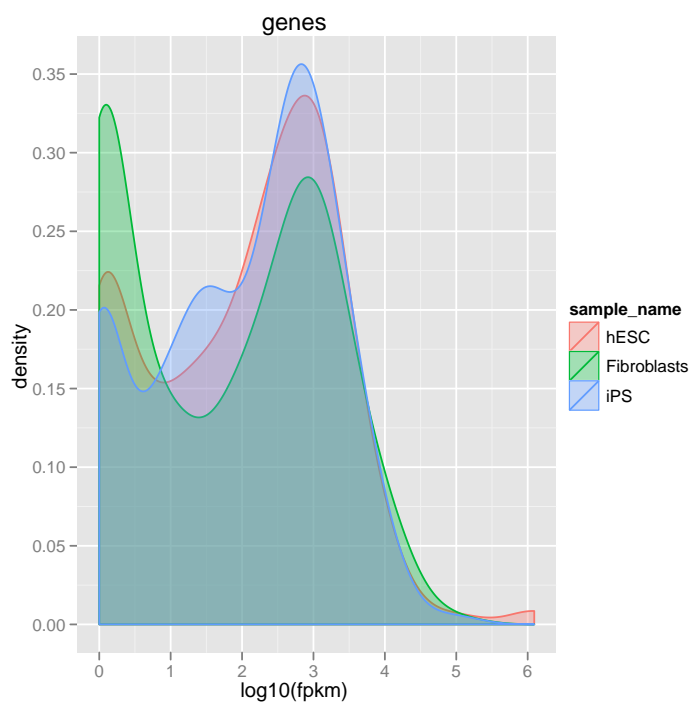


Figure 1: Density plot per sample of cufflinks output FPKM values.

Boxplots can be visualized using the *csBoxplot* method (Figure 2).

```
> b <- csBoxplot(cuff@genes)
> b
```

Pairwise comparisons can be made by using *csScatter*. You must specify the sample names to use for the *x* and *y* axes:

```
> s <- csScatter(cuff@genes, "hESC", "Fibroblasts",
+               smooth = T)
> s
```

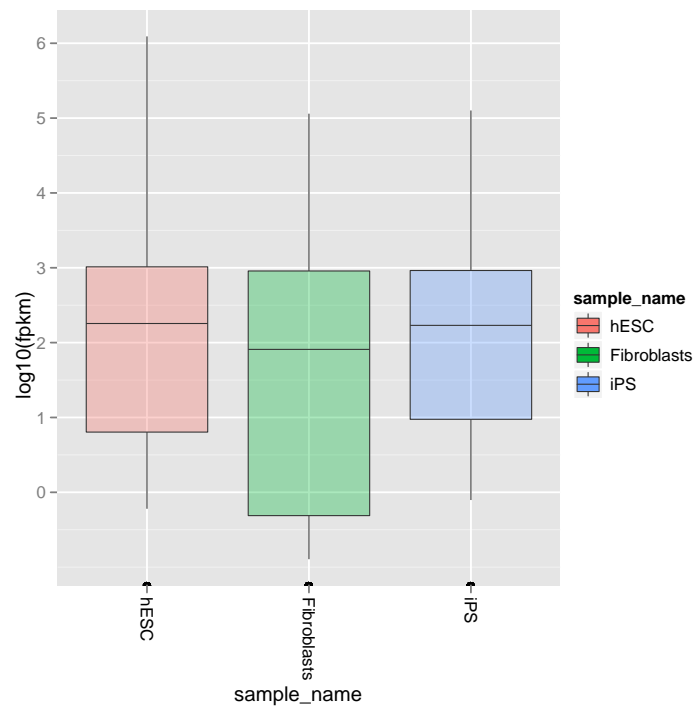


Figure 2: Box plot of FPKM values from cufflinks output.

Volcano plots are also available for the *CuffData* objects. Again, you must specify the comparisons by sample name.

```
> v <- csVolcano(cuff@genes, "hESC", "Fibroblasts")
> v
```

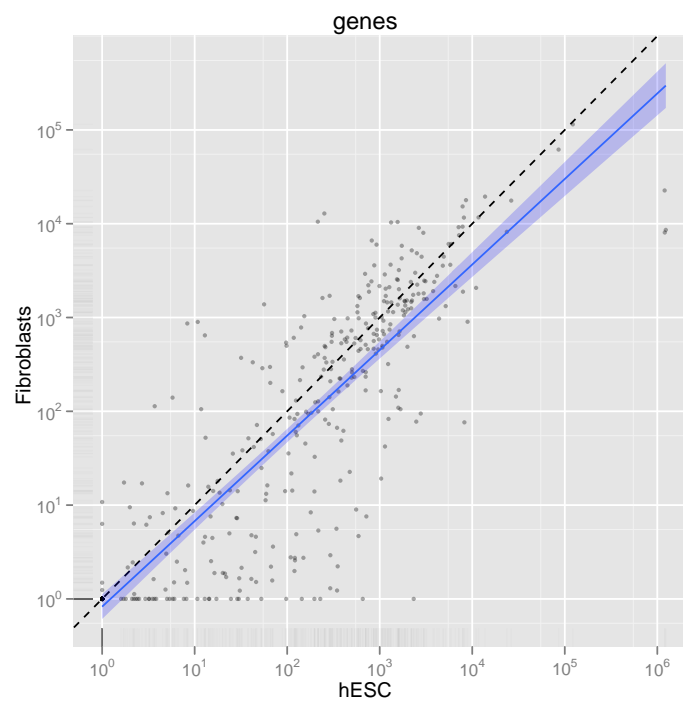


Figure 3: Scatter plot comparing the FPKM values of two samples from cufflinks output.

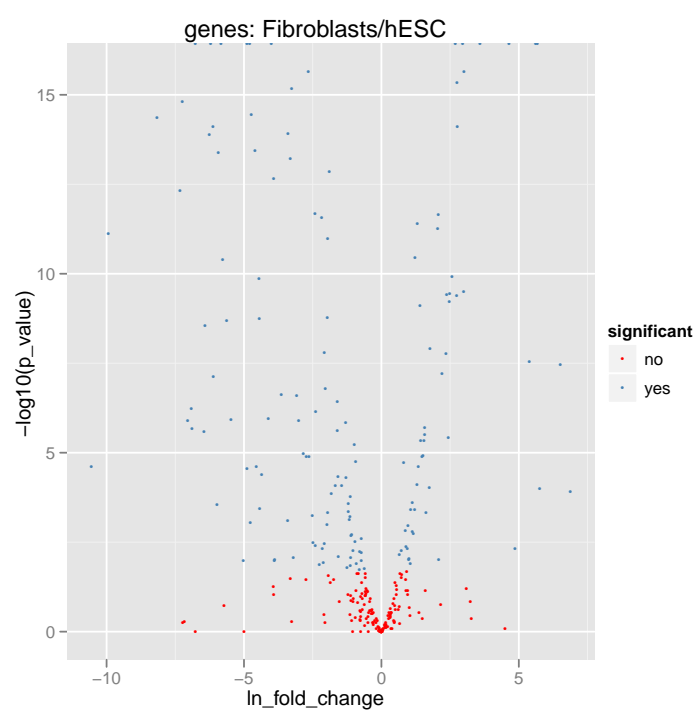


Figure 4: Volcano plot of \ln fold change vs significance.

5 Accessing Data

Feature-level information can be accessed directly from a *CuffData* object using the *fpkm*, *diffData*, or *features* methods:

```
> gene.features <- features(cuff@genes)
> head(gene.features)
```

	gene_id	class_code	nearest_ref_id	gene_short_name
1	XLOC_000001	<NA>	<NA>	<NA>
2	XLOC_000002	<NA>	<NA>	OR4F5
3	XLOC_000003	<NA>	<NA>	<NA>
4	XLOC_000004	<NA>	<NA>	<NA>
5	XLOC_000005	<NA>	<NA>	<NA>
6	XLOC_000006	<NA>	<NA>	OR4F16

	locus	length	coverage	gene_id
1	chr1:11873-29961	NA	NA	<NA>
2	chr1:69090-70008	NA	NA	<NA>
3	chr1:321083-321114	NA	NA	<NA>
4	chr1:321145-321223	NA	NA	<NA>
5	chr1:322036-328580	NA	NA	<NA>
6	chr1:367658-368595	NA	NA	<NA>

```
> gene.fpkm <- fpkm(cuff@genes)
> head(gene.fpkm)
```

	gene_id	sample_name	fpkm	conf_hi	conf_lo
1	XLOC_000001	Fibroblasts	16.401100	428.14700	0
2	XLOC_000001	hESC	0.723836	3.01108	0
3	XLOC_000001	iPS	54.067200	1402.31000	0
4	XLOC_000002	Fibroblasts	0.000000	0.00000	0
5	XLOC_000002	hESC	0.000000	0.00000	0
6	XLOC_000002	iPS	0.000000	0.00000	0

	quant_status
1	LOWDATA
2	OK
3	LOWDATA
4	OK
5	OK
6	OK

```
> isoform.fpkm <- fpkm(cuff@isoforms)
> head(isoform.fpkm)
```

	isoform_id	sample_name	fpkm	conf_hi	conf_lo
1	TCONS_000000001	Fibroblasts	11.910700	19.96650	3.85498
2	TCONS_000000001	hESC	0.000000	0.00000	0.00000

```

3 TCONS_00000001      iPS  9.563700 23.68410 0.00000
4 TCONS_00000002 Fibroblasts 0.000000 8.55378 0.00000
5 TCONS_00000002      hESC  0.723836 3.01108 0.00000
6 TCONS_00000002      iPS 32.934400 47.93760 17.93130
  quant_status
1          OK
2          OK
3      LOWDATA
4          OK
5          OK
6          OK

> gene.diff <- diffData(cuff@genes)
> head(gene.diff)

      gene_id sample_1  sample_2 status  value_1  value_2
1 XLOC_000001      hESC Fibroblasts    OK 7.23836e-01 16.4011
2 XLOC_000002      hESC Fibroblasts NOTEST 0.00000e+00 0.0000
3 XLOC_000003      hESC Fibroblasts NOTEST 0.00000e+00 0.0000
4 XLOC_000004      hESC Fibroblasts    OK 1.20000e+06 22616.4000
5 XLOC_000005      hESC Fibroblasts    OK 1.13903e+03 41.1644
6 XLOC_000006      hESC Fibroblasts NOTEST 0.00000e+00 0.0000
  ln_fold_change test_stat  p_value  q_value significant
1      4.50198 -0.246654 0.805176 0.893616          no
2      0.00000 0.000000 1.000000 1.000000          no
3      0.00000 0.000000 1.000000 1.000000          no
4     -5.72952 1.310270 0.190105 0.300329          no
5     -4.79027 10.857600 0.000000 0.000000          yes
6      0.00000 0.000000 1.000000 1.000000          no

```

Vectors of sample names and feature names are available by using the *samples* and *featureNames* methods:

```

> sample.names <- samples(cuff@genes)
> head(sample.names)

[1] "hESC"      "Fibroblasts" "iPS"

> gene.featurenames <- featureNames(cuff@genes)
> head(gene.featurenames)

[1] "XLOC_000001" "XLOC_000002" "XLOC_000003" "XLOC_000004"
[5] "XLOC_000005" "XLOC_000006"

```

To facilitate Bioconductor-like operations, an 'FPKM-matrix' can be returned easily using the *fpkmMatrix* method:

```

> gene.matrix <- fpkmMatrix(cuff@genes)
> head(gene.matrix)

```

	hESC Fibroblasts	iPS
XL0C_000001	7.23836e-01	16.4011 54.06720
XL0C_000002	0.00000e+00	0.0000 0.00000
XL0C_000003	0.00000e+00	0.0000 0.00000
XL0C_000004	1.20000e+06	22616.4000 0.00000
XL0C_000005	1.13903e+03	41.1644 944.30800
XL0C_000006	0.00000e+00	0.0000 9.00455

5.1 Writing your own SQL accessors

Since the cufflinks is a SQLite database backend, if you are familiar with SQL and/or RSQLite query construction, you can simply design your own SQL queries to access the data that you are after. PUT DATABASE SCHEMA HERE...

6 Creating Gene Sets

Gene Sets (stored in a *CuffGeneSet* object) can be created using the *getGenes* method on a *CuffSet* object. You must first create a vector of 'gene_ids' to identify the genes you wish to select:

```
> data(sampleData)
> myGeneIds <- sampleIDs
> myGeneIds

[1] "XLOC_001363" "XLOC_001297" "XLOC_001339" "XLOC_000132"
[5] "XLOC_001265" "XLOC_000151" "XLOC_001359" "XLOC_000069"
[9] "XLOC_000170" "XLOC_000105" "XLOC_001262" "XLOC_001348"
[13] "XLOC_001411" "XLOC_001369" "XLOC_000158" "XLOC_001370"
[17] "XLOC_001263" "XLOC_000115" "XLOC_000089" "XLOC_001240"

> myGenes <- sampleGeneSet
> myGenes

CuffGeneSet instance for genes c("XLOC_000069", "XLOC_000089", "XLOC_000105", "XLOC_000115",
Short name:      ESPN PGD MFN2 PRAMEF1 EFHD2 PADI1 NA FAM43B UBE2J2 C1orf86 SLC2A7 SPATA2
Slots:
      annotation
      fpkm
      diff
      isoforms      CuffFeatureSet instance of size 45
      TSS           CuffFeatureSet instance of size 18
      CDS           CuffFeatureSet instance of size 31
```

The same *fpkm*, *fpkmMatrix*, *features*, *diffData*, *samples*, and *featureNames* are available for instances of the *CuffGeneSet* class.

6.1 Geneset level plots

There are several plotting functions available for gene-set-level visualization:

```
> h <- csHeatmap(myGenes, cluster = "both")
> h

> s <- csScatter(myGenes, "Fibroblasts", "hESC", smooth = T)
> s

> v <- csVolcano(myGenes, cluster = "both")
> v
```

Similar plots can be made for all sub-level features of a *CuffGeneSet* class by specifying which slot you would like to plot (eg. *@isoforms*, *@TSS*, *@CDS*).

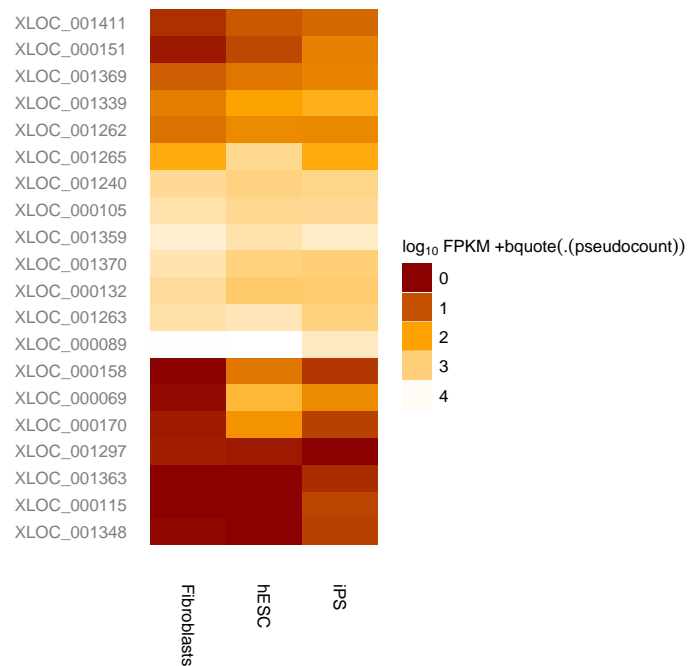


Figure 5: Heatmap of FPKM values for a random sample of 20 genes.

```
> ih <- csHeatmap(myGenes@isoforms, cluster = "both",
+   labRow = F)
> ih
```

Rudimentary k-means clustering is implemented as well.

```
> ic <- csCluster(myGenes, k = 4)
> ic
```

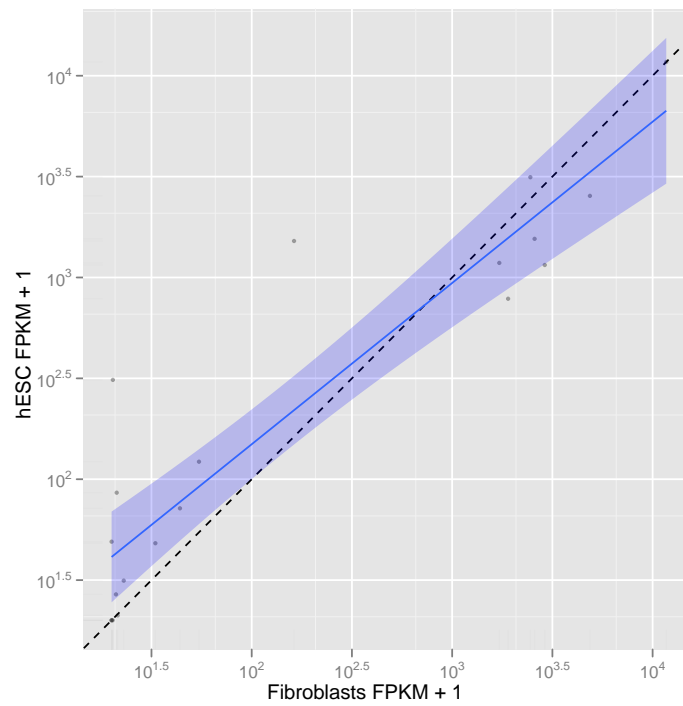


Figure 6: Scatterplot of FPKM values for a random sample of 20 genes between Fibroblasts and H1_hESC.

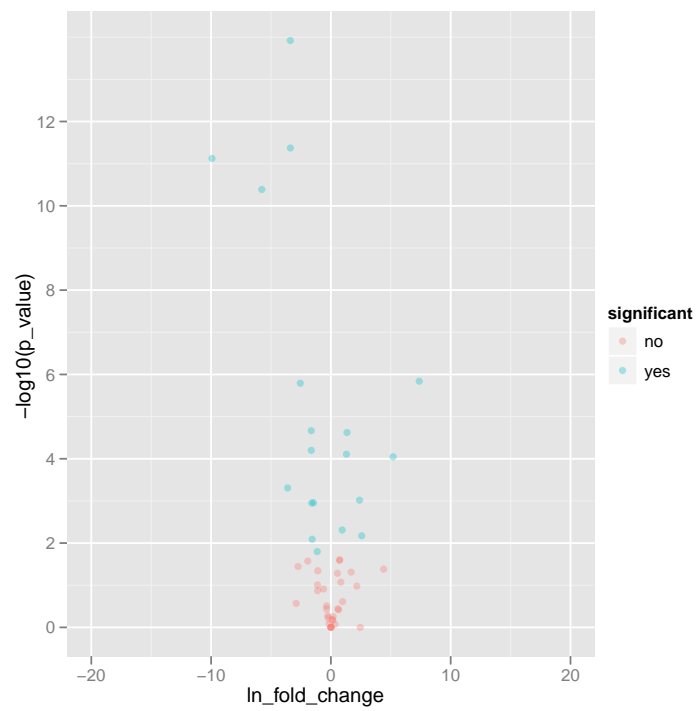


Figure 7: Volcano plot of FPKM vs significance values for a random sample of 20 genes between 2 conditions.

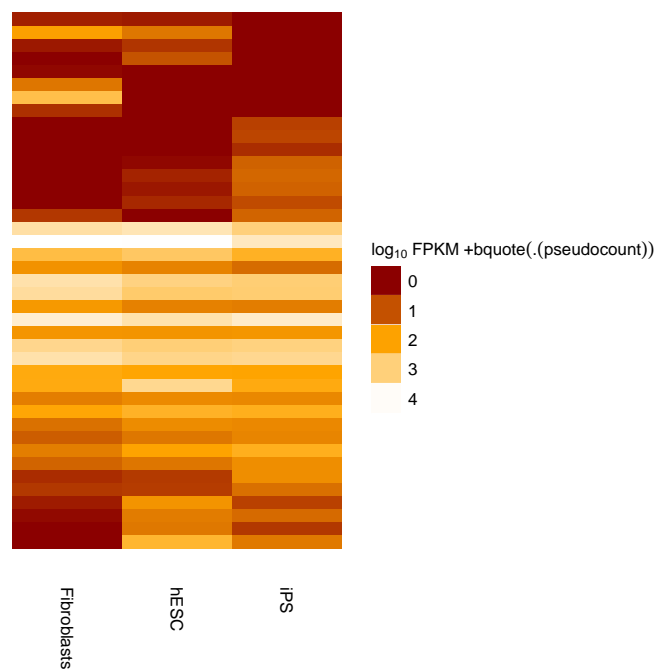


Figure 8: Heatmap of FPKM values of isoforms for a random sample of 20 genes.

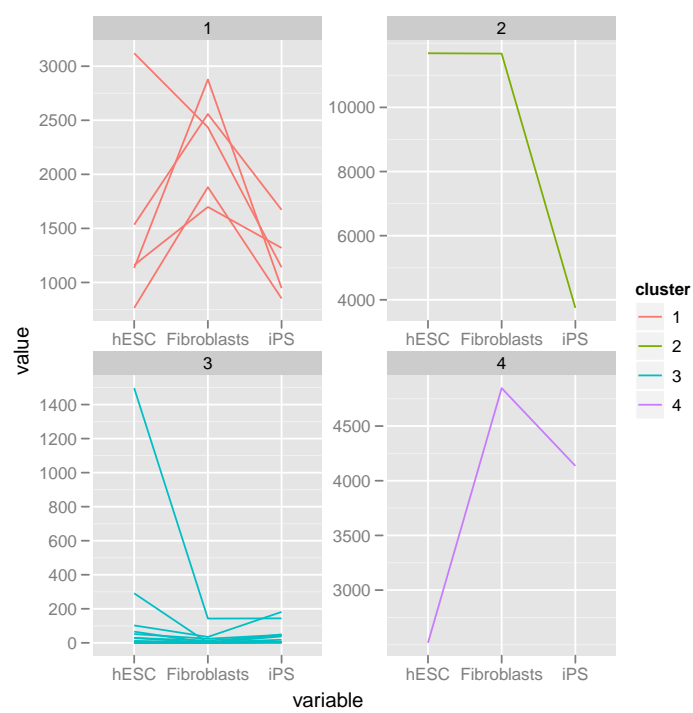


Figure 9: Basic k-means clustering of a CuffGeneSet set of 20 random genes.

7 Individual Genes

An individual CuffGene object can be created by using the `getGene` function for a given 'gene_id'.

```
> myGeneId <- "PINK1"
> myGene <- getGene(cuff, myGeneId)
> myGene
```

CuffGene instance for gene PINK1

Short name: PINK1

Slots:

```
      annotation
      fpkm
      diff
      isoforms      CuffFeature instance of size 2
      TSS           CuffFeature instance of size 2
      CDS           CuffFeature instance of size 2
```

```
> head(fpkm(myGene))
```

	gene_id	sample_name	fpkm	conf_hi	conf_lo
1	XLOC_000172	Fibroblasts	2919.340	4002.960	1835.730
2	XLOC_000172	hESC	693.465	813.869	573.062
3	XLOC_000172	iPS	1598.040	2282.380	913.710

quant_status

1	OK
2	OK
3	OK

```
> head(fpkm(myGene@isoforms))
```

	isoform_id	sample_name	fpkm	conf_hi	conf_lo
1	TCONS_00000480	Fibroblasts	2101.640	3111.330	1091.9400
2	TCONS_00000481	Fibroblasts	817.704	1391.700	243.7120
3	TCONS_00000480	hESC	573.512	668.688	478.3370
4	TCONS_00000481	hESC	119.953	152.675	87.2311
5	TCONS_00000480	iPS	1598.040	2282.380	913.7100
6	TCONS_00000481	iPS	0.000	0.000	0.0000

quant_status

1	OK
2	OK
3	OK
4	OK
5	OK
6	OK

7.1 Gene-level plots

```
> gl <- expressionPlot(myGene)
> gl
```

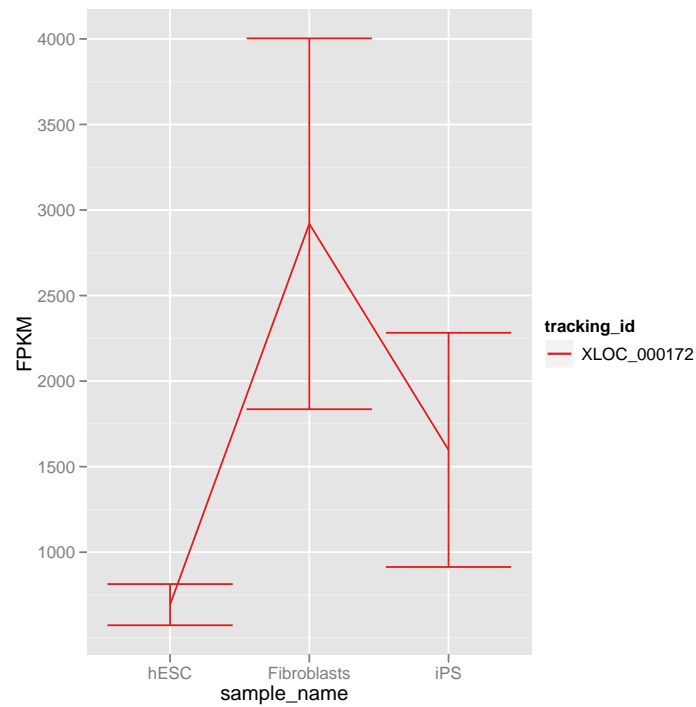


Figure 10: Line plot of FPKM expression values for a given gene

```
> gb <- expressionBarplot(myGene)
> gb

> igb <- expressionBarplot(myGene@isoforms)
> igb
```

8 Miscellaneous

- All plotting functions return ggplot objects and the resulting objects can be manipulated/faceted/alttered using standard ggplot2 methods.

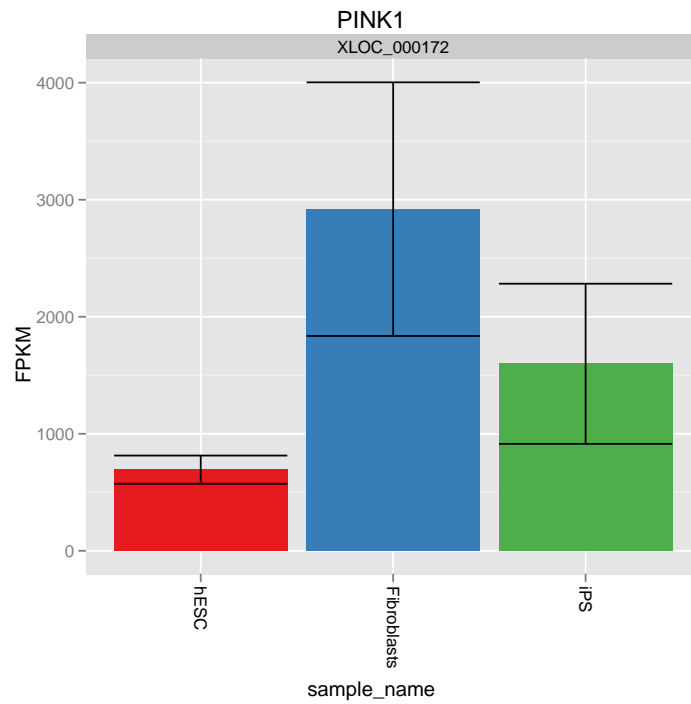


Figure 11: Bar plot of FPKM expression values for a given gene

- There are occasional DB connectivity issues that arise. Not entirely sure why yet. If necessary, just `readCufflinks` again and this should solve connectivity issues with a new RSQLite connection object. If connectivity continues to be a problem, try `cuff<-readCufflinks(rebuild=T)`
- I am still working on fully documenting each of the methods. There are a good number of arguments that exist, but might be hard to find without looking at the source.

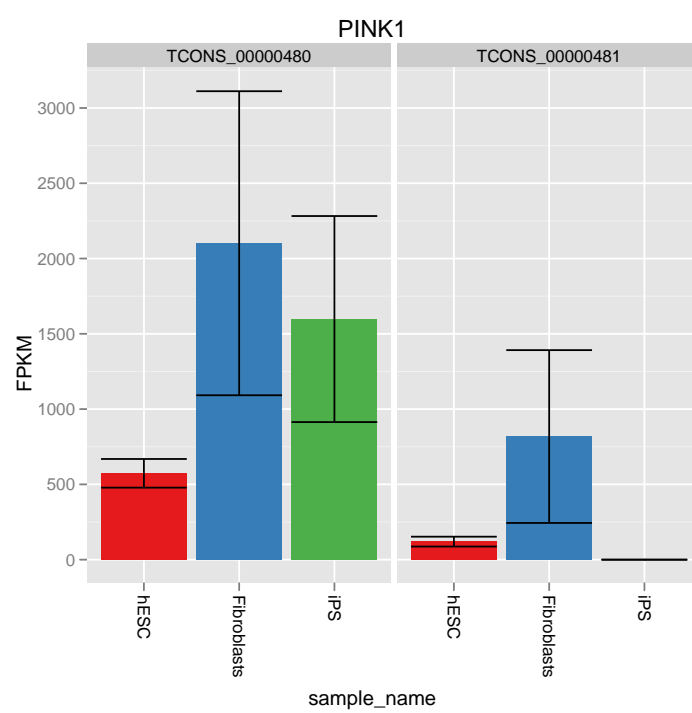


Figure 12: Bar plot of FPKM expression values for all isoforms of a given gene

9 Session info

```
> sessionInfo()

R version 2.12.1 (2010-12-16)
Platform: x86_64-apple-darwin9.8.0/x86_64 (64-bit)

locale:
[1] C/en_US.utf-8/C/C/en_US.utf-8/en_US.utf-8

attached base packages:
[1] grid      stats      graphics  grDevices  utils      datasets
[7] methods   base

other attached packages:
[1] cummeRbund_0.1.3 ggplot2_0.8.9    proto_0.3-8
[4] reshape_0.8.3    plyr_1.4         RSQLite_0.9-4
[7] DBI_0.2-5

loaded via a namespace (and not attached):
[1] RColorBrewer_1.0-2 digest_0.4.2      tools_2.12.1
```