

# DiffBind : differential binding analysis of ChIP-Seq peak data

Rory Stark

Gordon Brown

`rory.stark@cruk.cam.ac.uk`

`gordon.brown@cruk.cam.ac.uk`

University of Cambridge  
Cancer Research UK - Cambridge Institute

27 September 2013

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Processing overview</b>	<b>2</b>
<b>3</b>	<b>Example: obtaining differentially bound sites</b>	<b>4</b>
3.1	Reading in the peaksets . . . . .	4
3.2	Counting reads . . . . .	7
3.3	Establishing a contrast . . . . .	7
3.4	Performing the differential analysis . . . . .	7
3.5	Retrieving the differentially bound sites . . . . .	9
<b>4</b>	<b>Example: plotting</b>	<b>11</b>
4.1	Venn diagrams . . . . .	11
4.2	MA plots . . . . .	11
4.3	PCA plots . . . . .	11
4.4	Boxplots . . . . .	13
4.5	Heatmaps . . . . .	15
<b>5</b>	<b>Example: differential binding analysis using a blocking factor</b>	<b>16</b>
<b>6</b>	<b>Example: occupancy analysis and overlaps</b>	<b>22</b>
6.1	Overlap rates . . . . .	24
6.2	Deriving consensus peaksets . . . . .	24
6.3	A complete occupancy analysis: identifying sites unique to a sample group . . . . .	28

6.4	Comparison of occupancy and affinity based analyses . . . . .	31
<b>7</b>	<b>Technical notes</b>	<b>34</b>
7.1	Loading peaksets . . . . .	34
7.2	Merging peaks . . . . .	35
7.3	edgeR analysis . . . . .	35
7.4	DESeq analysis . . . . .	36
7.5	DESeq2 analysis . . . . .	37
<b>8</b>	<b>Vignette Data</b>	<b>38</b>
<b>9</b>	<b>Acknowledgements</b>	<b>38</b>
<b>10</b>	<b>Setup</b>	<b>38</b>

# 1 Introduction

This document offers an introduction and overview of the R Bioconductor package `DiffBind`, which provides functions for processing ChIP-Seq data enriched for genomic loci where specific protein/DNA binding occurs, including peak sets identified by ChIP-Seq peak callers and aligned sequence read datasets. It is designed to work with multiple peak sets simultaneously, representing different ChIP experiments (antibodies, transcription factor and/or histone marks, experimental conditions, replicates) as well as managing the results of multiple peak callers.

The primary emphasis of the package is on identifying sites that are differentially bound between two sample groups. It includes functions to support the processing of peak sets, including overlapping and merging peak sets, counting sequencing reads overlapping intervals in peak sets, and identifying statistically significantly differentially bound sites based on evidence of binding affinity (measured by differences in read densities). To this end it uses statistical routines developed in an RNA-Seq context (primarily the Bioconductor packages `edgeR` and `DESeq`). Additionally, the package builds on R graphics routines to provide a set of standardized plots to aid in binding analysis.

This guide includes a brief overview of the processing flow, followed by three sections of examples: the first focusing on the core task of obtaining differentially bound sites based on affinity data, the second working through the main plotting routines, and the third revisiting occupancy data (peak calls) in more detail, as well as comparing the results of an occupancy-based analysis with an affinity-based one. Finally, some technical aspects of how these analyses are accomplished are detailed.

# 2 Processing overview

`DiffBind` works primarily with peaksets, which are sets of genomic intervals representing candidate protein binding sites. Each interval consists of a chromosome, a start and end position, and usually

a score of some type indicating confidence in, or strength of, the peak. Associated with each peakset are metadata relating to the experiment from which the peakset was derived. Additionally, files containing mapped sequencing reads (BAM//BED) can be associated with each peakset (one for the ChIP data, and optionally another representing a control dataset).

Generally, processing data with DiffBind involves five phases:

1. **Reading in peaksets:** The first step is to read in a set of peaksets and associated metadata. Peaksets are derived either from ChIP-Seq peak callers, such as MACS (Zhang et al. [2008]), or using some other criterion (e.g. all the promoter regions in a genome). The easiest way to read in peaksets is using a comma-separated value (csv) sample sheet with one line for each peakset. A single experiment can have more than one associated peakset, e.g. if multiple peak callers are used for comparison purposes, and hence have more than one line in the sample sheet. Once the peaksets are read in, a merging function finds all overlapping peaks and derives a single set of unique genomic intervals covering all the supplied peaks.
2. **Occupancy analysis:** Peaksets, especially those generated by peak callers, provide an insight into the potential *occupancy* of the protein being ChIPed for at specific genomic loci. After the peaksets have been loaded, it can be useful to perform some exploratory plotting to determine how these occupancy maps agree with each other, e.g. between experimental replicates (re-doing the ChIP under the same conditions), between different peak callers on the same experiment, and within groups of samples representing a common experimental condition. DiffBind provides functions to enable overlaps to be examined, as well as functions to determine how well similar samples cluster together. Beyond quality control, the product of an occupancy analysis may be a *consensus peakset*, representing an overall set of candidate binding sites to be used in further analysis.
3. **Counting reads:** Once a consensus peakset has been derived, DiffBind can use the supplied sequence read files to count how many reads overlap each interval for each unique sample. The result of this is a *binding affinity matrix* containing a (normalized) read count for each sample at every potential binding site. With this matrix, the samples can be re-clustered using affinity, rather than occupancy, data. The binding affinity matrix is used for QC plotting as well as for subsequent differential analysis.
4. **Differential binding affinity analysis:** The core functionality of DiffBind is the differential binding affinity analysis, which enables binding sites to be identified that are statistically significantly differentially bound between sample groups. To accomplish this, first a contrast (or contrasts) is established, dividing the samples into groups to be compared. Next the core analysis routines are executed, by default using `edgeR`. This will assign a p-value and FDR to each candidate binding site indicating the significance of their being differentially bound.
5. **Plotting and reporting:** Once one or more contrasts have been run, DiffBind provides a number of functions for reporting and plotting the results. MA plots give an overview of the results of the analysis, while correlation heatmaps and PCA plots show how the groups

cluster based on differentially bound sites. Boxplots show the distribution of reads within differentially bound sites corresponding to whether they gain or lose affinity between the two sample groups. A reporting mechanism enables differentially bound sites to be extracted for further processing, such as annotation and/or pathway analysis.

### 3 Example: obtaining differentially bound sites

This section offers a quick example of how to use DiffBind to identify significantly differentially bound sites using affinity (read count) data.

The dataset for this example consists of ChIPs against the transcription factor ERa using five breast cancer cell lines (Ross-Innes et al. [2012]). Three of these cell lines are responsive to tamoxifen, while two others are resistant to tamoxifen treatment. There are at least two replicates for each of the cell lines, with one cell line having three replicates, for a total of eleven sequenced libraries. Note that this experiment includes two types of MCF7 cells: the regular tamoxifen responsive line as well as MCF7 cells specially treated with tamoxifen until a tamoxifen resistant cell line is obtained. For each sample, we have one peakset originally derived using the MACS peak caller (Zhang et al. [2008]), for a total of eleven peaksets. Note that to save space in the package, only data for chromosome 18 is used. The metadata and peak data are available in the `extra` subdirectory of the DiffBind package directory; you can make this your working directory by entering:

```
> library(DiffBind)
> setwd(system.file("extra", package="DiffBind"))
```

Obtaining the sites significantly differentially bound (DB) between the samples that respond to tamoxifen and those that are resistant can be done in a five-step script:

```
> tamoxifen = dba(sampleSheet="tamoxifen.csv")
> tamoxifen = dba.count(tamoxifen)
> tamoxifen = dba.contrast(tamoxifen, categories=DBA_CONDITION)
> tamoxifen = dba.analyze(tamoxifen)
> tamoxifen.DB = dba.report(tamoxifen)
```

The following subsections describe these steps in more detail.

#### 3.1 Reading in the peaksets

Table 1 shows the sample sheet, saved in a file called `tamoxifen.csv`. The peaksets are read in using the following DiffBind function:

```
> tamoxifen = dba(sampleSheet="tamoxifen.csv")
```

Table 1: Tamoxifen dataset sample sheet (tamoxifen.csv).

SampleID	Tissue	Factor	Condition	Replicate	bamReads	bamControl	Peaks
BT474.1-	BT474	ER	Resistant	1	BT474_ER.1.bed.gz	BT474_Input.bed.gz	BT474_ER.1.bed.gz
BT474.2-	BT474	ER	Resistant	2	BT474_ER.2.bed.gz	BT474_Input.bed.gz	BT474_ER.2.bed.gz
MCF7.1+	MCF7	ER	Responsive	1	MCF7_ER.1.bed.gz	MCF7_Input.bed.gz	MCF7_ER.1.bed.gz
MCF7.2+	MCF7	ER	Responsive	2	MCF7_ER.2.bed.gz	MCF7_Input.bed.gz	MCF7_ER.2.bed.gz
MCF7.3+	MCF7	ER	Responsive	3	MCF7_ER.3.bed.gz	MCF7_Input.bed.gz	MCF7_ER.3.bed.gz
T47D.1+	T47D	ER	Responsive	1	T47D_ER.1.bed.gz	T47D_Input.bed.gz	T47D_ER.1.bed.gz
T47D.2+	T47D	ER	Responsive	2	T47D_ER.2.bed.gz	T47D_Input.bed.gz	T47D_ER.2.bed.gz
MCF7.1-	MCF7	ER	Resistant	1	TAMR_ER.1.bed.gz	TAMR_Input.bed.gz	TAMR_ER.1.bed.gz
MCF7.2-	MCF7	ER	Resistant	2	TAMR_ER.2.bed.gz	TAMR_Input.bed.gz	TAMR_ER.2.bed.gz
ZR75.1+	ZR75	ER	Responsive	1	ZR75_ER.1.bed.gz	ZR75_Input.bed.gz	ZR75_ER.1.bed.gz
ZR75.2+	ZR75	ER	Responsive	2	ZR75_ER.2.bed.gz	ZR75_Input.bed.gz	ZR75_ER.2.bed.gz

The result is a `DBA` object; the metadata associated with this object can be displayed simply as follows:

```
> tamoxifen

11 Samples, 2602 sites in matrix (3557 total):
      ID Tissue Factor Condition Replicate Caller Intervals
1  BT474.1-  BT474    ER  Resistant         1    raw      1084
2  BT474.2-  BT474    ER  Resistant         2    raw      1115
3   MCF7.1+  MCF7    ER  Responsive         1    raw      1513
4   MCF7.2+  MCF7    ER  Responsive         2    raw      1037
5   MCF7.3+  MCF7    ER  Responsive         3    raw      1372
6   T47D.1+  T47D    ER  Responsive         1    raw       509
7   T47D.2+  T47D    ER  Responsive         2    raw       347
8   MCF7.1-  MCF7    ER  Resistant         1    raw      1148
9   MCF7.2-  MCF7    ER  Resistant         2    raw       933
10  ZR75.1+  ZR75    ER  Responsive         1    raw      2111
11  ZR75.2+  ZR75    ER  Responsive         2    raw      1975
```

This shows how many peaks are in each peakset, as well as (in the first line) total number of unique peaks after merging overlapping ones (3557) and the default binding matrix of 11 samples by the 2602 sites that overlap in at least two of the samples. This object is available for loading using `data(tamoxifen_peaks)`.

Using only this peak caller data, a correlation heatmap can be generated which gives an initial clustering of the samples using the cross-correlations of each row of the binding matrix:

```
> plot(tamoxifen)
```

The resulting plot (Figure 1) shows that while the replicates for each cell line cluster together appropriately, the cell lines do not cluster into groups corresponding to those that are responsive (MCF7+, T47D, and ZR75) vs. those resistant (BT474 and MCF7-) to tamoxifen treatment. It also shows that the two most highly correlated cell lines are the two MCF7-based ones, even though they respond differently to tamoxifen treatment.

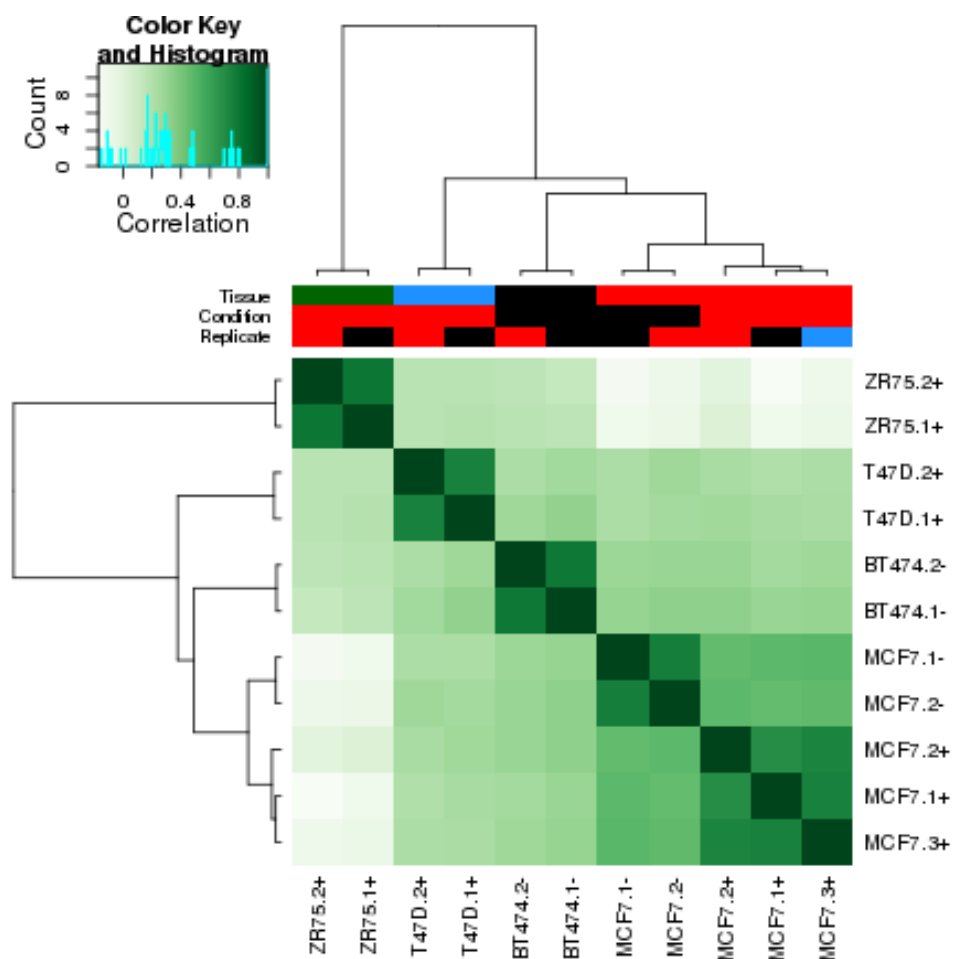


Figure 1: Correlation heatmap, using occupancy (peak caller score) data. Generated by: `plot(tamoxifen)`; can also be generated by: `dba.plotHeatmap(tamoxifen)`.

## 3.2 Counting reads

The next step is to calculate a binding matrix with scores based on read counts for every sample (affinity scores), rather than confidence scores for only those peaks called in a specific sample (occupancy scores). These reads are obtained using the `dba.count` function:<sup>1</sup>

```
> tamoxifen = dba.count(tamoxifen, minOverlap=3)
```

If you do not have the raw reads available to you, this object is available loading using `data(tamoxifen_counts)`. The `dba.count` call plots a new correlation heatmap based on the affinity scores, seen in Figure 2a. While this shows overall higher correlation levels (evidenced by the shift in the scale), and a slightly different clustering, responsiveness to tamoxifen treatment does not appear to form a basis for clustering when using all of the affinity scores. (Note that the clustering can change based on what scoring metric is used; see Section 4.4 for more details).

## 3.3 Establishing a contrast

Before running the differential analysis, we need to tell DiffBind which cell lines fall in which groups. This is done using the `dba.contrast` function, as follows:

```
> tamoxifen = dba.contrast(tamoxifen, categories=DBA_CONDITION)
```

The uses the *condition* metadata (Responsive vs. Resistant) to set up a contrast with 4 samples in the Resistant group and 7 samples in the Responsive group.<sup>2</sup>

## 3.4 Performing the differential analysis

The main differential analysis function is invoked as follows:

```
> tamoxifen = dba.analyze(tamoxifen)
```

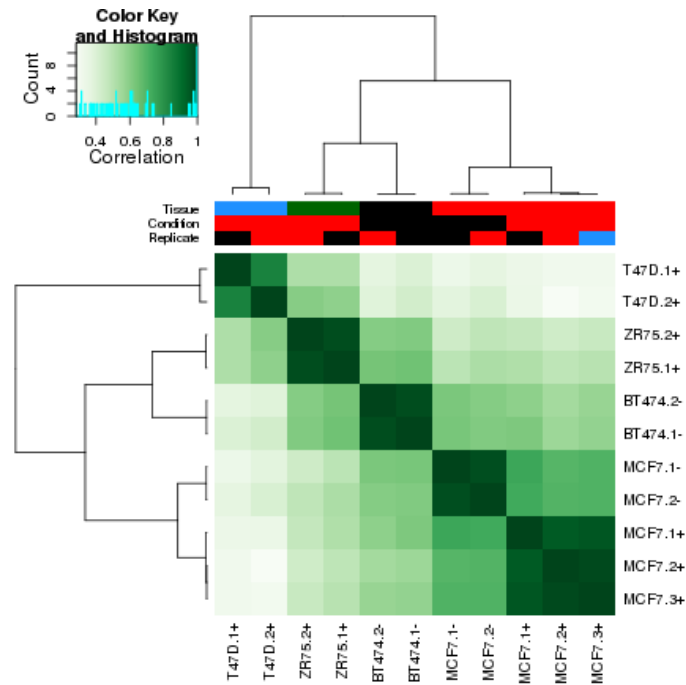
This will run an `edgeR` analysis (see subsequent section discussing the technical details of the `edgeR` analysis) on the binding affinity matrix. Displaying the resultant `DBA` object shows that 238 of the 1654 sites are identified as being significantly differentially bound (DB) using the default threshold of  $FDR \leq 0.1$

```
> tamoxifen
```

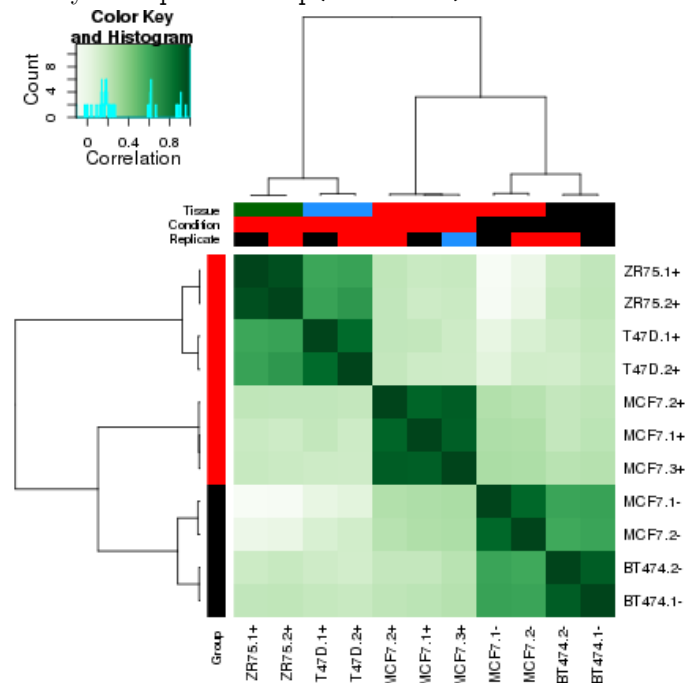
---

<sup>1</sup>Note that due to space limitations the reads are not shipped with the package. See Section 8 for options to obtain the full dataset. Alternatively, you can get the result of the `dba.count` call by loading the supplied R object by invoking `data(tamoxifen_counts)`

<sup>2</sup>This step is actually optional: if the main analysis function `dba.analyze` is invoked with no contrasts established, DiffBind will set up a default set of contrasts automatically, which will include the one we are interested in.



(a) Correlation heatmap, using affinity (read count) data. Generated by: `tamoxifen = dba.count(tamoxifen)`; can also be generated by: `dba.plotHeatmap(tamoxifen)`



(b) Correlation heatmap, using only significantly differentially bound sites. Generated by: `tamoxifen = dba.analyze(tamoxifen)`; can also be generated by: `dba.plotHeatmap(tamoxifen, contrast=1)`

Figure 2: Correlation heatmap plots, generated using `dba.plotHeatmap`.



11 Samples, 1654 sites in matrix:

	ID	Tissue	Factor	Condition	Replicate	Caller	Intervals	SN
1	BT474.1-	BT474	ER	Resistant	1	counts	1654	0.19
2	BT474.2-	BT474	ER	Resistant	2	counts	1654	0.18
3	MCF7.1+	MCF7	ER	Responsive	1	counts	1654	0.35
4	MCF7.2+	MCF7	ER	Responsive	2	counts	1654	0.22
5	MCF7.3+	MCF7	ER	Responsive	3	counts	1654	0.28
6	T47D.1+	T47D	ER	Responsive	1	counts	1654	0.14
7	T47D.2+	T47D	ER	Responsive	2	counts	1654	0.08
8	MCF7.1-	MCF7	ER	Resistant	1	counts	1654	0.25
9	MCF7.2-	MCF7	ER	Resistant	2	counts	1654	0.16
10	ZR75.1+	ZR75	ER	Responsive	1	counts	1654	0.35
11	ZR75.2+	ZR75	ER	Responsive	2	counts	1654	0.24

1 Contrast:

	Group1	Members1	Group2	Members2	DB.edgeR
1	Resistant	4	Responsive	7	238

By default, `dba.analyze` plots a correlation heatmap if it finds any significantly differentially bound sites, shown in Figure 2b. Using only the differentially bound sites, we now see that the four tamoxifen resistant samples (representing two cell lines) cluster together, although the tamoxifen-responsive MCF7 replicates cluster closer to them than to the other tamoxifen responsive samples. Comparing Figure 2a, which uses all 1654 consensus binding sites, with Figure 2b, which uses only the 238 differentially bound sites, demonstrates how the differential binding analysis isolates sites that help distinguish between the Resistant and Responsive sample groups.

### 3.5 Retrieving the differentially bound sites

The final step is to retrieve the differentially bound sites as follows:

```
> tamoxifen.DB = dba.report(tamoxifen)
```

These are returned as a `GRanges` object, appropriate for downstream processing:

```
> tamoxifen.DB
```

GRanges with 238 ranges and 6 metadata columns:

	seqnames	ranges	strand		Conc
	<Rle>	<IRanges>	<Rle>		<numeric>
1433	chr18	[62640747, 62642453]	*		7.23866506975617
7	chr18	[ 384853, 386517]	*		7.0786431629282
1606	chr18	[72497301, 72498984]	*		8.14521029208161
156	chr18	[ 7635601, 7636863]	*		6.46472243896863

```

1238 chr18 [53719311, 53721628] * | 6.82899613594856
...
243 chr18 [10005095, 10005860] * | 5.52517288446464
435 chr18 [18016530, 18017512] * | 5.55255473591273
967 chr18 [41561813, 41562543] * | 4.47989303013818
1377 chr18 [58864423, 58865252] * | 4.90182477121527
1177 chr18 [51567856, 51568947] * | 4.98767575896437

Conc_Resistant Conc_Responsive Fold p.value
<numeric> <numeric> <numeric> <numeric>
1433 3.04664715030695 7.86174997158677 -4.81510282127981 2.67570911112791e-08
7 8.33585259098193 4.79603648089405 3.53981611008788 2.73305594926472e-07
1606 4.47793289609017 8.75539103745019 -4.27745814136002 2.9850692789474e-07
156 7.61678962756677 4.73510012334251 2.88168950422426 4.73932446490167e-07
1238 3.99836527307059 7.40537545704353 -3.40701018397294 7.77481632584023e-07
...
243 6.49131254859287 4.38937190587354 2.10194064271933 0.0138922686351343
435 6.42856759447956 4.61657648602423 1.81199110845533 0.0139446594423061
967 2.89702643139354 4.94526727345692 -2.04824084206338 0.0140207023589831
1377 2.69508792541359 5.43553122217969 -2.7404432967661 0.0142297285301061
1177 3.42539689954277 5.45018237516861 -2.02478547562584 0.0142361214261219

FDR
<numeric>
1433 4.42562286980557e-05
7 0.0001645768195793
1606 0.0001645768195793
156 0.000195971066623684
1238 0.000257190924058795
...
243 0.0981466668832952
435 0.0981466668832952
967 0.0982637360243987
1377 0.098935062347923
1177 0.098935062347923
---
seqlengths:
chr18
NA

```

The value columns show the mean read concentration over all the samples (the default calculation uses log2 normalized ChIP read counts with control read counts subtracted) and the mean concentration over the first (Resistant) group and second (Responsive) group. The Fold column shows the difference in mean concentrations between the two groups (Conc\_Resistant - Conc\_Responsive),

with a positive value indicating increased binding affinity in the Resistant group and a negative value indicating increased binding affinity in the Responsive group. The final two columns give confidence measures for identifying these sites as differentially bound, with a raw p-value and a multiple testing corrected FDR in the final column.

## 4 Example: plotting

Besides the correlation heatmaps automatically generated by the core functions, a number of other plots are available using the affinity data. This sections covers Venn diagrams, MA plots, PCA plots, Boxplots, and Heatmaps.

### 4.1 Venn diagrams

Venn diagrams are useful for examining overlaps between peaksets, particularly when determining how best to derive consensus peaksets for further analysis. Section 6.2, which discusses consensus peaksets, shows a number of Venn plots in context, and the help page for `dba.plotVenn` has a number of additional examples.

### 4.2 MA plots

MA plots are a useful way to visualize the effect of normalization on data, as well as seeing which of the datapoints are being identified as differentially bound. An MA plot can be obtained for the resistant-responsive contrast as follows:

```
> dba.plotMA(tamoxifen)
```

The plot is shown in Figure 3. Each point represents a binding site, with point in red representing sites identified as differentially bound. The plot shows how the differentially bound sites have an absolute log fold difference of at least 2. This same data can also be shown with the concentrations of each sample groups plotted against each other plot using `dba.plotMA(tamoxifen, bXY=T)`.

### 4.3 PCA plots

While the correlation heatmaps already seen are good for showing clustering, plots based on principal components analysis can be used to give a deeper insight into how samples are associated. A PCA plot corresponding to Figure 2a, which includes normalized read counts for all the binding sites, can be obtained as follows:

```
> dba.plotPCA(tamoxifen, DBA_CONDITION)
```

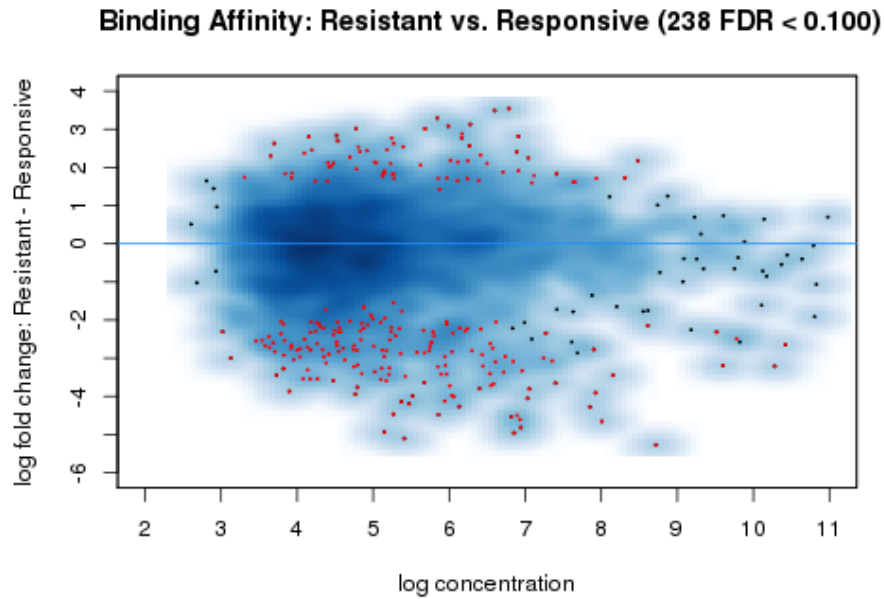


Figure 3: MA plot of Resistant-Responsive contrast, with sites identified as significantly differentially bound shown in red. Generated by: `dba.plotMA(tamoxifen)`

```

      Legend
Resistant  "black"
Responsive "red"

```

This draws the plot and returns a color legend. The resulting plot (Figure 4a) shows the four Resistant samples (black) not separable from the Responsive samples (red) in either the first (horizontal) or the second (vertical) components when looking at all the binding sites.

A PCA plot using only the differentially bound sites (corresponding to Figure 2b), using an FDR threshold of 0.05, can be drawn as follows:

```
> dba.plotPCA(tamoxifen, contrast=1, th=.05)
```

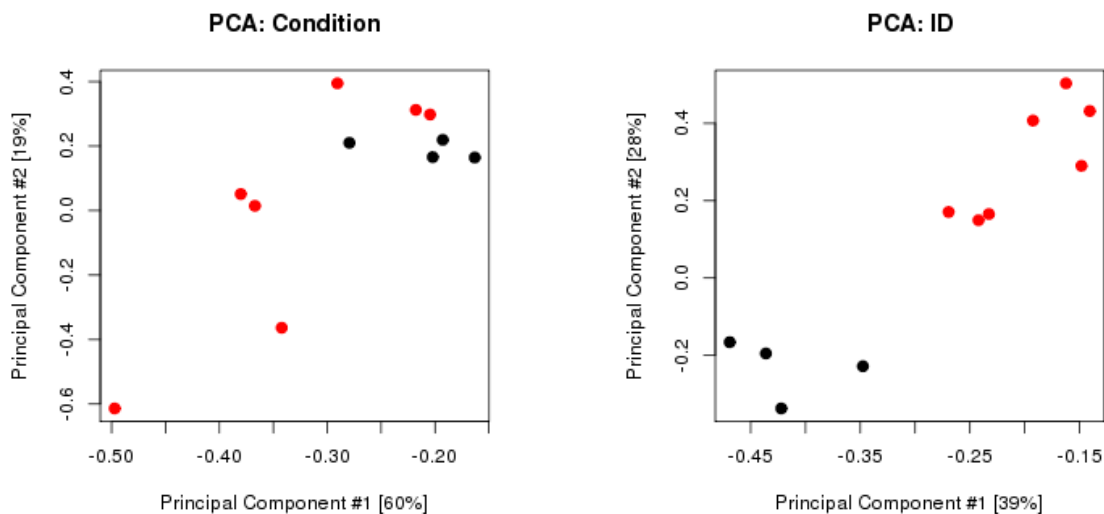
```

      Legend
Resistant  "black"
Responsive "red"

```

This plot (Figure 4b) shows that the differential analysis identifies sites than can be used to separate the sample groups along both the first and second components.

The `dba.plotPCA` function is customizable. For example, if you want to see where each of the unique cell lines lies, type `dba.plotPCA(tamoxifen, attributes=c(DBA_TISSUE, DBA_CONDITION))`. If your installation of R supports 3D graphics using the `rgl` package, try `dba.plotPCA(tamoxifen, b3D=T)`. Seeing the first three principal components can be a useful exploratory exercise.



(a) PCA plot using affinity data for all sites.  
Generated by: `dba.plotPCA(tamoxifen)`

(b) PCA plot using affinity data for only differentially bound sites. Generated by:  
`dba.plotPCA(tamoxifen, contrast=1)`

Figure 4: Principal Component Analysis (PCA) plots, generated using `dba.plotPCA`.

## 4.4 Boxplots

Boxplots provide a way to view how read distributions differ between classes of binding sites. Consider the example, where the 238 differentially bound sites are identified. The MA plot (Figure 3) shows that these are not distributed evenly between those that increase binding affinity in the Responsive group vs. those that increase binding affinity in the Resistant groups. This can be seen quantitatively using the sites returned in the report:

```
> sum(tamoxifen.DB$Fold<0)
```

```
[1] 167
```

```
> sum(tamoxifen.DB$Fold>0)
```

```
[1] 71
```

But how are reads distributed amongst the different classes of differentially bound sites and sample groups? These data can be more clearly seen using a boxplot:

```
> pvals = dba.plotBox(tamoxifen)
```

The default plot (Figure 5) shows in the first two boxes that amongst differentially bound sites overall, the Responsive samples have a somewhat higher mean read concentration. The next two

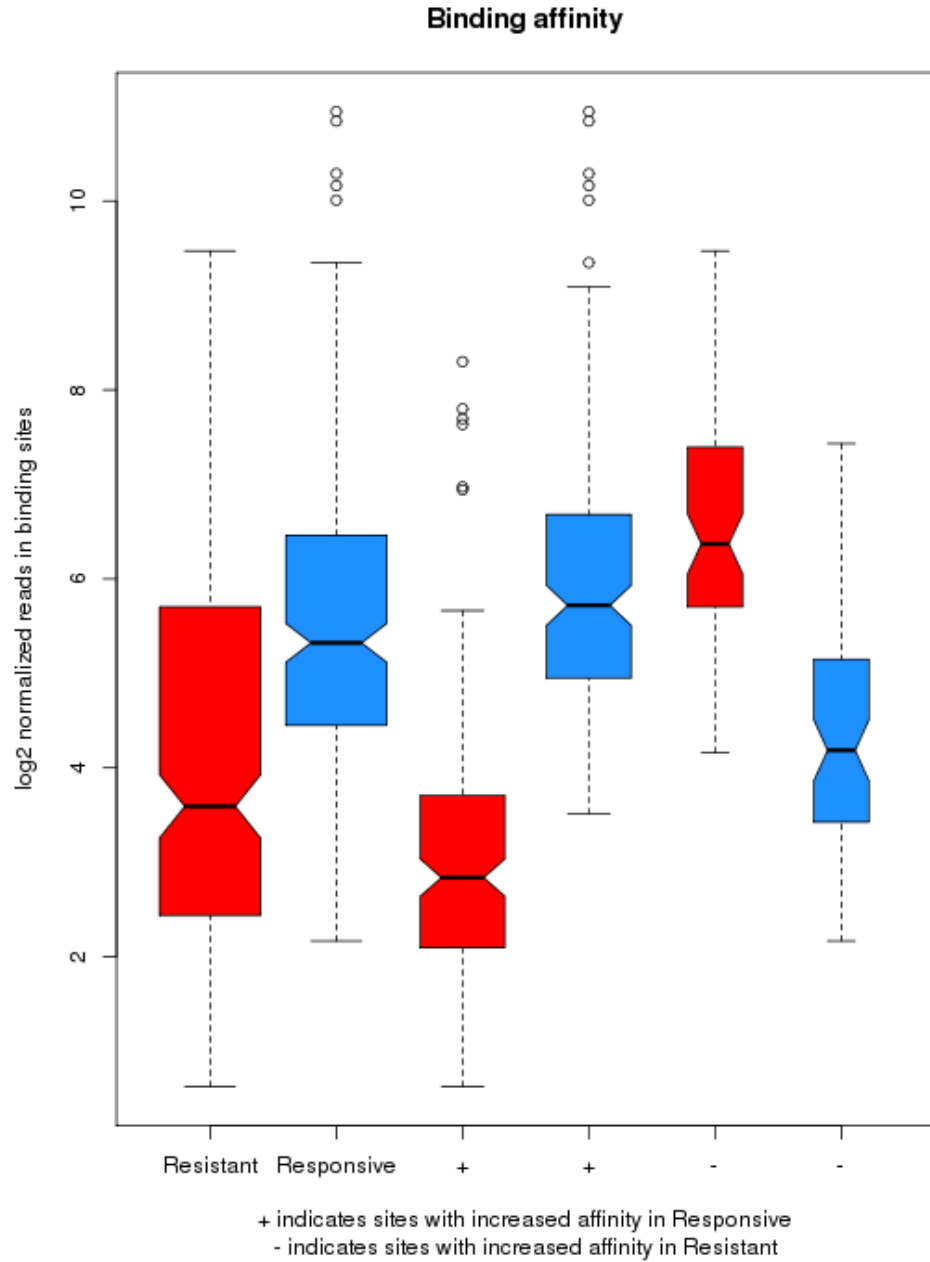


Figure 5: Box plots of read distributions for significantly differentially bound (DB) sites. Tamoxifen resistant samples are shown in red, and responsive samples are shown in blue. Left two boxes show distribution of reads over all DB sites in the Resistant and Responsive groups; middle two boxes show distributions of reads in DB sites that increase in affinity in the Responsive group; last two boxes show distributions of reads in DB sites that increase in affinity in the Resistant group. Generated by: `dba.plotBox(tamoxifen)`

boxes show the distribution of reads in differentially bound sites that exhibit increased affinity in the Responsive samples, while the final two boxes show the distribution of reads in differentially bound sites that exhibit increased affinity in the Resistant samples.

`dba.plotBox` returns a matrix of p-values (computed using a two-sided Wilcoxon ‘Mann-Whitney’ test, paired where appropriate) indicating which of these distributions are significantly different from another distribution.

```
> pvals
```

	Resistant.DB	Responsive.DB	Resistant.DB+	Responsive.DB+
Resistant.DB	1.000000e+00	3.708203e-18	1.837432e-06	3.728625e-20
Responsive.DB	3.708203e-18	1.000000e+00	1.204767e-42	1.444915e-03
Resistant.DB+	1.837432e-06	1.204767e-42	1.000000e+00	3.750193e-29
Responsive.DB+	3.728625e-20	1.444915e-03	3.750193e-29	1.000000e+00
Resistant.DB-	5.620287e-17	8.499579e-09	5.036766e-30	1.940145e-04
Responsive.DB-	3.364967e-02	2.257211e-08	2.741703e-12	2.926444e-14
	Resistant.DB-	Responsive.DB-		
Resistant.DB	5.620287e-17	3.364967e-02		
Responsive.DB	8.499579e-09	2.257211e-08		
Resistant.DB+	5.036766e-30	2.741703e-12		
Responsive.DB+	1.940145e-04	2.926444e-14		
Resistant.DB-	1.000000e+00	2.482200e-13		
Responsive.DB-	2.482200e-13	1.000000e+00		

The significance of the overall difference in distribution of concentrations amongst the differentially bound sites in the two groups is shown to be p-value=8.011987e-13, while those between the Resistant and Responsive groups in the individual cases (increased in Responsive or Resistant) have p-values computed as 3.750193e-29 and 2.482200e-13.

## 4.5 Heatmaps

DiffBind provides two types of heatmaps. This first, correlation heatmaps, we have already seen. For example, the heatmap shown in Figure 2a can be generated as follows:

```
> corvals = dba.plotHeatmap(tamoxifen)
```

The effect of different scoring methods (normalization) can be examined in these plots by setting the `score` parameter to a different value. The default value, `DBA_SCORE_TMM_MINUS_FULL`, uses the TMM normalization procedure from `edgeR`, with control reads subtracted first and using the full library size (total reads in library). Another scoring method is to use RPKM fold (RPKM of the ChIP reads divided by RPKM of the control reads; a correlation heatmap for all the data using this scoring method can be obtained by typing `dba.plotHeatmap(tamoxifen, score=DBA_SCORE_RPKM_FOLD)`).

Another way to view the patterns of binding affinity directly in the differentially bound sites is via a binding affinity heatmap. This can be plotted for the example case as follows:

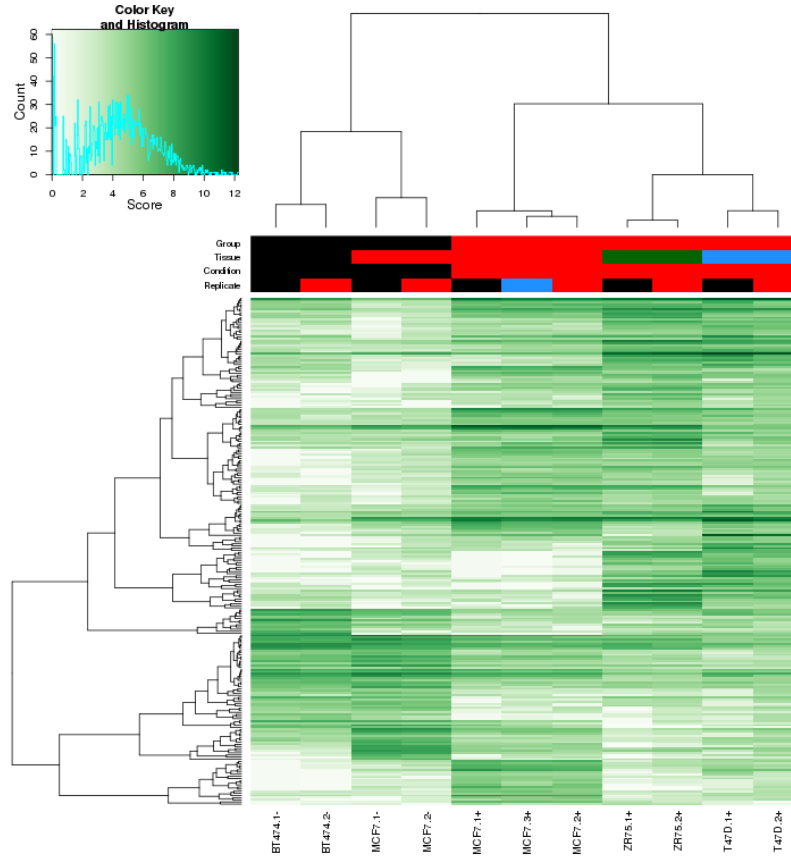


Figure 6: Binding affinity heatmap showing affinities for differentially bound sites. Samples cluster first by whether they are responsive to tamoxifen treatment, then by cell line. Clusters of binding sites show distinct patterns of affinity levels. Generated by: `dba.plotHeatmap(tamoxifen, contrast=1, correlations=FALSE)`

```
> corvals = dba.plotHeatmap(tamoxifen, contrast=1, correlations=FALSE)
```

Figure 6 shows the affinities and clustering of the differentially bound sites (rows), as well as the sample clustering (columns). This plot can be tweaked to get more contrast, for example by using row-scaling `dba.plotHeatmap(tamoxifen, contrast=1, correlations=FALSE, scale=row)`.

## 5 Example: differential binding analysis using a blocking factor

The previous example showed how to perform a differential binding analysis using a single factor with two values; that is, finding the significantly differentially bound sites between two sets of samples. This section extends the example by including a second factor, potentially with multiple



values, that represents a confounding condition. Examples of experiments where it is appropriate to use a blocking factor include ones where there are potential batch effects, with samples from the two conditions prepared together, or a matched design (e.g. matched normal and tumor pairs, where the primary factor of interest is to discover sites consistently differentially bound between normal and tumor samples. In the current example, the confounding effect we want to control for is the presence of two sets of samples, one tamoxifen responsive and one resistant, that are both derived from the same MCF7 cell line.

In the previous analysis, the two MCF7-derived cell lines tended to cluster together. While the differential binding analysis was able to identify sites that could be used to separate the resistance from the responsive samples, the confounding effect of the common ancestry could still be seen even when considering only the significantly differentially bound sites (Figure 2a).

Using the generalized linear modelling (GLM) functionality included in `edgeR` and `DESeq`, the confounding factor can be explicitly modelled. This is done by specifying a blocking factor to `dba.contrast`. There are a number of ways to specify this factor. If it is encapsulated in a piece of metadata (eg. `DBA_REPLICATE`, or `DBA_TREATMENT` etc.), simply specifying the metadata field is sufficient. In the current case, there is no specific metadata field that captures the factor we want to block (although an unused metadata field, such as `DBA_TREATMENT`, could be used to specify this factor). An alternate way of specifying the confounded sampled is to use a mask:

```
> data(tamoxifen_counts)
> tamoxifen = dba.contrast(tamoxifen, categories=DBA_CONDITION,
+                          block=tamoxifen$masks$MCF7)
```

Now when the analysis is run, it will be run using both the single-factor comparison as well as fitting a linear model with the second, blocking factor, for comparison:

```
> tamoxifen = dba.analyze(tamoxifen)
> tamoxifen
```

11 Samples, 1654 sites in matrix:

	ID	Tissue	Factor	Condition	Replicate	Caller	Intervals	SN
1	BT474.1-	BT474	ER	Resistant	1	counts	1654	0.19
2	BT474.2-	BT474	ER	Resistant	2	counts	1654	0.18
3	MCF7.1+	MCF7	ER	Responsive	1	counts	1654	0.35
4	MCF7.2+	MCF7	ER	Responsive	2	counts	1654	0.22
5	MCF7.3+	MCF7	ER	Responsive	3	counts	1654	0.28
6	T47D.1+	T47D	ER	Responsive	1	counts	1654	0.14
7	T47D.2+	T47D	ER	Responsive	2	counts	1654	0.08
8	MCF7.1-	MCF7	ER	Resistant	1	counts	1654	0.25
9	MCF7.2-	MCF7	ER	Resistant	2	counts	1654	0.16
10	ZR75.1+	ZR75	ER	Responsive	1	counts	1654	0.35
11	ZR75.2+	ZR75	ER	Responsive	2	counts	1654	0.24

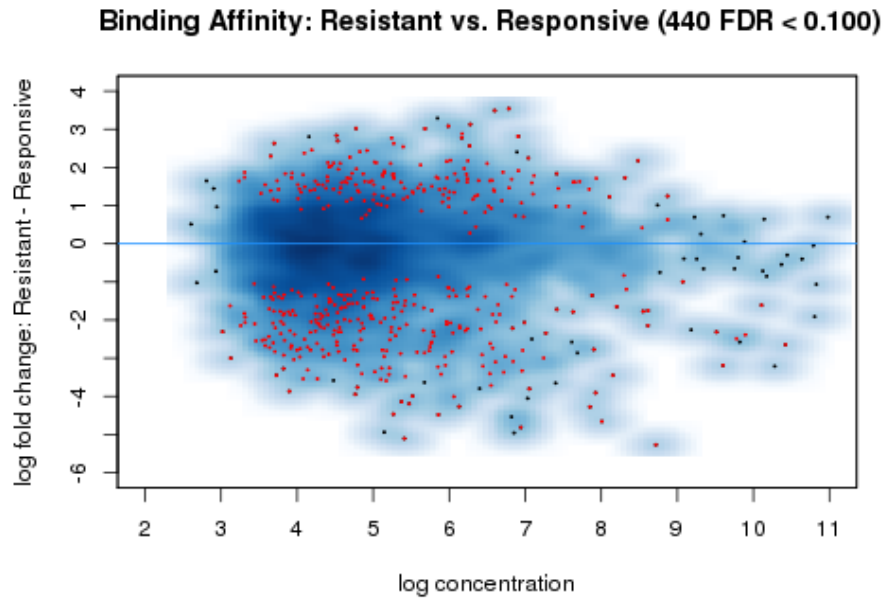


Figure 7: MA plot of Resistant-Responsive contrast, using MCF7 origin as a blocking factor, with sites identified as significantly differentially bound shown in red. Generated by: `dba.plotMA(tamoxifen, method=DBA_EDGER_BLOCK)`

1 Contrast:

	Group1	Members1	Group2	Members2	Block1Val	InBlock1	Block2Val	InBlock2
1	Resistant	4	Responsive	7	true	5	false	6
	DB.edgeR	DB.edgeR.block						
1	238		440					

This indicates that where the standard, single-factor `edgeR` analysis identifies 238 differentially bound sites, the analysis using the blocking factor finds 440 such sites. An MA plot shows how the analysis has changed:

```
> dba.plotMA(tamoxifen, method=DBA_EDGER_BLOCK)
```

The resulting plot is shown in Figure 7. Comparing this to Figure 3, at least two differences can be observed. The analysis has become more sensitive, with sites being identified as significantly differentially bound with lower magnitude fold changes (as low as twofold, as this plot is on a log2 scale). But it is not merely lowering a fold threshold: some sites with higher fold changes are no longer found to be significant. These were identified as significantly differentially bound in the earlier analysis because the confounding factor was not being modelled.

Consider the resulting separation and clustering using the newly discovered differentially bound sites:

```
> dba.plotHeatmap(tamoxifen,contrast=1,method=DBA_EDGER_BLOCK,
+                 attributes=c(DBA_TISSUE,DBA_CONDITION,DBA_REPLICATE))
```

```
> dba.plotPCA(tamoxifen,contrast=1,method=DBA_EDGER_BLOCK,
+             attributes=c(DBA_TISSUE,DBA_CONDITION))
```

```

Legend
BT474:Resistant "black"
MCF7:Resistant "red"
MCF7:Responsive "dodgerblue"
T47D:Responsive "darkgreen"
ZR75:Responsive "cyan"
```

Frequently, as more sites are included in these plots, the result is often worse clustering/separation along the grouping of primary interest. As the correlations used for the heatmap and the count scores used for the PCA plots do not take the blocking factor into account, the separation seen in these plots will not necessarily sharpen after completing a successful block analysis. Compare the resulting correlation heatmap (Figure 8 to Figure 2b). The PCA plot (Figure 9, compared to Figure 4b), shows less distance between the two sample groups along the first component, as more binding sites common to the two MCF7 cell lines are able to be identified as being significantly differentially bound between the Resistant and Responsive sample groups.

It is also interesting to compare the performance of `edgeR` with that of `DESeq` and `DESeq2` on this dataset:

```
> tamoxifen = dba.analyze(tamoxifen,method=DBA_ALL_METHODS)
```

```
.
.
.
.
```

```
> tamoxifen
```

```
11 Samples, 1654 sites in matrix:
```

	ID	Tissue	Factor	Condition	Replicate	Caller	Intervals	SN
1	BT474.1-	BT474	ER	Resistant	1	counts	1654	0.19
2	BT474.2-	BT474	ER	Resistant	2	counts	1654	0.18
3	MCF7.1+	MCF7	ER	Responsive	1	counts	1654	0.35
4	MCF7.2+	MCF7	ER	Responsive	2	counts	1654	0.22
5	MCF7.3+	MCF7	ER	Responsive	3	counts	1654	0.28
6	T47D.1+	T47D	ER	Responsive	1	counts	1654	0.14
7	T47D.2+	T47D	ER	Responsive	2	counts	1654	0.08
8	MCF7.1-	MCF7	ER	Resistant	1	counts	1654	0.25

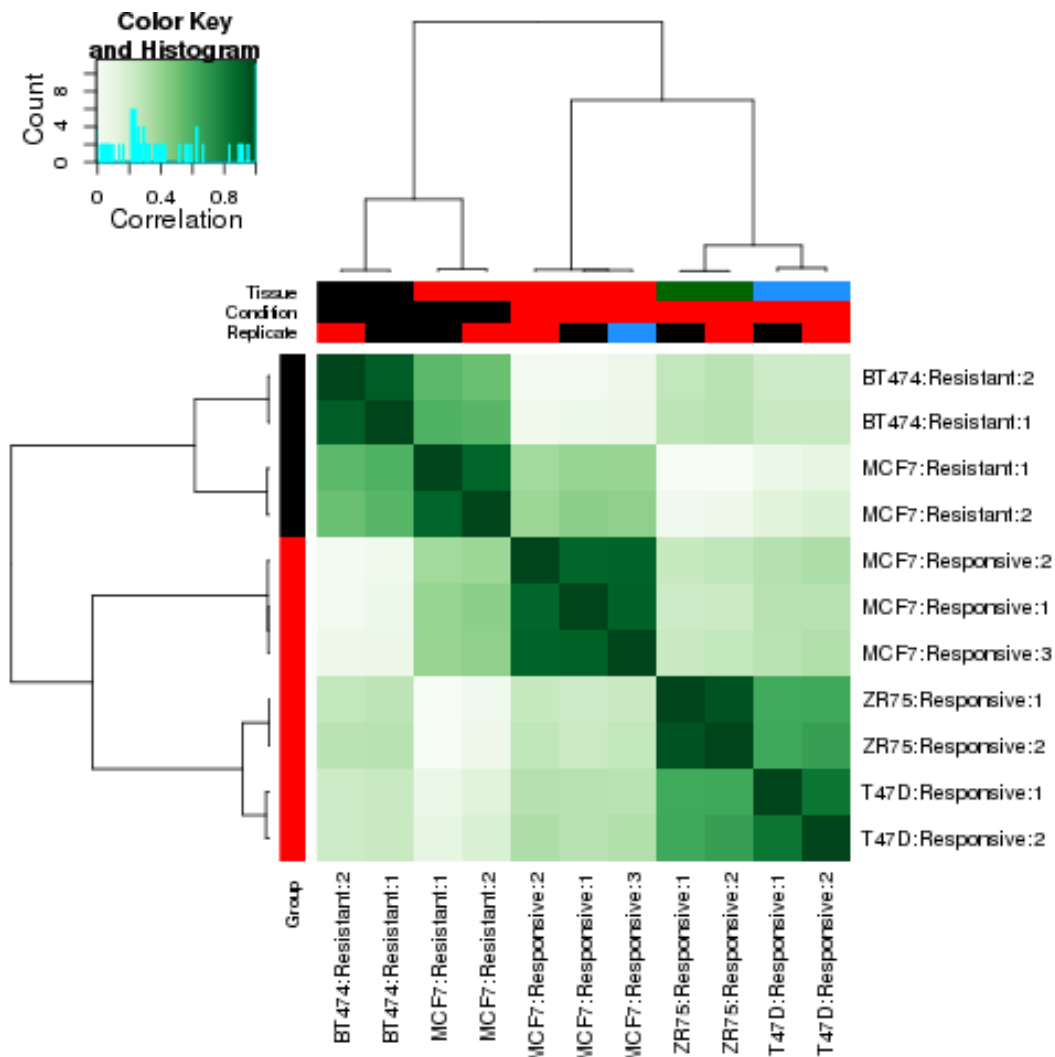


Figure 8: Correlation heatmap of using scores for significantly differentially bound sites for the Resistant-Responsive contrast, using MCF7 origin as a blocking factor. Generated by: `dba.plotHeatmap(tamoxifen, contrast=1, method=DBA_EDGER_BLOCK, attributes=c(DBA_TISSUE,DBA_CONDITION,DBA_REPLICATE))`

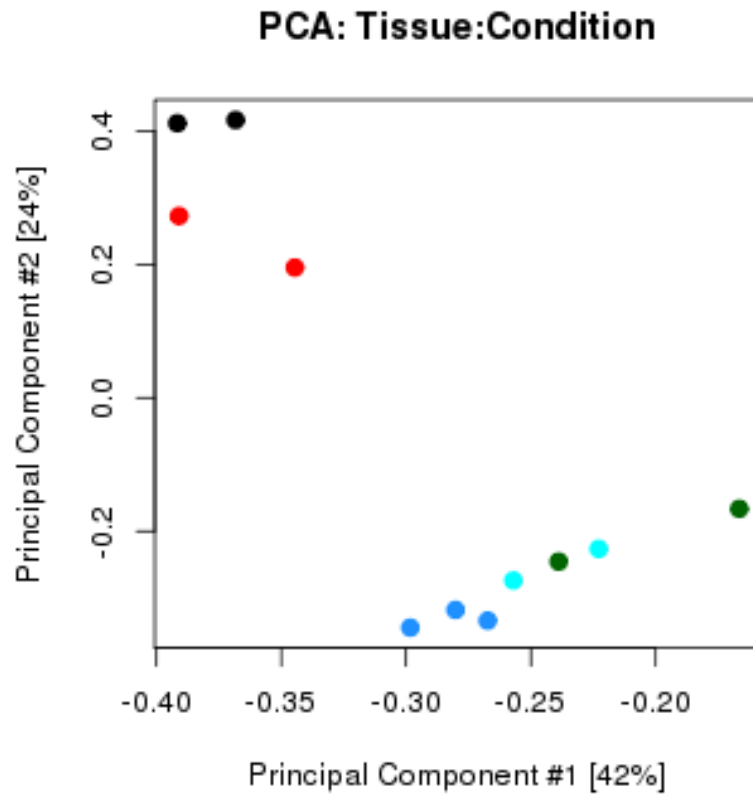


Figure 9: Plot of first two principal components, using scores for significantly differentially bound sites for the Resistant-Responsive contrast, using MCF7 origin as a blocking factor. Resistant samples in black and red; MCF7 samples in red and blue. Generated by: `dba.plotPCA(tamoxifen,contrast=1,method=DBA_EDGER_BLOCK, attributes=c(DBA_TISSUE,DBA_CONDITION))`

9	MCF7.2-	MCF7	ER Resistant	2 counts	1654	0.16
10	ZR75.1+	ZR75	ER Responsive	1 counts	1654	0.35
11	ZR75.2+	ZR75	ER Responsive	2 counts	1654	0.24

1 Contrast:

	Group1	Members1	Group2	Members2	Block1Val	InBlock1	Block2Val	InBlock2
1	Resistant	4	Responsive	7	true	5	false	6
	DB.edgeR	DB.edgeR.block	DB.DESeq	DB.DESeq.block	DB.DESeq2	DB.DESeq2.block		
1	238	440	10	455	198	409		

DESeq has a much more conservative approach to the single factor analysis, identifying only 10 sites as differentially bound, while DESeq2 identifies a similar number of sites as edgeR. When modelling the confounding factor, the greater sensitivity results in many more sites being identified using all three packages. You can check this by looking at the identified sites using `dba.report`, and performing MA, heatmap, and PCA plots.

We can also compare the sites identified using edgeR, DESeq, and DESeq2. An easy way to do this is to use a special feature of the `dba.report` function:

```
> tam.block = dba.report(tamoxifen,method=DBA_ALL_BLOCK,bDB=TRUE,bAll=TRUE)
> tam.block
```

3 Samples, 420 sites in matrix (536 total):

	ID	Tissue	Factor	Condition	Treatment	Intervals
1	Resistant_vs_Responsive	all	DB	edgeR	block	440
2	Resistant_vs_Responsive	all	DB	DESeq	block	455
3	Resistant_vs_Responsive	all	DB	DESeq2	block	409

```
> dba.plotVenn(tam.block,1:3,label1="edgeR",label2="DESeq",label3="DESeq2")
```

The overlap is shown in Figure 10. Agreement is quite good, with DESeq2 identifying the fewest sites not identified by at least one other method.

## 6 Example: occupancy analysis and overlaps

In this section, we look at the tamoxifen resistance ER-binding dataset in some more detail, showing what a pure occupancy-based analysis would look like, and comparing it to the results obtained using the affinity data. For this we will start by re-loading the peaksets:

```
> data(tamoxifen_peaks)
```

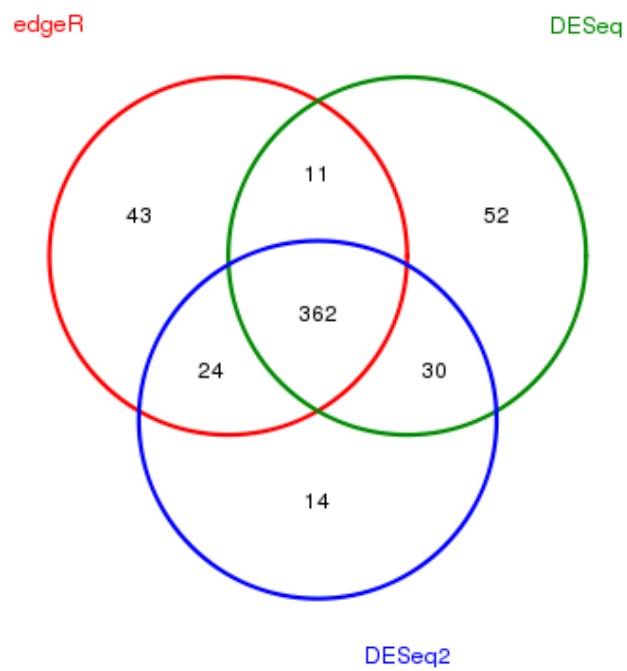


Figure 10: Venn diagram showing overlap of differentially bound peaks identified using `edgeR`, `DESeq`, and `DESeq2` to do a multi-factor (blocking) analysis. Generated by plotting the result of: `dba.plotVenn(tam.block,1:3, label1="edgeR",label2="DESeq",label3="DESeq2")`

## 6.1 Overlap rates

One reason to do an occupancy-based analysis is to determine what candidate sites should be used in a subsequent affinity-based analysis. In the example so far, we took all sites that were identified in peaks in at least three of the eleven peaksets, reducing the number of sites from 3557 overall to the 1654 sites used in the differential analysis. We could have used a more stringent criterion, such as only taking sites identified in five or six of the peakset, or a less stringent one, such as including all 3557 sites. In making the decision of what criteria to use many factors come into play, but it helps to get an idea of the rates at which the peaksets overlap (for more details on how overlaps are determined, see Section 7.2 on peak merging). A global overview can be obtained using the RATE mode of the `dba.overlap` function as follows:

```
> olap.rate = dba.overlap(tamoxifen,mode=DBA_OLAP_RATE)
> olap.rate
```

```
[1] 3557 2602 1654 1299 1002 764 620 455 352 187 118
```

`olap.rate` is a vector containing the number of peaks that appear in at least one, two, three, and so on up to all eleven peaksets.

These values can be plotted to show the overlap rate drop-off curve:

```
> plot(olap.rate,type='b',ylab='# peaks', xlab='Overlap at least this many peaksets')
```

The rate plot is shown in Figure 11. These curves typically exhibit a roughly geometric drop-off, with the number of overlapping sites halving as the overlap criterion becomes stricter by one site. When the drop-off is extremely steep, this is an indication that the peaksets do not agree very well. For example, if there are replicates you expect to agree, there may be a problem with the experiment. In the current example, peak agreement is high and the curve exhibits a better than geometric drop-off.

## 6.2 Deriving consensus peaksets

When performing an overlap analysis, it is often the case that the overlap criteria are set stringently in order to lower noise and drive down false positives.<sup>3</sup> The presence of a peak in multiple peaksets is an indication that it is a "real" binding site, in the sense of being identifiable in a repeatable manner. The use of biological replicates (performing the ChIP multiple times), as in the tamoxifen dataset, can be used to guide derivation of a consensus peakset. Alternatively, an inexpensive but less powerful way to help accomplish this is to use multiple peak callers for each ChIP dataset and look for agreement between peak callers (Li et al. [2011]).

Consider for example the standard (tamoxifen responsive) MCF7 cell line, represented by three replicates in this dataset. How well do the replicates agree on their peak calls? The overlap

---

<sup>3</sup>It is less clear that limiting the potential binding sites in this way is appropriate when focusing on affinity data, as the differential binding analysis method will identify only sites that are significantly differentially bound, even if operating on peaksets that include incorrectly identified sites.



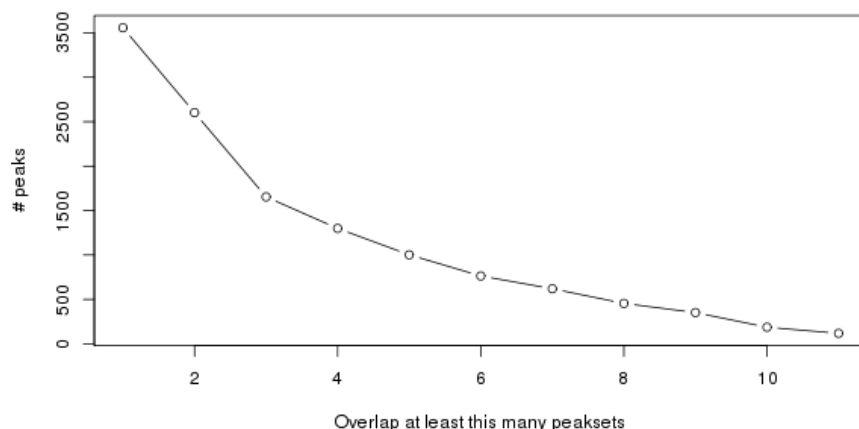


Figure 11: Overlap rate plot, showing how the number of overlapping peaks decreases as the overlap criteria becomes more stringent. X axis shows the number of peaksets in which the site is identified, while the Y axis shows the number of overlapping sites. Generated by plotting the result of: `dba.overlap(tamoxifen,mode=DBA_OLAP_RATE)`

rate for just the MCF7 samples can be isolated using a *sample mask*. A set of sample masks are automatically associated with a *DBA* object in the `$masks` field:

```
> names(tamoxifen$masks)

[1] "BT474"      "MCF7"      "T47D"      "ZR75"      "ER"
[6] "Resistant"  "Responsive" ""           "raw"       "Replicate.1"
[11] "Replicate.2" "Replicate.3" "All"      "None"
```

Arbitrary masks can be generated using the `dba.mask` function, or simply by specifying a vector of peakset numbers. In this case, a mask that isolates the MCF7 samples can be generated by combining to pre-defined masks (MCF7 and Responsive) and passed into the `dba.overlap` function:

```
> dba.overlap(tamoxifen,tamoxifen$masks$MCF7 & tamoxifen$masks$Responsive,
+             mode=DBA_OLAP_RATE)

[1] 1767 1222 874
```

There are 874 peaks (out of 1767) identified in all three replicates. A finer grained view of the overlaps can be obtained with the `dba.plotVenn` function:

```
> dba.plotVenn(tamoxifen, tamoxifen$masks$MCF7 & tamoxifen$masks$Responsive)
```

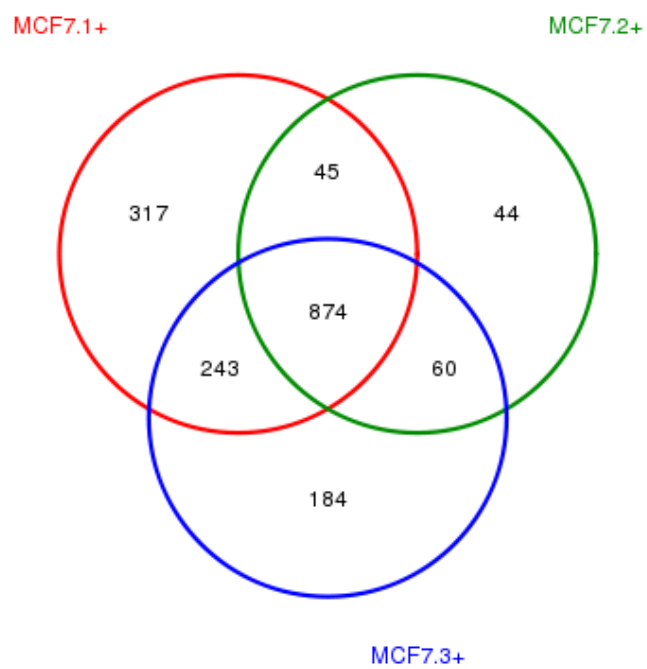


Figure 12: Venn diagram showing how the ER peak calls for three replicates of responsive MCF7 cell line overlap. Generated by plotting the result of: `dba.venn(tamoxifen,tamoxifen$mask$MCF7 & tamoxifen$mask$Responsive)`

The resultant plot is shown as Figure 12. This plot shows the 874 consensus peaks identified as common to all replicates, but further breaks down how the replicates relate to each other. The same can be done for each of the replicated cell line experiments, and rather than applying a global cutoff (3 of 11), each cell line could be dealt with individually in deriving a final peakset. A separate consensus peakset for each of the replicated sample types can be added to the DBA object using `dba.peakset`:

```
> tamoxifen = dba.peakset(tamoxifen, consensus = c(DBA_TISSUE,DBA_CONDITION),
+                         minOverlap=0.66)
> tamoxifen
```

16 Samples, 2602 sites in matrix (3557 total):

	ID	Tissue	Factor	Condition	Replicate	Caller	Intervals
1	BT474.1-	BT474	ER	Resistant	1	raw	1084
2	BT474.2-	BT474	ER	Resistant	2	raw	1115
3	MCF7.1+	MCF7	ER	Responsive	1	raw	1513
4	MCF7.2+	MCF7	ER	Responsive	2	raw	1037
5	MCF7.3+	MCF7	ER	Responsive	3	raw	1372
6	T47D.1+	T47D	ER	Responsive	1	raw	509
7	T47D.2+	T47D	ER	Responsive	2	raw	347
8	MCF7.1-	MCF7	ER	Resistant	1	raw	1148
9	MCF7.2-	MCF7	ER	Resistant	2	raw	933
10	ZR75.1+	ZR75	ER	Responsive	1	raw	2111
11	ZR75.2+	ZR75	ER	Responsive	2	raw	1975
12	BT474:Resistant	BT474	ER	Resistant	1-2	raw	902
13	MCF7:Responsive	MCF7	ER	Responsive	1-2-3	raw	1222
14	T47D:Responsive	T47D	ER	Responsive	1-2	raw	298
15	MCF7:Resistant	MCF7	ER	Resistant	1-2	raw	795
16	ZR75:Responsive	ZR75	ER	Responsive	1-2	raw	1633

This adds a new consensus peakset for each set of samples that share the same Tissue and Condition values. The exact effect could be obtained by calling `tamoxifen = dba.peakset(tamoxifen, consensus = -DBA_REPLICATE)` on the original set of peaks; this tells DiffBind to generate a consensus peakset for every set of samples that have identical metadata values *except* the Replicate number.

From this, a new DBA object can be generated consisting of only the five consensus peaksets (the `$Consensus` mask filters peaksets previously formed using `dba.peakset`) :

```
> tamoxifen_consensus = dba(tamoxifen, mask = tamoxifen$masks$Consensus)
> tamoxifen_consensus
```

5 Samples, 1163 sites in matrix (2444 total):

	ID	Tissue	Factor	Condition	Replicate	Caller	Intervals
--	----	--------	--------	-----------	-----------	--------	-----------

1	BT474:Resistant	BT474	ER Resistant	1-2	raw	902
2	MCF7:Responsive	MCF7	ER Responsive	1-2-3	raw	1222
3	T47D:Responsive	T47D	ER Responsive	1-2	raw	298
4	MCF7:Resistant	MCF7	ER Resistant	1-2	raw	795
5	ZR75:Responsive	ZR75	ER Responsive	1-2	raw	1633

Alternatively, a master consensus peakset could be generated, and reads counted, directly using `dba.count`: `tamoxifen = dba.count(tamoxifen, peaks=tamoxifen$mask$Consensus)`

Finally, consider an analysis where we wished to treat all five MCF7 samples together to look for binding sites specific to that cell line irrespective of tamoxifen resistant/responsive status. We can create consensus peaksets for each cell type, and look at how the resultant peaks overlap (Figure 13):

```
> data(tamoxifen_peaks)
> tamoxifen = dba.peakset(tamoxifen, consensus = DBA_TISSUE, minOverlap=0.66)
> dba.plotVenn(tamoxifen, tamoxifen$mask$Consensus)
```

### 6.3 A complete occupancy analysis: identifying sites unique to a sample group

Occupancy-based analysis, in addition to offering many ways of deriving consensus peaksets, can also be used to identify sites unique to a group of samples. This is analogous to, but not the same as, finding differentially bound sites. In these subsections, the two approaches are directly compared.

Returning to the original tamoxifen dataset:

```
> data(tamoxifen_peaks)
```

We can derive consensus peaksets for the Resistant and Responsive groups. First we examine the overlap rates:

```
> dba.overlap(tamoxifen, tamoxifen$mask$Resistant, mode=DBA_OLAP_RATE)

[1] 1875 1298 597 436

> dba.overlap(tamoxifen, tamoxifen$mask$Responsive, mode=DBA_OLAP_RATE)

[1] 3208 2293 1217 807 584 265 161
```

Requiring that consensus peaks overlap in at least one third of the samples in each group results in 1298 sites for the Resistant group and 1217 sites for the Responsive group:

```
> tamoxifen = dba.peakset(tamoxifen, consensus = DBA_CONDITION, minOverlap = 0.33)
> dba.plotVenn(tamoxifen, tamoxifen$mask$Consensus)
```

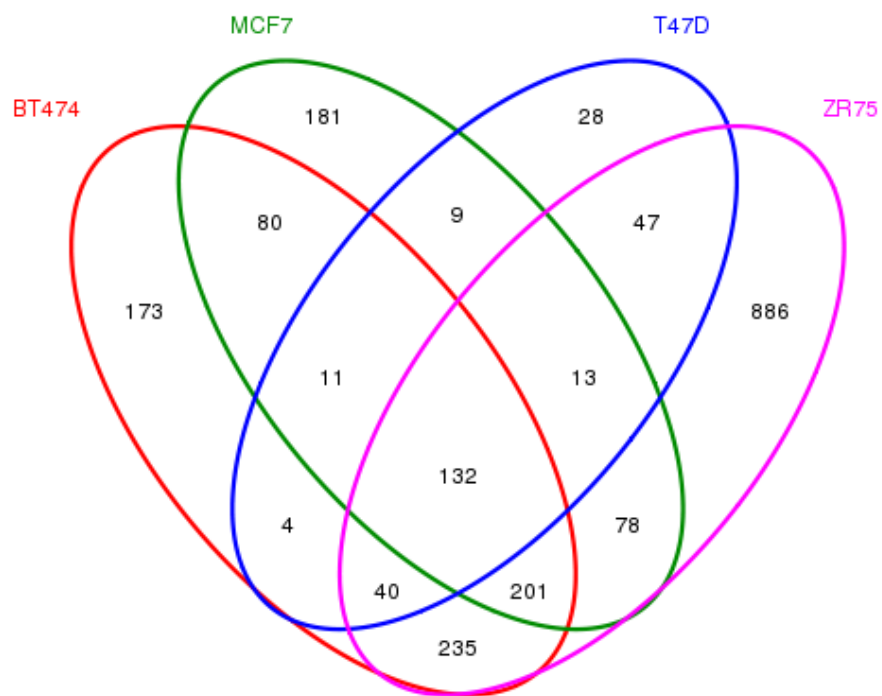


Figure 13: Venn diagram showing how the consensus peaks for each cell type overlap. Generated by plotting the result of: `dba.venn(tamoxifen,tamoxifen$mask$Consensus)`

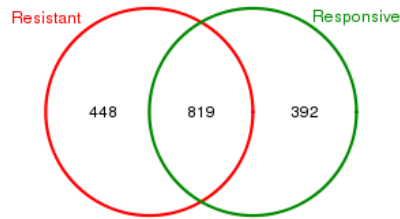


Figure 14: Venn diagram showing how the ER peak calls for two response groups overlap. Generated by plotting the result of: `dba.plotVenn(tamoxifen, tamoxifen$mask$Consensus)`

Figure 14 shows that 448 sites are unique to the Resistant group, and 392 sites are unique to the Responsive group, with 819 sites being identified in both groups (meaning in at least half the Resistant samples and at least three of the seven Responsive samples). If our primary interest is in finding binding sites that are different between the two groups, it may seem reasonable to consider the 819 common sites to be uninteresting, and focus on the 840 sites that are unique to a specific group. These unique sites can be obtained using `dba.overlap`:

```
> tamoxifen.OL = dba.overlap(tamoxifen, tamoxifen$mask$Consensus)
```

The sites unique to the Resistant group are accessible in `tamoxifen.OL$onlyA`, with the Responsive-unique sites in `tamoxifen.OL$onlyB`:

```
> tamoxifen.OL$onlyA
```

GRanges with 448 ranges and 1 metadata column:

	seqnames	ranges	strand	score
	<Rle>	<IRanges>	<Rle>	<numeric>
4	chr18	[301531, 302172]	*	0.128978267622691
6	chr18	[346557, 347362]	*	0.0254536579689849
7	chr18	[361121, 362106]	*	0.0178065034135036
8	chr18	[384853, 386517]	*	0.0887190625790319
11	chr18	[479200, 480032]	*	0.086873645309346
...	...	...	...	...
1649	chr18	[75157775, 75158504]	*	0.0346388152202919
1650	chr18	[75163026, 75163816]	*	0.0306614720699587

```

1652    chr18 [75401417, 75402162]    * | 0.0494002153710507
1653    chr18 [75525519, 75526188]    * | 0.064589591798652
1657    chr18 [75826088, 75826939]    * | 0.0252808318204148
---
seqlengths:
  chr18
    NA

> tamoxifen.OL$onlyB

GRanges with 392 ranges and 1 metadata column:
      seqnames      ranges strand |      score
      <Rle>        <IRanges> <Rle> | <numeric>
    5    chr18    [ 336592,  337347]    * | 0.0466440809907735
   10    chr18    [ 439109,  440079]    * | 0.0331046835979069
   27    chr18    [ 988767,  989698]    * | 0.047010413661554
   28    chr18   [1065304, 1066051]    * | 0.0695151519760949
   33    chr18   [1231653, 1232311]    * | 0.0552311386757906
   ...    ...          ...      ...  ...
1645    chr18 [74850775, 74851581]    * | 0.0736038759005841
1646    chr18 [74860617, 74861878]    * | 0.056640563146553
1647    chr18 [74906349, 74907306]    * | 0.0356814694433001
1655    chr18 [75641971, 75642647]    * | 0.061984513945994
1659    chr18 [76087923, 76089259]    * | 0.125980115522818
---
seqlengths:
  chr18
    NA

```

The scores associated with each site are derived from the peak caller confidence score, and are a measure of confidence in the peak call (occupancy), not a measure of how strong or distinct the peak is.

## 6.4 Comparison of occupancy and affinity based analyses

So how does this occupancy-based analysis compare to the previous affinity-based analysis?

First, different criteria were used to select the overall consensus peakset. We can compare them to see how well they agree:

```

> tamoxifen = dba.peakset(tamoxifen,tamoxifen$mask$Consensus,
+                         minOverlap=1,sampID="OL Consensus")
> tamoxifen = dba.peakset(tamoxifen,!tamoxifen$mask$Consensus,
+                         minOverlap=3,sampID="Consensus_3")
> dba.plotVenn(tamoxifen,14:15)

```



Figure 15: Venn diagram showing how the ER peak calls for two different ways of deriving consensus peaksets. Generated by plotting the result of: `dba.plotVenn(tamoxifen,14:15)`

Figure 15 shows that the two sets agree on about 85% of their sites, so the results should be directly comparable between the differing parameters used to establish the consensus peaksets.<sup>4</sup>

Next re-load the affinity analysis:

```
> data(tamoxifen_analysis)
```

To compare the sites unique to each sample group identified from the occupancy analysis with those sites identified as differentially bound based on affinity (read count) data, we use a feature of `dba.report` that facilitates evaluating the occupancy status of sites. Here we obtain a report of all the sites (`th=1`) with occupancy statistics (`bCalled=T`):

```
> tamoxifen.rep = dba.report(tamoxifen,bCalled=T,th=1)
```

The `bCalled` option adds two columns to the report (`Called1` and `Called2`), one for each group, giving the number of samples within the group in which the site was identified as a peak in the original peaksets generated by the peak caller. We can use these to recreate the overlap criteria used in the occupancy analysis:

```
> onlyResistant = tamoxifen.rep$Called1>=2 & tamoxifen.rep$Called2<3
> sum(onlyResistant )
```

```
[1] 313
```

---

<sup>4</sup>Alternatively, we could re-run the analysis using the newly derived consensus peakset by passing it into the counting function: `> tamoxifen = dba.count(tamoxifen, peaks = tamoxifen$masks$Consensus)`



```
> onlyResponsive = tamoxifen.rep$Called2>=3 & tamoxifen.rep$Called1<2
> sum(onlyResponsive)
```

```
[1] 391
```

```
> bothGroups = tamoxifen.rep$Called1>= 2 & tamoxifen.rep$Called2>=3
> sum(bothGroups)
```

```
[1] 821
```

Comparing these numbers verifies the similarity with those seen in Figure 14, showing again how the basic analysis is not oversensitive to differences in how the consensus peaksets are formed. This overlap analysis suggests that 704 of the sites are uniquely bound in either the Responsive or Resistant groups, while 821 sites are common to both.

Completing a full differential analysis and focusing on only those sites identified as significantly differentially bound ( $\text{FDR} \leq 0.1$ ), however, shows a different story than that obtainable using only occupancy data:

```
> tamoxifen.DB = dba.report(tamoxifen,bCalled=T,th=.1)
> onlyResistant.DB = tamoxifen.DB$Called1>=2 & tamoxifen.DB$Called2<3
> sum(onlyResistant.DB)
```

```
[1] 33
```

```
> onlyResponsive.DB = tamoxifen.DB$Called2>=3 & tamoxifen.DB$Called1<2
> sum(onlyResponsive.DB)
```

```
[1] 131
```

```
> bothGroups.DB = tamoxifen.DB$Called1>=2 & tamoxifen.DB$Called2>=3
> sum(bothGroups.DB)
```

```
[1] 71
```

There are a number of notable differences in the results. First, overall there are many fewer sites identified as differentially bound ( $33+131+71 = 238$ ) than are unique to one condition ( $313+391 = 704$ ). Indeed, most of the sites identified in the occupancy analysis as unique to a sample group are not found to be significantly differentially bound using the affinity data. While partly this is a result of the stringency of the statistical tests, it shows how the affinity analysis can discriminate between sites where peak callers are making occupancy decisions that do not reflect significant differences in read densities at these sites. Note that only about  $39152.3961661342\%$  of sites unique to one condition are identifiable as significantly differentially bound ( $33+131 = 164$  out of 704). Secondly, differentially bound sites are as likely to be called in the consensus of both response groups as they are to be unique to one group, as *nearly one third of the sites identified as significantly differentially*

*bound are called as peaks in both response groups.* Indeed, sites identified in both sample groups are almost as likely to be identified as significantly differentially bound (8.64799025578563%, or 71 out of 821) as sites identified only in the Resistant sample group (10.5431309904153%, or 33 out of 313). The final advantage of a quantitative analysis is that the differentially bound peaks identified using the affinity analysis are associated with significance statistics (p-value and FDR) that can be used to rank them for further examination, while the occupancy analysis yields a relatively unordered list of peaks, as the peak caller statistics refer only to the significance of occupancy, and not of differential binding.

## 7 Technical notes

This section includes some technical notes explaining some of the technical details of DiffBind processing.

### 7.1 Loading peaksets

There are a number of ways to get peaksets loaded into a DBA object. Peaksets can be read in from files or loaded from interval sets already stored in an R object. Samples can be specified either in a sample sheet (using `dba`) or loaded one at a time (using `dba.peakset`).

When loading in peaksets from files, specifying what peak caller generated the file enables peaks from supported peak callers to be read in. See the help page for `dba.peakset` for a list of supported peak callers. Any string to indicate the peak caller; if it is not one of the supported callers, a default "raw" format is assumed, consisting of a text file with three or four columns (indicating the chromosome, start position, and end position, with a score for each interval found in the fourth column, if present). You can further control how peaks are read using the `PeakFormat`, `ScoreCol`, and `bLowerBetter` fields if you want to override the defaults for the specified peak caller identifier. For example, with the tamoxifen dataset used in this tutorial, the peaks were called using the MACS peak caller, but the data are supplied as simple text files, not the expected MACS "xls" format. To maintain the peak caller in the metadata, we could specify the `PeakCaller` as "macs" but the `PeakFormat` as "raw". If we wanted to use peak scores in a different column than the fourth, the `scorecol` parameter could be set to indicate the appropriate column number. When handling scoring, DiffBind by default assumes that a higher score indicates a "better" peak. If this is not the case, for example if the score is a p-value or FDR, we could set `bLowerScoreBetter` to `TRUE`.

When using a sample sheet, values for fields missing in the sample sheet can be supplied when calling `dba`. In addition to the minimal sample sheet used for the tutorial, an equivalent sample sheet with all the metadata fields is included, called "tamoxifen\_allfields.csv". See the help page for `dba` for an example using this sample sheet.

## 7.2 Merging peaks

When forming the global binding matrix consensus peaksets, `DiffBind` first identifies all unique peaks amongst the relevant peaksets. As part of this process, it merges overlapping peaks, replacing them with a single peak representing the narrowest region that covers all peaks that overlap by at least one base. There are at least two consequences of this that are worth noting.

First, as more peaksets are included in analysis, the average peak width tends to become longer as more overlapping peaks are detected and the start/end points are adjusted outward to account for them. Secondly, peak counts may not appear to add up as you may expect due to merging. For example, if one peakset contains two small peaks near to each other, while a second peakset includes a single peak that overlaps both of these by at least one base, these will all be replaced in the merged matrix with a single peak. As more peaksets are added, multiple peaks from multiple peaksets may be merged together to form a single, wider peak.

## 7.3 edgeR analysis

When `dba.analyze` is invoked using the default `method=DBA_EDGER`, a standardized differential analysis is performed using the `edgeR` package (Robinson et al. [2010]). This section details the precise steps in that analysis.

For each contrast, a separate analysis is performed. First, a matrix of counts is constructed for the contrast, with columns for all the samples in the first group, followed by columns for all the samples in the second group. The raw read count is used for this matrix; if the `bSubControl` parameter is set to `TRUE` (as it is by default), the raw number of reads in the control sample (if available) will be subtracted (with a minimum final read count of 1). Next the library size is computed for each sample for use in subsequent normalization. By default, this is the total number of reads in the library (calculated from the source BAM//BED file). Alternatively, if the `bFullLibrarySize` parameter is set to `FALSE`, the total number of reads in peaks (the sum of each column) is used. Note that “effective” library size (`bFullLibrarySize = FALSE`) may be more appropriate for situations when the overall signal (binding rate) is expected to be directly comparable between the samples. Next comes a call to `edgeR`’s `DGEList` function. The `DGEList` object that results is next passed to `calcNormFactors` with all other parameters retained as defaults (`method="TMM"`), returning an updated `DGEList` object. This is passed to `estimateCommonDisp` with default parameters.

If the method is `DBA_EDGER_CLASSIC`, then if `bTagwise` is `TRUE` (most useful when there are at least three members in each group of a contrast), the resulting `DGEList` object is then passed to `estimateTagwiseDisp`, with the prior set to 50 divided by two less than the total number of samples in the contrast, and `trend="none"`. The final steps are to perform testing to determine the significance measure of the differences between the sample groups by calling `exactTest` (Robinson and Smyth [2007]) using the `DGEList` with the `dispersion` set based on the `bTagwise` parameter.

If the method is `DBA_EDGER_GLM` (the default), then a design matrix is generated with two coefficients (the Intercept and one of the groups). Next `estimateGLMCommonDisp` is called; if `bTagwise=TRUE`, `estimateGLMTagwiseDisp` is called as well. The model is fitted by calling `glmFit`,

and the specific contrast fitted by calling `glmLRT`, specifying that the second coefficient be dropped. Finally, an `exactTest` (McCarthy et al. [2012]) is performed, using either common or tagwise dispersion depending on the value specified for `bTagwise`.

This final `DGEList` for contrast `n` is stored in the `DBA` object as

```
DBA$contrasts[[n]]$edgeR
```

and may be examined and manipulated directly for further customization. Note however that if you wish to use this object directly with `edgeR` functions, then the `bReduceObjects` parameter should be set to `FALSE`, otherwise the default value of `TRUE` will result in essential object fields being stripped.

If a blocking factor has been added to the contrast, an additional `edgeR` analysis is carried out. This follows the `DBA_EDGER_GLM` case detailed above, except a more complex design matrix is generated that includes all the unique values for the blocking factor. These coefficients are all included in the `glmLRT` call. The resultant object is accessible as

```
DBA$contrasts[[n]]$edgeR$block.
```

## 7.4 DESeq analysis

When `dba.analyze` is invoked using `method=DBA_DESEQ`<sup>5</sup>, a standardized differential analysis is performed using the `DESeq` package (Anders and Huber [2010]). This section details the precise steps in that analysis.

For each contrast, a separate analysis is performed. First, a matrix of counts is constructed for the contrast, with columns for all the samples in the first group, followed by columns for all the samples in the second group. The raw read count is used for this matrix; if the `bSubControl` parameter is set to `TRUE` (as it is by default), the raw number of reads in the control sample (if available) will be subtracted. Next the library size is computed for each sample for use in subsequent normalization. By default, this is the total number of reads in the library (calculated from the source BAM//BED file). Alternatively, if the `bFullLibrarySize` parameter is set to `FALSE`, the total number of reads in peaks (the sum of each column) is used. Note that "effective" library size (`bFullLibrarySize = FALSE`) may be more appropriate for situations when the overall signal (binding rate) is expected to be directly comparable between the samples. The first step concludes with a call to `DESeq`'s `newCountDataSet` function, which returns a `CountDataSet` object. If `bFullLibrarySize` is set to `TRUE`, then `sizeFactors` is called with the number of reads in the BAM/BED files for each ChIP sample, divided by the minimum of these; otherwise, `estimateSizeFactors` is invoked. Next, `estimateDispersions` is called with the `CountDataSet` object and `fitType` set to `local`. If there are no replicates, (only one sample in each group), `method` is set to `blind`. Otherwise, if `bTagwise` is `TRUE`, `method` is set to `per-condition`; if it is `FALSE`, `method` is set to `pooled` (or `pooled-CR` for a blocking analysis).

If the method is `DBA_DESEQ_CLASSIC`, `nbinomTest` is called, and the result (reordered by adjusted p-value) saved for reporting.

---

<sup>5</sup>Note that `DESeq` can be made the default analysis method for a `DBA` object by setting `DBA$config$AnalysisMethod=DBA_DESEQ`.

If the method is `DBA_DESEQ_GLM` (the default), two models are fitted using `fitNbinomGLMs`: a full model is fitted with all the coefficients, and a second model is fitted with the second coefficient dropped. These are tested against each other using `nbinomGLMTest`, with the resulting p values adjusted using `p.adjust` (with `method="BH"`).

The final results are accessible within the `DBA` object as

```
DBA$contrasts[[n]]$DESeq1$DEdata
```

and may be examined and manipulated directly for further customization. Note however that if you wish to use this object directly with `DESeq` functions, then the `bReduceObjects` parameter should be set to `FALSE`, otherwise the default value of `TRUE` will result in essential object fields being stripped.

If a blocking factor has been added to the contrast, an additional `DESeq` analysis is carried out. This follows the `DBA_DESEQ_GLM` case detailed above, except a more complex design is generated when `newCountDataSet` is called that includes all the unique values for the blocking factor. These coefficients are all included in the `fitNbinomGLMs` calls. The resultant object is accessible as

```
DBA$contrasts[[n]]$DESeq1$block$DEdata.
```

## 7.5 DESeq2 analysis

When `dba.analyze` is invoked using `method=DBA_DESEQ2`<sup>6</sup>, a standardized differential analysis is performed using the `DESeq2` package (Anders and Huber [2010]). This section details the precise steps in that analysis.

For each contrast, a separate analysis is performed. First, a matrix of counts is constructed for the contrast, with columns for all the samples in the first group, followed by columns for all the samples in the second group. The raw read count is used for this matrix; if the `bSubControl` parameter is set to `TRUE` (as it is by default), the raw number of reads in the control sample (if available) will be subtracted. Next the library size is computed for each sample for use in subsequent normalization. By default, this is the total number of reads in peaks (the sum of each column). Alternatively, if the `bFullLibrarySize` parameter is set to `TRUE`, the total number of reads in the library (calculated from the source BAM/BED file) is used. The first step concludes with a call to `DESeq2`'s `DESeqDataSetFromMatrix` function, which returns a `DESeqDataSet` object.

If `bFullLibrarySize` is set to `TRUE`, then `sizeFactors` is called with the number of reads in the BAM/BED files for each ChIP sample, divided by the minimum of these; otherwise, `estimateSizeFactors` is invoked.

`estimateDispersions` is then called with the `DESeqDataSet` object and `fitType` set to `local`. Next the model is fitted and tested using `nbinomWaldTest`. The final results (as a `DESeqDataSet`) are accessible within the `DBA` object as

```
DBA$contrasts[[n]]$DESeq2$DEdata
```

and may be examined and manipulated directly for further customization. Note however that if you wish to use this object directly with `DESeq2` functions, then the `bReduceObjects` parameter

---

<sup>6</sup>Note that `DESeq2` can be made the default analysis method for a `DBA` object by setting `DBA$config$AnalysisMethod=DBA_DESEQ2`.

should be set to FALSE, otherwise the default value of TRUE will result in essential object fields being stripped.

If a blocking factor has been added to the contrast, an additional DESeq2 analysis is carried out by setting the `design` to include all the unique values for the blocking factor. This occurs before the dispersions are calculated. The resultant `DESeqDataSet` object is accessible as

```
DBA$contrasts[[n]]$DESeq2$block$DEdata.
```

## 8 Vignette Data

Due to space limitations, the aligned reads associated with the cell line data used in this vignette are not included as part of the `DiffBind` package.

If you would like to get access to the reads, you can email `rory.stark@cruk.cam.ac.uk` and ask to be added to the Dropbox containing the vignette data.

Alternatively, you can use all of the original data archived at GEO. A script is included with the package that will download the data. This can be run by executing:

```
> source(file.path(system.file("extra", package="DiffBind"), "tamoxifen_GEO.R"))
```

Note that the data are more than 20 GB in size, and that this script can take a very long time to run (it took me about 12 hours). You may have a more efficient way to download these files from GEO.

Also included is a sample sheet that uses the downloaded data. This sample sheet can also be examined to see which GEO accession IDs correspond to which samples. The sample sheet will be copied into your working directory when you execute the above script; otherwise it can be found at:

```
> cat(file.path(system.file("extra", package="DiffBind"), "tamoxifen_GEO.csv"))
```

## 9 Acknowledgements

This package was developed at Cancer Research UK's Cambridge Research Institute with the help and support of many people there. We wish to acknowledge everyone the Bioinformatics Core under the leadership of Matthew Eldridge, as well as the Nuclear Receptor Transcription Laboratory under the leadership of Jason Carroll. Researchers who contributed ideas and/or pushed us in the right direction include Caryn-Ross Innes, Vasiliki Theodorou, and Tamir Chandra among many others. We also thank members of the Gordon Smyth laboratory at the WEHI, Melbourne, particularly Mark Robinson and Davis McCarthy, for helpful discussions.

## 10 Setup

This vignette was built on:

```

> sessionInfo()

R version 3.0.3 (2014-03-06)
Platform: x86_64-apple-darwin10.8.0 (64-bit)

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

attached base packages:
[1] parallel stats graphics grDevices utils datasets methods
[8] base

other attached packages:
[1] DESeq2_1.2.10 RcppArmadillo_0.4.100.2.1
[3] Rcpp_0.11.1 DESeq_1.14.0
[5] lattice_0.20-27 locfit_1.5-9.1
[7] Biobase_2.22.0 DiffBind_1.8.5
[9] GenomicRanges_1.14.4 XVector_0.2.0
[11] IRanges_1.20.7 BiocGenerics_0.8.0

loaded via a namespace (and not attached):
[1] amap_0.8-12 annotate_1.40.1 AnnotationDbi_1.24.0
[4] bitops_1.0-6 caTools_1.16 DBI_0.2-7
[7] edgeR_3.4.2 gdata_2.13.2 genefilter_1.44.0
[10] geneplotter_1.40.0 gplots_2.12.1 grid_3.0.3
[13] gtools_3.3.1 KernSmooth_2.23-12 limma_3.18.13
[16] MASS_7.3-31 RColorBrewer_1.0-5 RSQLite_0.11.4
[19] splines_3.0.3 stats4_3.0.3 survival_2.37-7
[22] tools_3.0.3 XML_3.98-1.1 xtable_1.7-3
[25] zlibbioc_1.8.0

```

## References

- Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, Oct 2010. doi: 10.1186/gb-2010-11-10-r106.
- Q. Li, J.B. Brown, H. Huang, and P. Bickel. Measuring reproducibility of high-throughput experiments. *Annals of Applied Statistics*, 2011.
- D.J. McCarthy, Y. Chen, and G.K. Smyth. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic Acids Research*, 2012.

- Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–40, Jan 2010. doi: 10.1093/bioinformatics/btp616. URL <http://bioinformatics.oxfordjournals.org/cgi/content/full/26/1/139>.
- M.D. Robinson and G.K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, 2007.
- C.S. Ross-Innes, R. Stark, A.E. Teschendorff, K.A. Holmes, H.R. Ali, M.J. Dunning, G.D. Brown, O. Gojis, I.O. Ellis, A.R. Green, et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, 481(7381):389–393, 2012.
- Y. Zhang, T. Liu, C.A. Meyer, J. Eeckhoute, D.S. Johnson, B.E. Bernstein, C. Nussbaum, R.M. Myers, M. Brown, W. Li, et al. Model-based analysis of chip-seq (MACS). *Genome Biol*, 9(9):R137, 2008.