ChIPQC Report

Overview

This report was generated using ChIPQC
The report provides both general and ChIP-seq specific quality metrics and diagnostic graphics to allow for the quantitative assessment of ChIP-seq quality.
The report is split into three main sections:

- **QC Summary** - Overview of results.
- **QC Results** - Full QC results and figures.
- **QC files and versions** - Files and program versions used in QC

QC Summary

Table 1. Summary of ChIP-seq filtering and quality metrics.

ID	Tissue	Factor	Condition	Replicate	Reads	Dup%	ReadL	FragL	RelCC	SSD	RiP%	RiBL%
BT4741	BT474	ER	Resistant	1	776353	8.4	28	207	1.6	1.9	15	1.8
BT4742	BT474	ER	Resistant	2	782419	10	28	201	1.3	1.6	14	1.7
BT474c	BT474	Control	Resistant	c1	598010	3.3	28	98	0.24	1.1	3	1.7
TAMR1	MCF7	ER	Resistant	1	747610	16	28	205	1.8	2.2	19	2
TAMR2	MCF7	ER	Resistant	2	728601	5.8	28	203	1.5	1.4	12	1.6
TAMRc	MCF7	Control	Resistant	c4	779102	6.2	28	96	0.2	1.4	2.2	2.1
MCF71	MCF7	ER	Responsive	1	438994	21	28	188	1.5	2.5	26	1.7
MCF72	MCF7	ER	Responsive	2	465700	4.8	28	209	1.7	1.6	16	2.2
MCF73	MCF7	ER	Responsive	3	577273	10	28	207	1.9	2.2	22	1.8
MCF7c	MCF7	Control	Responsive	c2	485192	1.7	28	96	0.082	1.4	2.6	2.5
T47D1	T47D	ER	Responsive	1	507492	7.9	28	207	1.2	1.6	9.6	2.2
T47D2	T47D	ER	Responsive	2	1831766	9.6	28	216	1.2	2.3	5.7	2.3
T47Dc	T47D	Control	Responsive	c3	400396	32	36	126	0.16	6	1.2	8.6
ZR751	ZR75	ER	Responsive	1	804427	16	28	212	2.3	3.5	31	1.4
ZR752	ZR75	ER	Responsive	2	2918549	24	28	214	2.4	4.4	21	1.2
ZR75c	ZR75	Control	Responsive	c5	1023987	20	36	127	0.29	5.4	1.5	5.4

Table 1contains a summary of filtering and quality metrics generated by the ChIPQC package. Further information on these metrics, their associated figures and additional quality measures can be found within the related QC Results subsections.
A short description of **Table 1** metrics is provided below:

- **ID -** Unique sample ID.
- **Tissue/Factor/Condition** - Metadata associated to sample.
- **Replicate** - Number of replicate within sample group
- **Reads** - Number of sample reads within analysed chromosomes.
- **Dup%** - Percentage of MapQ filter passing reads marked as duplicates
- **FragLen** - Estimated fragment length by cross-coverage method
- **SSD** - SSD score (htSeqTools)
- **FragLenCC** - Cross-Coverage score at the fragment length
- **RelativeCC** - Cross-coverage score at the fragment length over Cross-coverage at the read length
- **RIP%** - Percentage of reads wthin peaks
- **RIBL%** - Percentage of reads wthin Blacklist regions

QC Results

Mapping, Filtering and Duplication rate

This section presents the mapping quality, duplication rate and distribution of reads in known genomic features.

Table 2. Number and percantage of mapped,duplicated and MapQ filter passing reads

ID	Tissue	Factor	Condition	Replicate	Unmapped	Mapped	Pass MapQ Filter and Dup	Total Dup%	Pass MapQ Filter%	Pass MapQ Filter and Dup%
BT4741	BT474	ER	Resistant	1	0	776353	55046	8.2	84	8.4
BT4742	BT474	ER	Resistant	2	0	782419	68448	9.9	85	10
BT474c	BT474	Control	Resistant	c1	0	598010	16164	3.7	82	3.3
TAMR1	MCF7	ER	Resistant	1	0	747610	95986	14	83	16
TAMR2	MCF7	ER	Resistant	2	0	728601	34703	5.7	82	5.8
TAMRc	MCF7	Control	Resistant	c4	0	779102	38680	6.3	81	6.2
MCF71	MCF7	ER	Responsive	1	0	438994	73806	18	79	21
MCF72	MCF7	ER	Responsive	2	0	465700	17830	4.9	79	4.8
MCF73	MCF7	ER	Responsive	3	0	577273	47976	9.3	81	10
MCF7c	MCF7	Control	Responsive	c2	0	485192	6063	2.6	74	1.7
T47D1	T47D	ER	Responsive	1	0	507492	31493	7.5	79	7.9
T47D2	T47D	ER	Responsive	2	0	1831766	141288	9.6	81	9.6
T47Dc	T47D	Control	Responsive	c3	0	400396	70612	26	56	32
ZR751	ZR75	ER	Responsive	1	0	804427	111809	15	88	16
ZR752	ZR75	ER	Responsive	2	0	2918549	612998	22	88	24
ZR75c	ZR75	Control	Responsive	c5	0	1023987	151561	19	74	20

Table 2 shows the absolute number of total, mapped, passing MapQ filter and duplicated reads. The percent of mapped reads passing quality filter and marked as duplicates (Non-Redundant Fraction?) are also included.

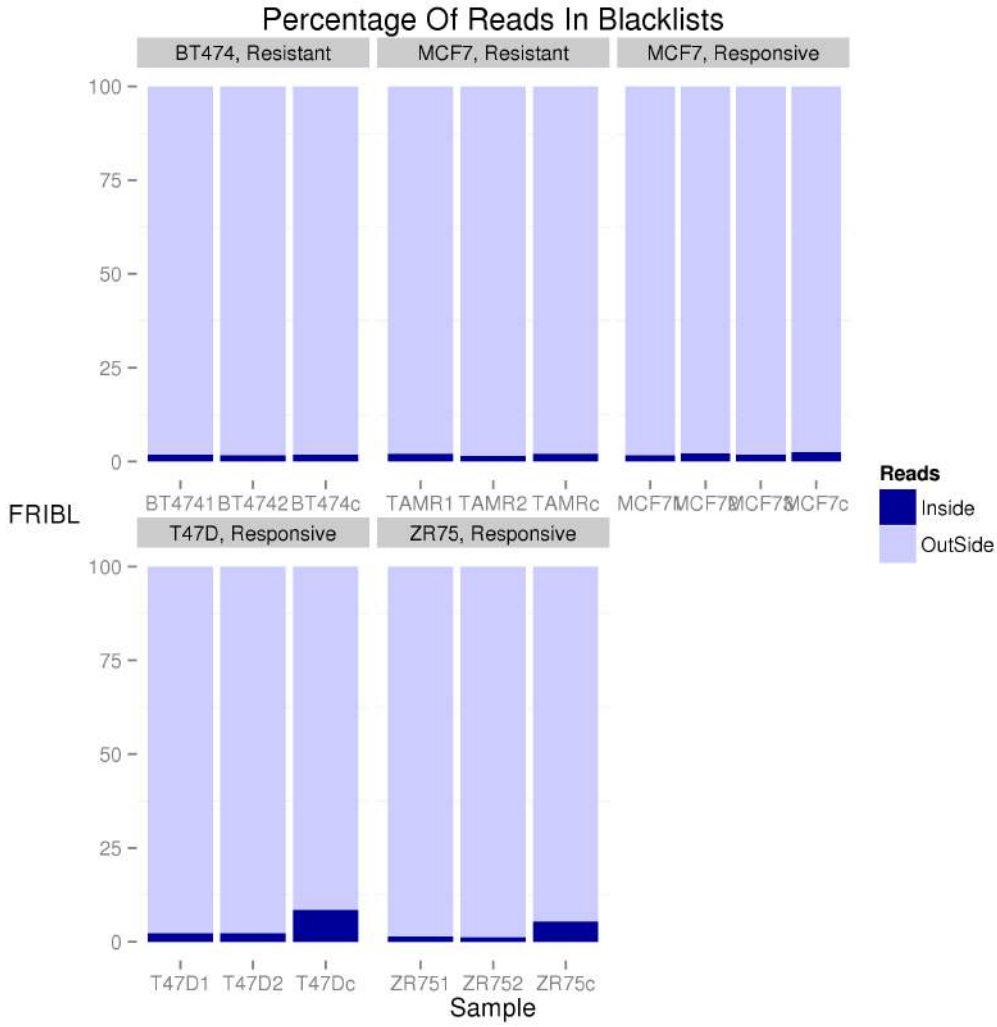
Description of read filtering and flag metrics:

- **Total Dup%-**Percentage of all **mapped** reads which are marked as **duplicates.**
- **Pass MapQ Filter%-**Percentage of all **mapped** reads which**pass MapQ quality** filter
- **Pass MapQ Filter and Dup%-**Percentage of all reads which pass **MapQ filter** and are marked as**duplicates.**

Duplication rates (Dup %) are dependent on the ChIP library complexity and the number of reads sequenced Higher duplication rates maybe due to low ChIP efficiency when read counts are lower or conversely saturation of ChIP signal when sequencing large number of reads. Since this metric is dependent on both read depth and the properties of the ChIP itself, comparison between biological or technical replicates of similat total read counts can best identify problematic libraries .

Highly mappable (multimappable) positions within the genome can attract large levels of duplication and so assessment of duplication before and after MapQ quality filtering can identify contribution of these positions to the duplication rate.

Figure 1. Barplot of the percentage of reads in blacklists

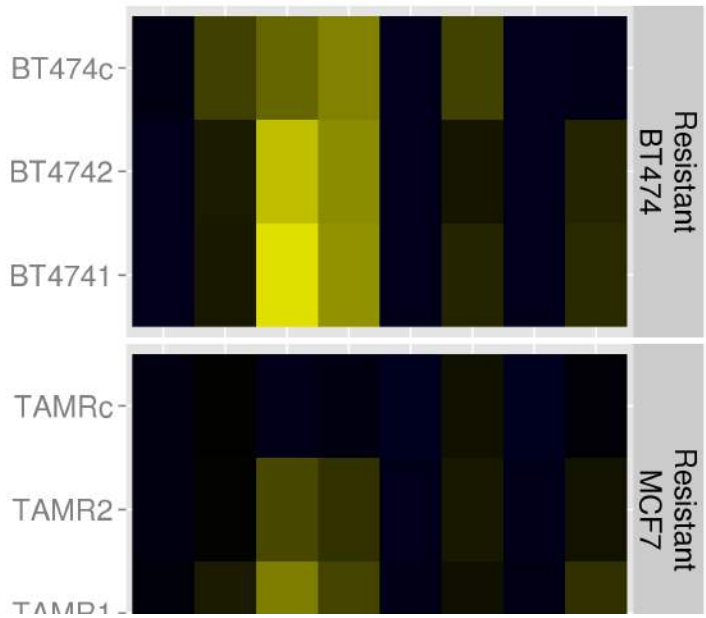


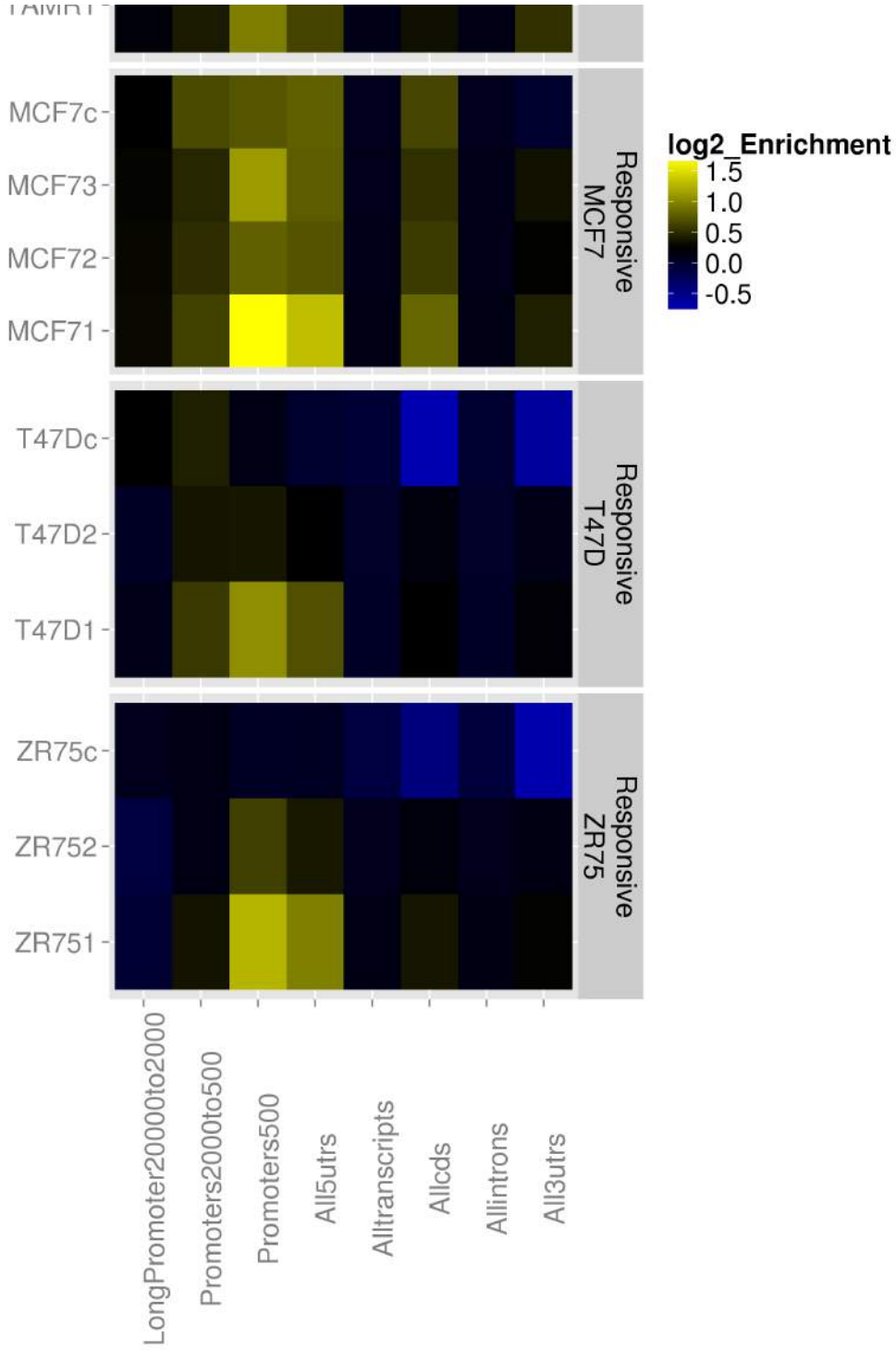
Genomic regions of high, anomalous signal have been seen to contribute directly to the Encode RCS and NSC metrics and can confound fragment length estimation, calculation of ChIP enrichment metrics (i.e. SSD) and comparison of signal between samples.

The identifaction of genomic stretches of artefact signal has been previously described for single samples using Input controls and more recently work as part of the Encode consortium has identified conserved regions of high artefact signal for many model organisms.

The percentage of total ChIP signal within known artefact regions can therefore be useful to evaluate the level of such confounding, abbarant signal in a sample. **(Figure 1)**

Figure 2. Heatmap of log2 enrichment of reads in genomic features





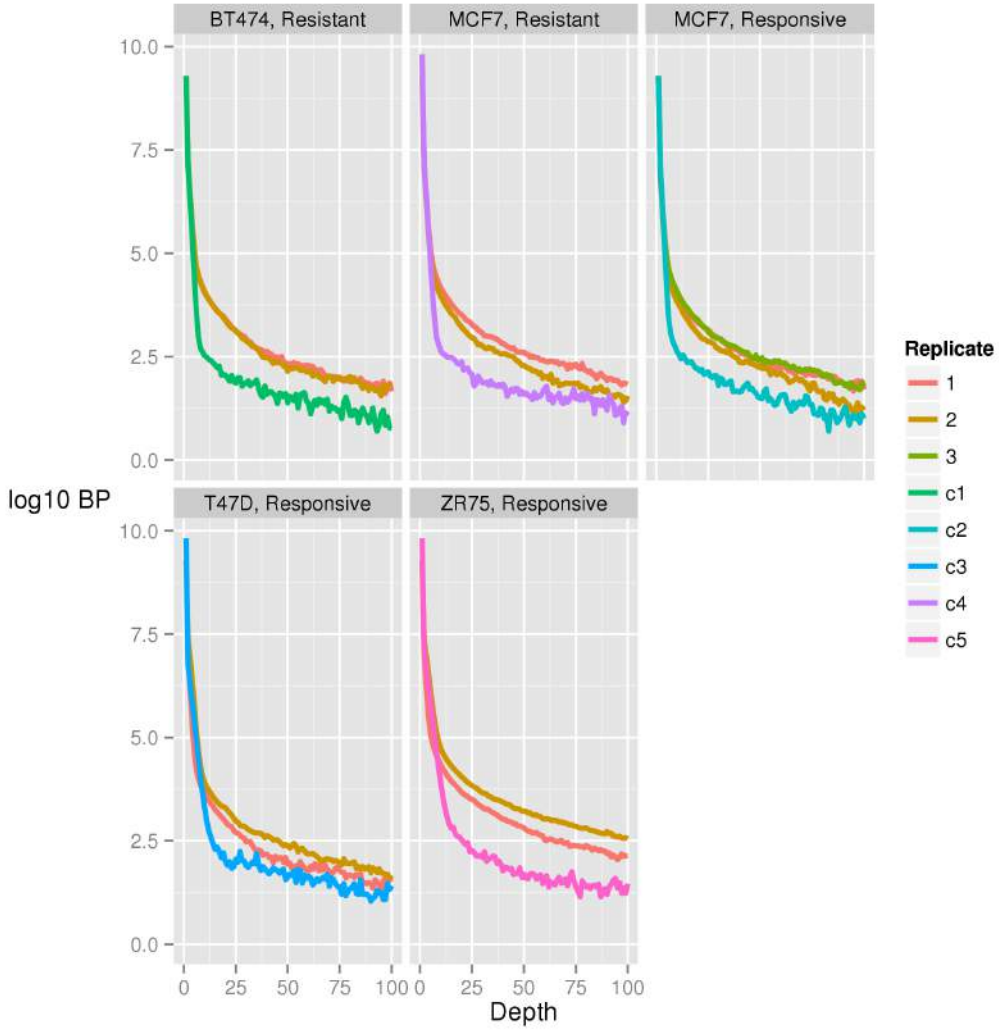
The distribution of reads across known genomic features such as genes and their subcomponents may allow further evaluation of ChIP-seq success and quality. A transcription factor know to preferentially bind at a genomic feature should show relative enrichment against other transcription factors showing no such preference. In addition,a replicate showing a differing enrichment patterns across genomic features compared to those within its sample group would highlight a potential outlier sample worthy of further investigation

Figure 2 shows the log2 enrichment of specified genomic features within samples with regions of greater enrichment showing bright yellow and lower enrichment seen in black

ChIP signal Distribution and Structure

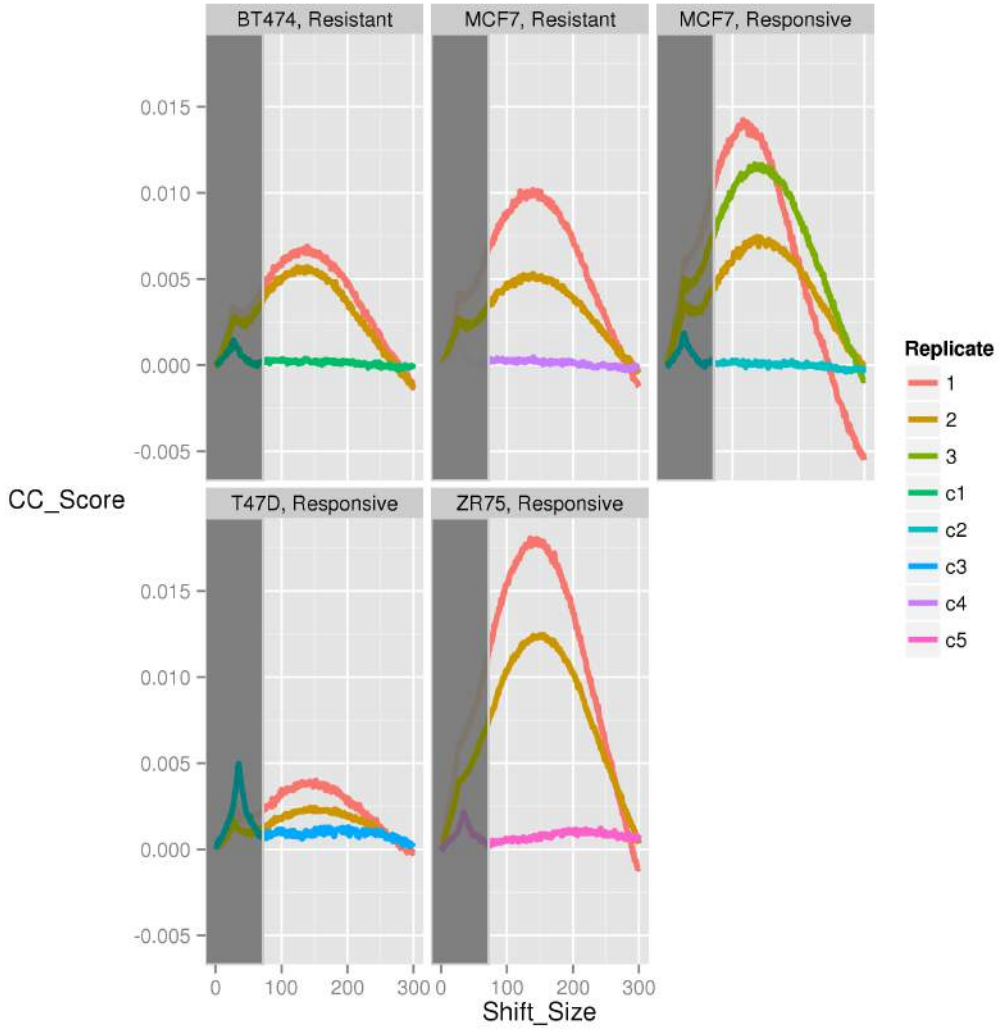
In this section, metrics relating to genome wide depths of coverage and, the relationship between Watson and Crick reads are presented. The metrics are the SSD metric and cross-coverage metrics, Relative_CC and fragmentLength_CC.

Figure 3. Plot of the log2 base pairs of genome at differing read depths



SSD is the standard deviation of coverage normalised to the total number of reads. Evaluation of the number of bases at differing read depths,**(figure 3)**alongside the use of the SSD metric allow for an assessment of the distribution of ChIP-seq or input signal. Successfull Histone and transcription factor ChIP-seq samples will show a higher proportion of genomic positions at greater depths and equivalence of sample and input SSD scores highlights either an unsuccessful ChIP or high levels of anomalous input signal

Figure 4. Plot of CrossCoverage score after successive strand shifts



An important measure of ChIP successive is the degree to which Watson and Crick reads cluster around the centres of transcription factor bindind sites or epigentic marks.

Transcription factor binding sites identified by ChIP-seq will show two distinct peaks of Watson and Crick strands separated by the fragment length. Here the method of cross-coverage (ChIPseq package) analysis is used to investigate this spatial clustering of Watson and Crick reads.

To investigate this spatial clustering, reads on the positive strand are shifted in 1bp steps and the total proportion genome now covered by both strands combined is assessed. **Figure 4** shows the CCov_Score (described below) after successive shifts. The points of highest outside of the read-length exclusion region, 2* the read length, (marked in grey) is considered the fragment length

Following the methodology first presented for cross-correlation by Encode to calculate the Relative Strand Cross Correlation (NSC) and Normalised Strand Cross Correlation, the Relative Cross Coverage score and Fragment Length Cross Coverage score are calculated.

The calculation of cross-coverage (CCov),Relative CCov and Fragment Length CCov scores are explained below:

- **CCov_Score-** 1-(Total covered genome size at strand shift)/(covered genome size with no shift)
- **Fragment Length CCov-** (CCov of fragment length strand shift)/(Minimum CCov)
- **Relative CCov-** (CCov of fragment length strand shift)/(CCov of read length strand shift)

Peak Profile and ChIP Enrichment

Following the identification of genome wide enrichment (peak calling), the percentage of ChIP signal within enriched regions, as well the average profile across these regions can be used to further evaluate ChIP quality

Figure 5. Plot of the average signal profile across peaks

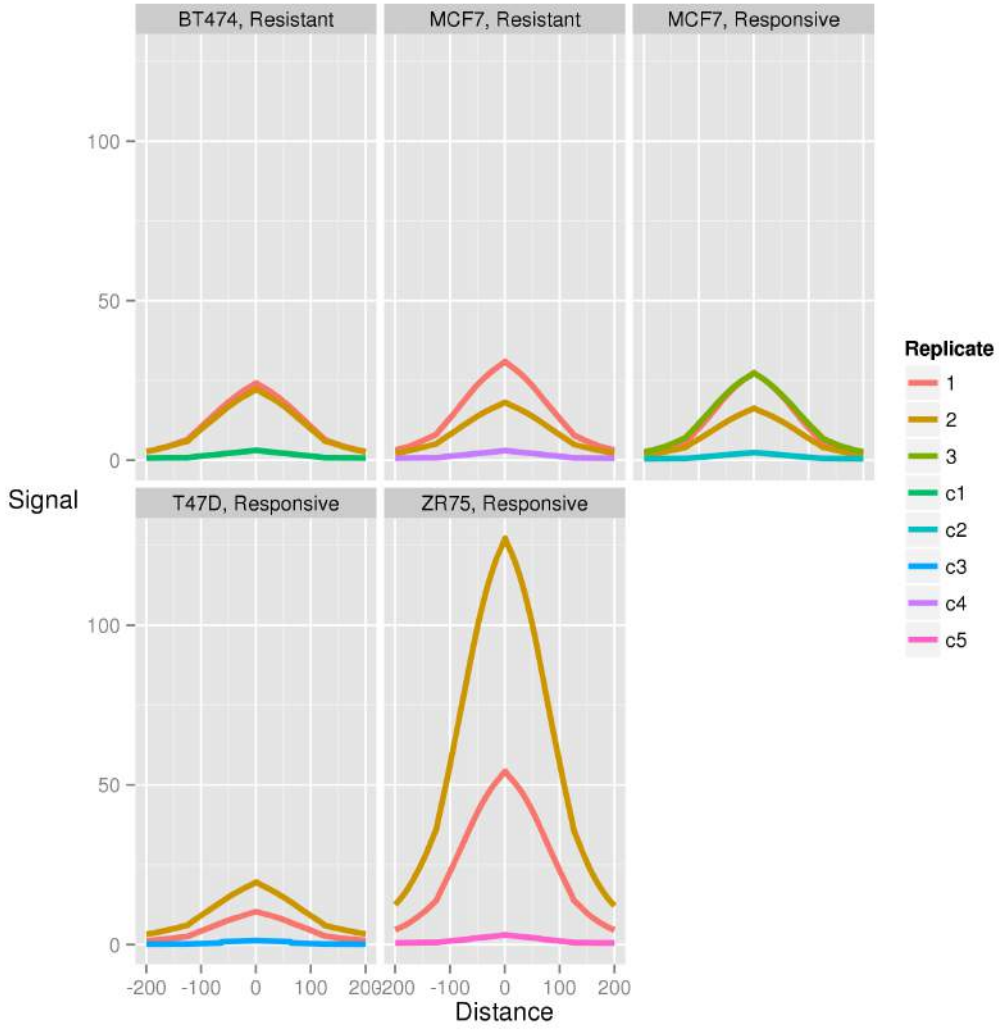


Figure5 represents the mean read depth across and around peaks. By identying the average pattern of enrichment across peaks, differences in both mean peak height and shape may be found. This not only assits in a better characterisation of ChIP enrichment but can aid in the identification of outliers.

Figure 6. Barplot of the percentage number of reads in peaks

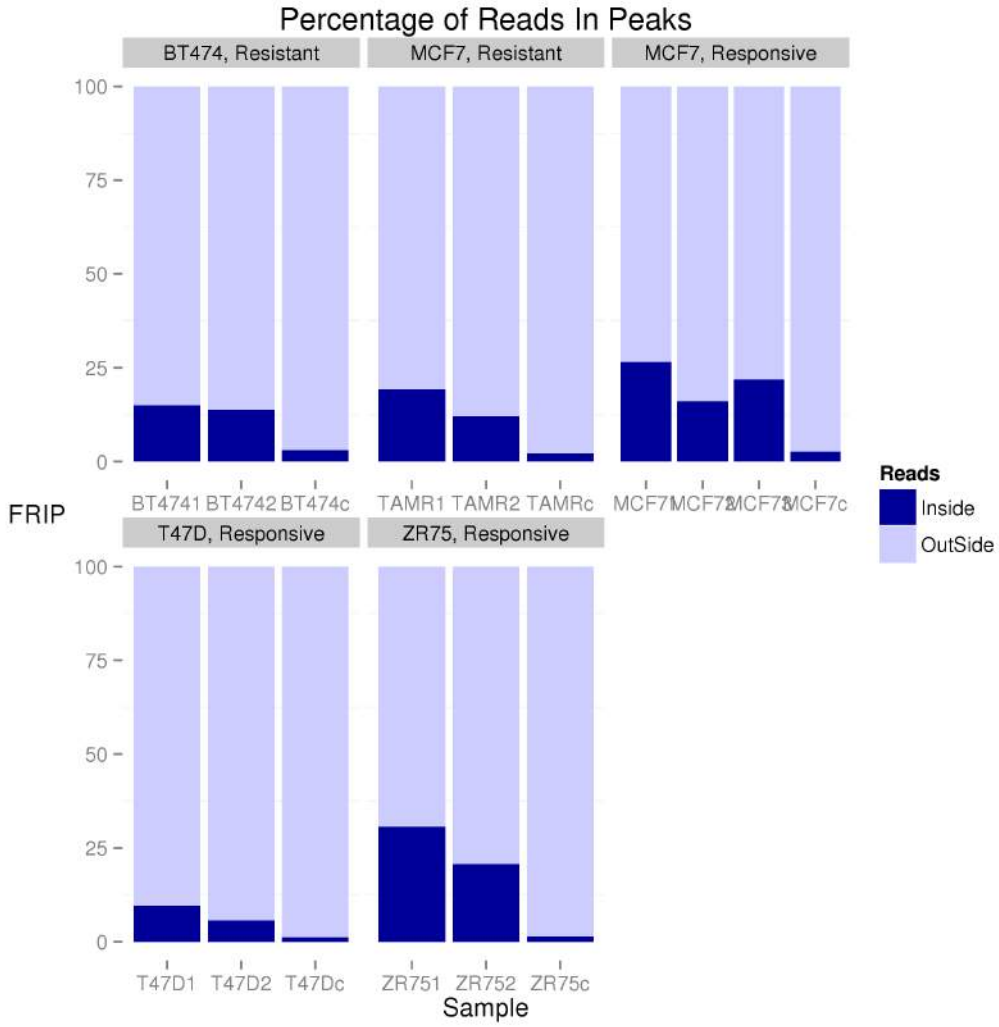


Figure6 shows the total percentage of reads contained within enriched regions or peaks. The higher efficiency ChIP-seq will show a higher percentage of reads in enriched regions/peaks and longer epigenetic marks will often have a higher ranges of efficiencies than punctate marks or transcription factors.

Figure 7. Density plot of the number of reads in peaks

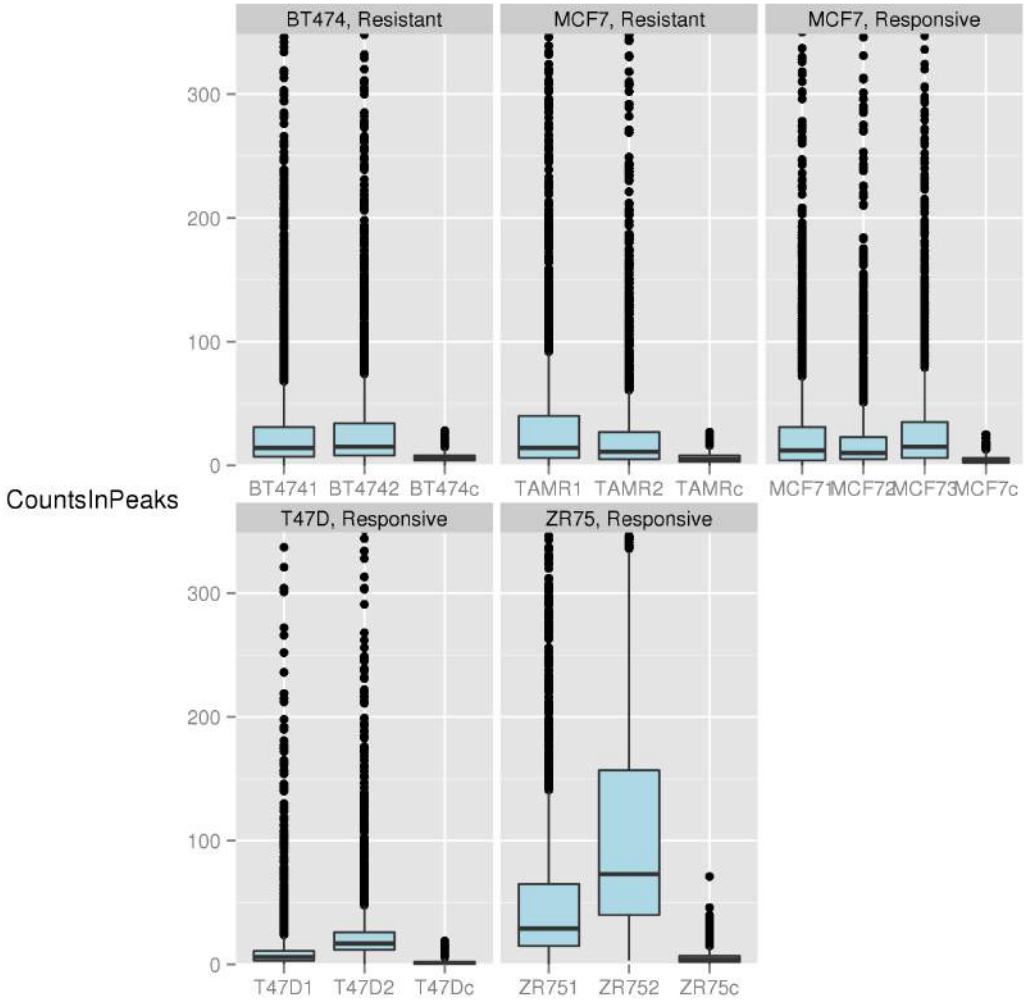


Figure7 shows the distribution of reads in all peaks. Evaluation of the distibution can allow for greater characteriation of the variability and range of signal in peaks within a sample and so better characterise the signal across peaks than the RIP score may allow.

Figure 8. Plot of correlation between peaksets

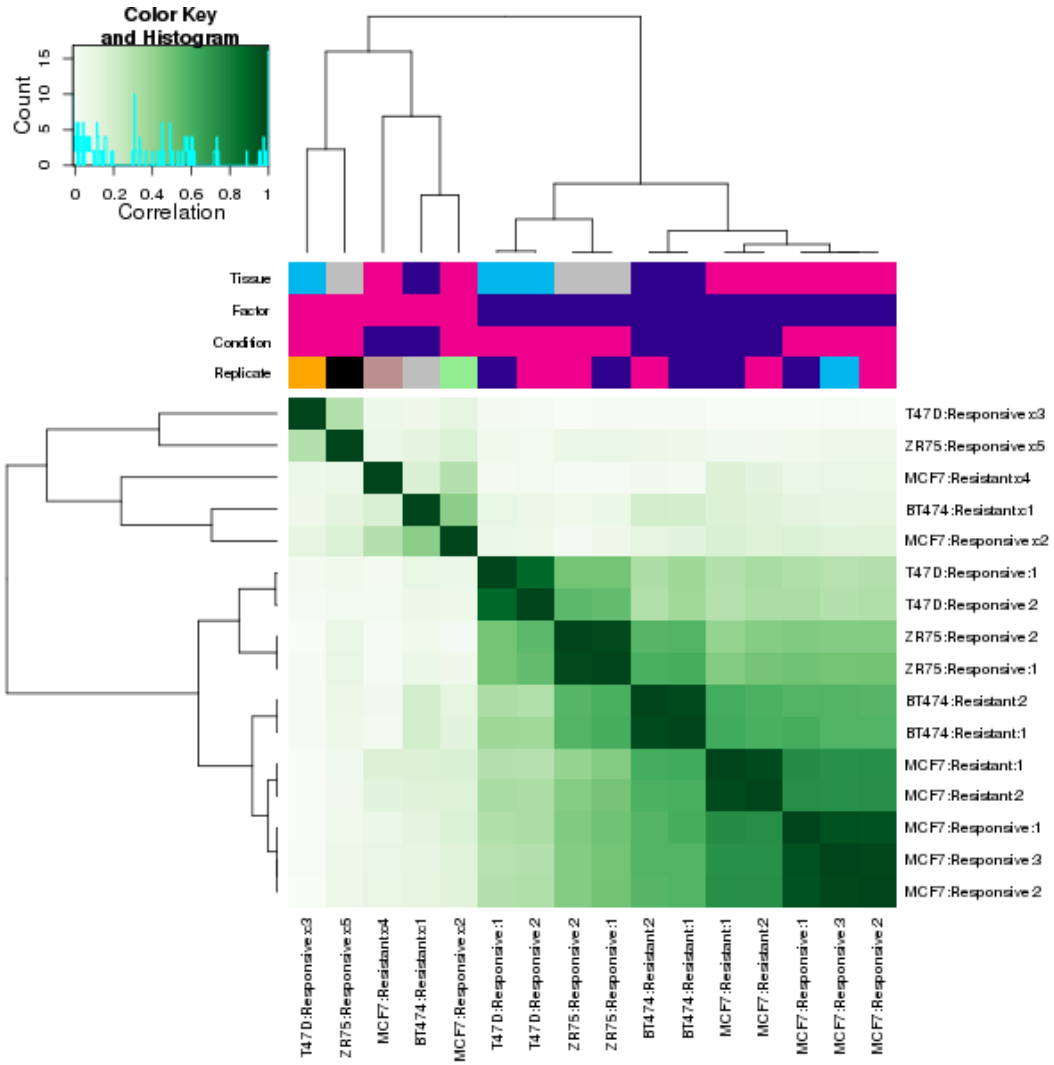


Figure 9. PCA of peaksets

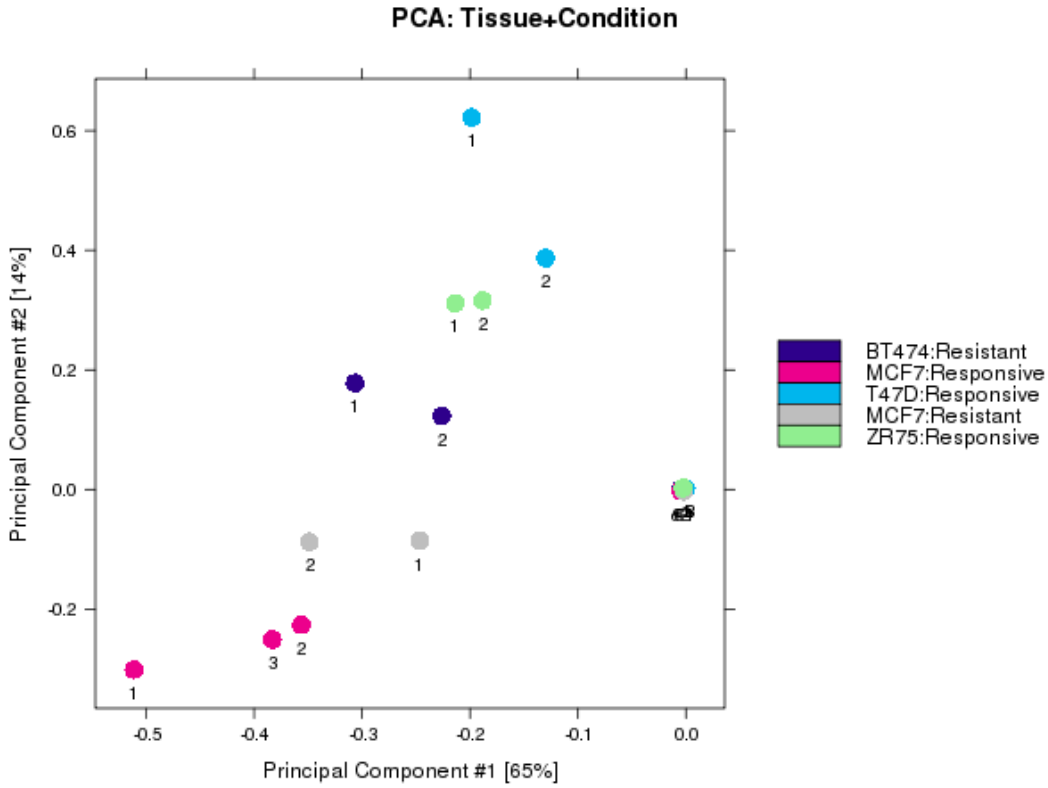


Figure8 and 9 shows the correlation between samples as a heatmap and by principal component analysis. Replicate samples of high quality can be expected to cluster together in the heatmap and be spatially grouped within the PCA plot.

Files and Versions

R Version Information

- Version: 3.1.0
- Version_String :R version 3.1.0 alpha (2014-03-18 r65213)

ChIPQC Version Information

- Version: ChIPQC:1.0.1
- Author: Tom Carroll, Wei Liu, Ines de Santiago, Rory Stark
- Maintainer: Tom Carroll , Rory Stark