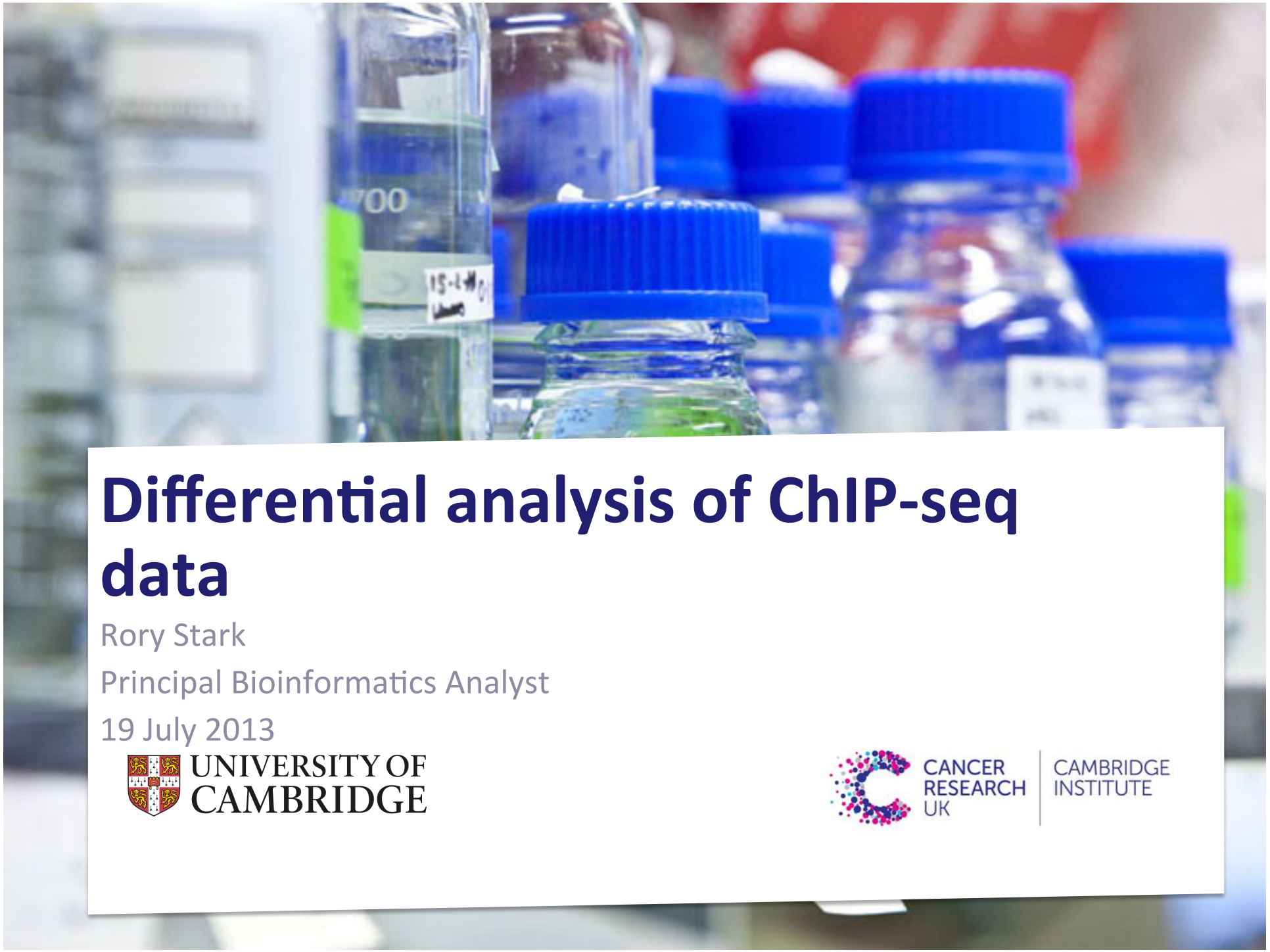


# Differential analysis of ChIP-seq data

Rory Stark

Principal Bioinformatics Analyst

19 July 2013

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

# Link to zipped copy of working directory:

http://goo.gl/b4pMP

# Analysis of ChIP-seq data

## EXPERIMENTAL DESIGN

- Controls and replicates

## QC/READ PROCESSING

- Library QC
- Alignment and filtering
- QC measures and assessment

## PEAK CALLING

- Peak callers

## DIFFERENTIAL BINDING ANALYSIS

- Occupancy-based analysis
- Affinity-based analysis

## VALIDATION AND DOWNSTREAM ANALYSIS

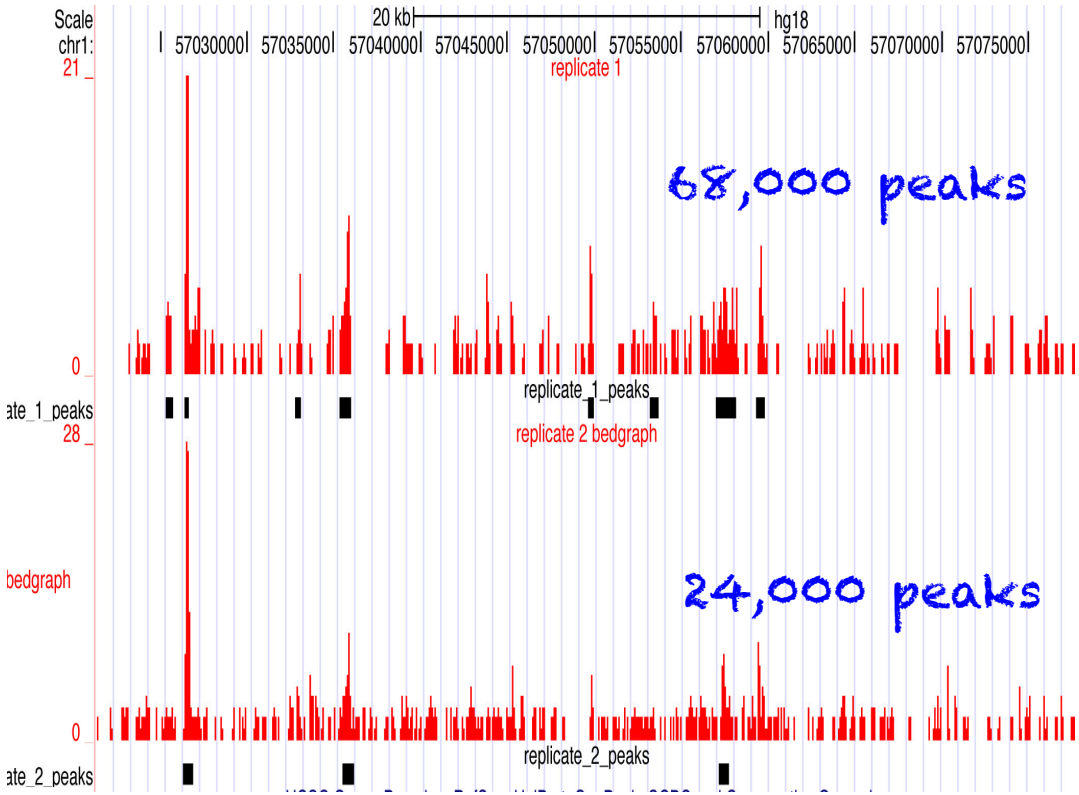
- Motif analysis
- Annotation
- Integrating binding and expression data

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

# Differential Binding Analysis


# Differential binding analysis: Observations

- ChIP-seq variability
  - Biological
  - Experimental
  - Technical
- Peak calling is noisy
  - Profusion of peak callers
  - Highly parametric
  - Callers have low agreement on marginal peaks
- Many samples involved
  - Conditions and treatments (contrasts)
  - Factors, marks, antibodies
  - Replicates ***required*** to capture variance




CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

# Differential binding analysis: Goals

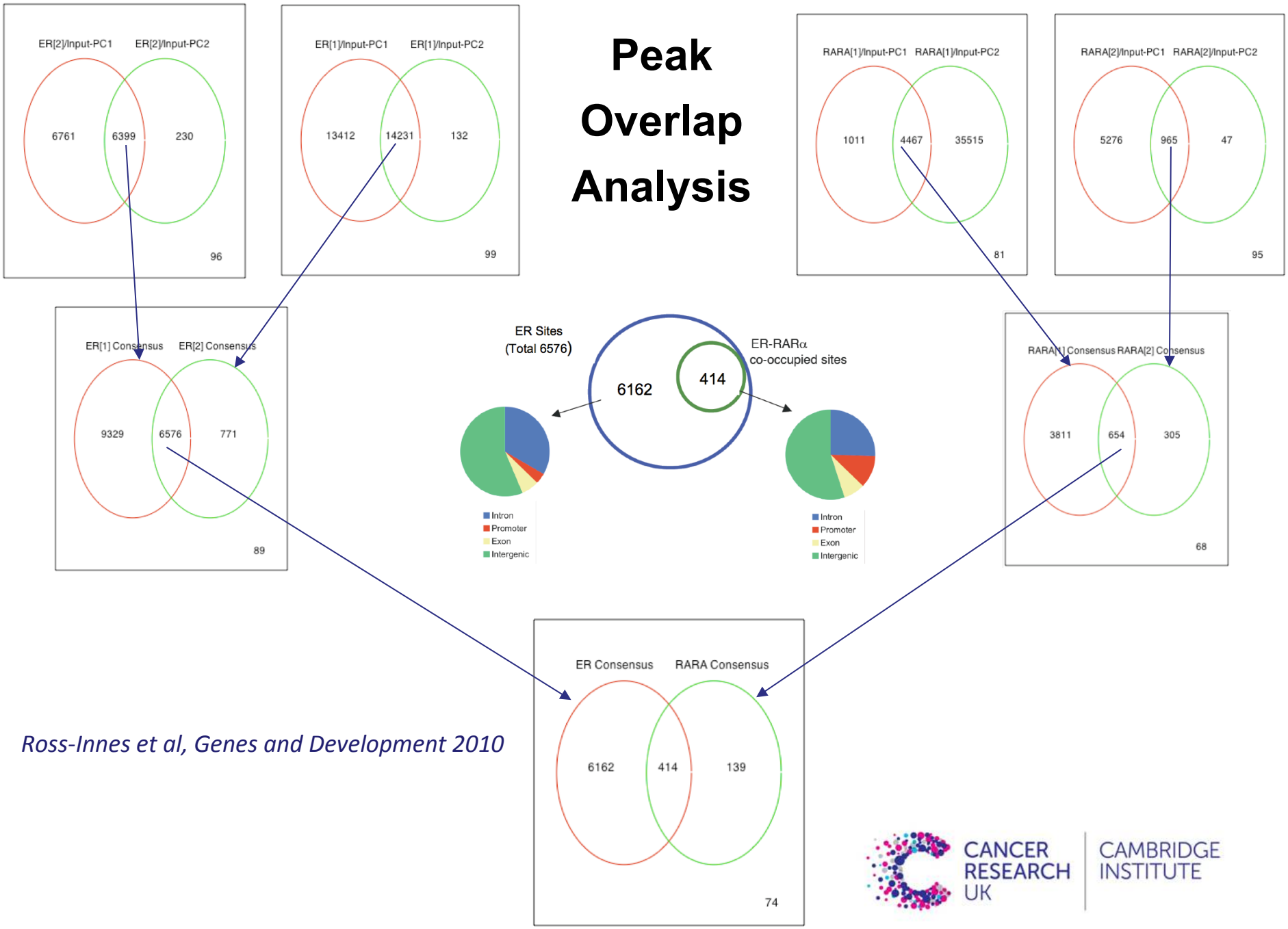
- Be robust to noise
  - Noisy experiments
  - Noisy peak calling
- Determine DB without requiring global binding maps for each ChIP
- Exploit quantitative affinity (read scores) beyond binary occupancy (peak calls)
- Functionally link differential regulatory events (DB) with differential mRNA levels (DE)

# Types of differential binding analysis

- Overlap analysis (peaks/site occupancy)
- Quantitative analysis
  - Binding site count density (ChipDiff, DiffBind)
  - Binding profile (MMDiff)
  - PCA of multiple factors (dPCA)

# Occupancy Analysis


# Peak Overlap Analysis




*Ross-Innes et al, Genes and Development 2010*

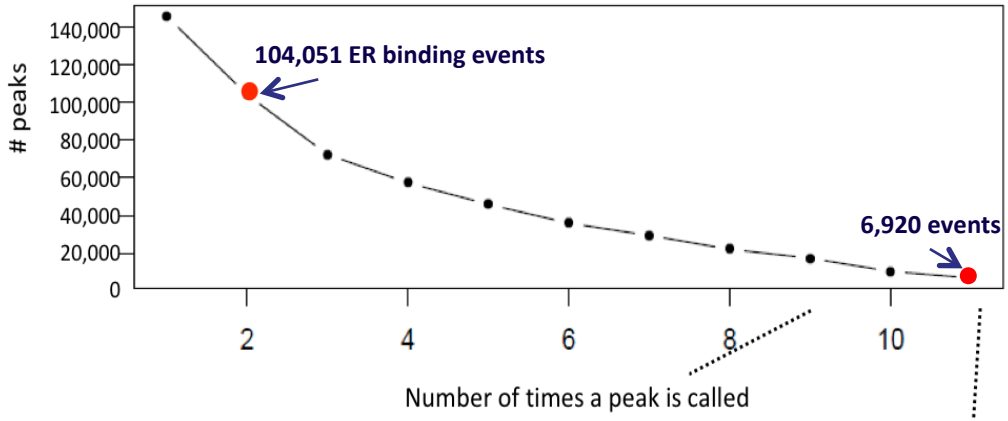
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

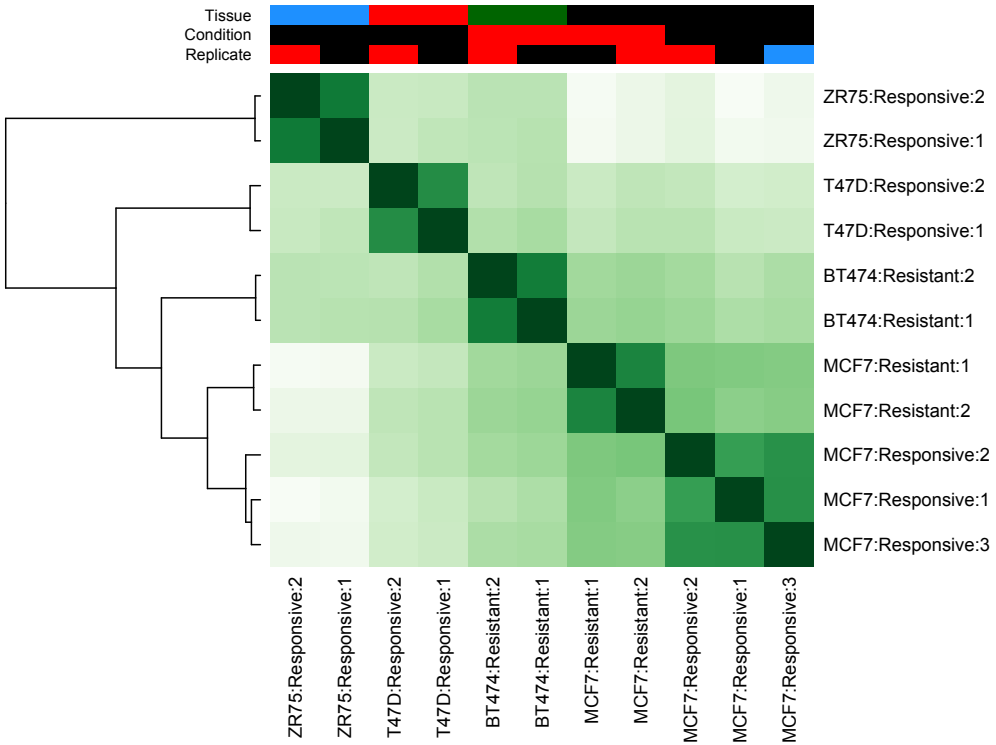
# Example domain: Tamoxifen resistance in breast cancer cell lines

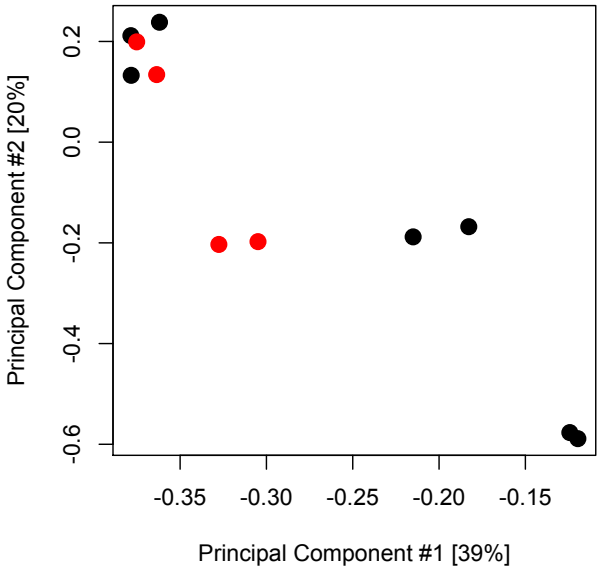
```
11 Samples, 112554 sites in matrix (157722 total):
        ID Tissue Factor  Condition  Treatment Replicate Peak.caller Intervals
1   MCF7+1   MCF7     ER Responsive Full-Media         1        macs     79605
2   MCF7+2   MCF7     ER Responsive Full-Media         2        macs     49199
3   MCF7+3   MCF7     ER Responsive Full-Media         3        macs     66428
4   T47D+1   T47D     ER Responsive Full-Media         1        macs     32825
5   T47D+2   T47D     ER Responsive Full-Media         2        macs     29144
6   ZR75+1   ZR75     ER Responsive Full-Media         1        macs     85577
7   ZR75+2   ZR75     ER Responsive Full-Media         2        macs     81340
8  BT474-1  BT474     ER  Resistant Full-Media         1        macs     41715
9  BT474-2  BT474     ER  Resistant Full-Media         2        macs     42143
10  MCF7-1   MCF7     ER  Resistant Full-Media         1        macs     66306
11_ MCF7-2   MCF7     ER  Resistant Full-Media         2        macs     45839
```

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

# Occupancy Clustering

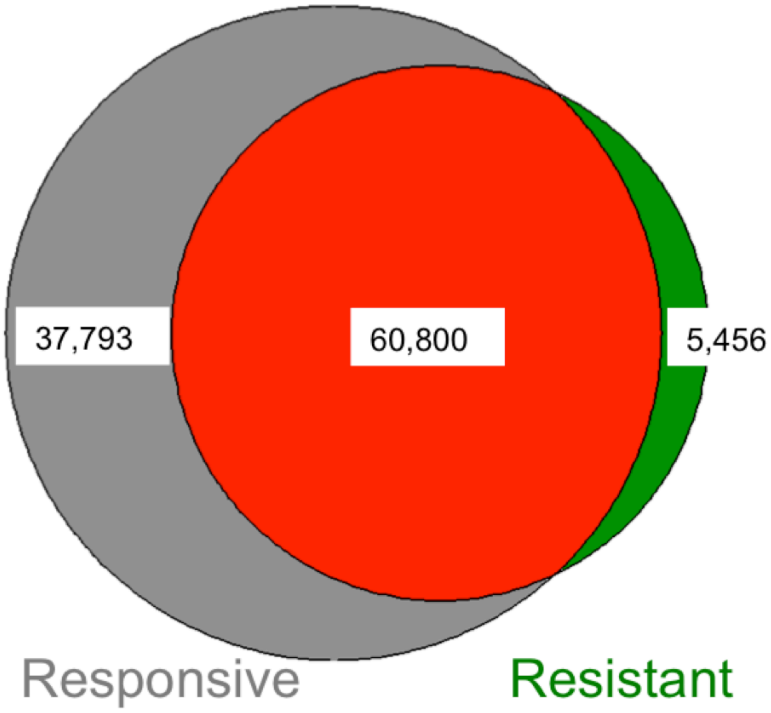


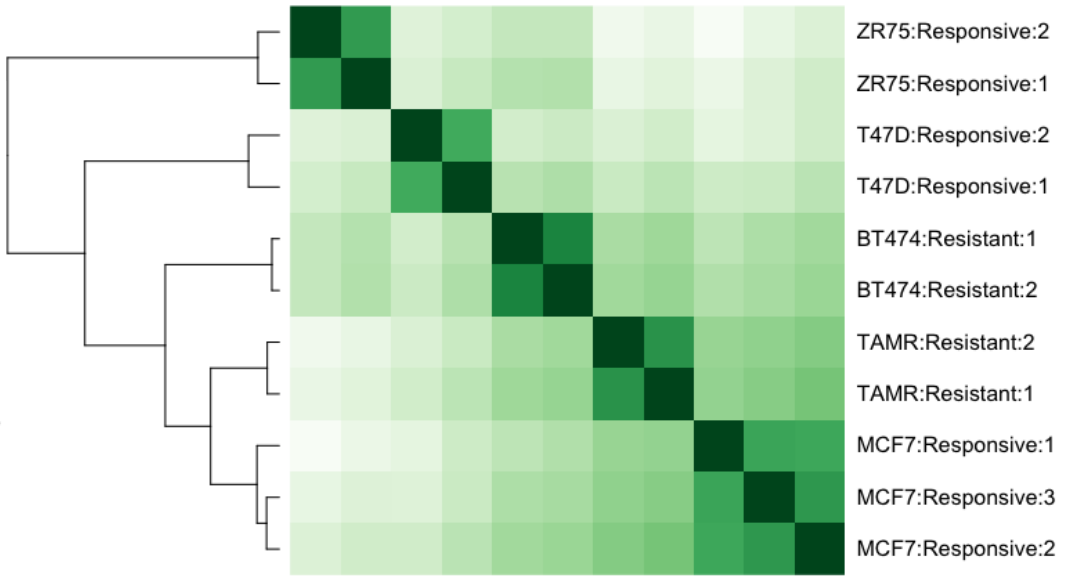



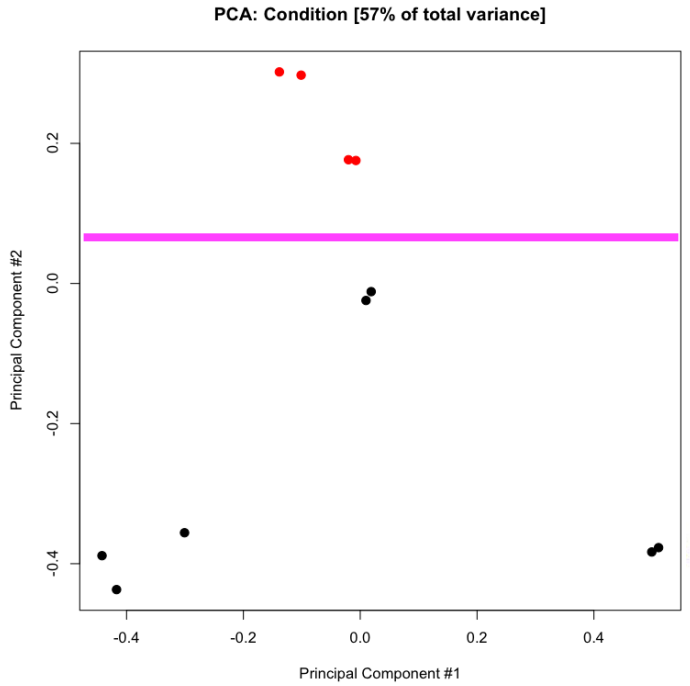




CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

# Occupancy Analysis




Observations:

- Overall loss of ER binding activity in resistant cell lines
- Using only status-unique sites:
  - Does not cluster by status
  - Separable in second principal component

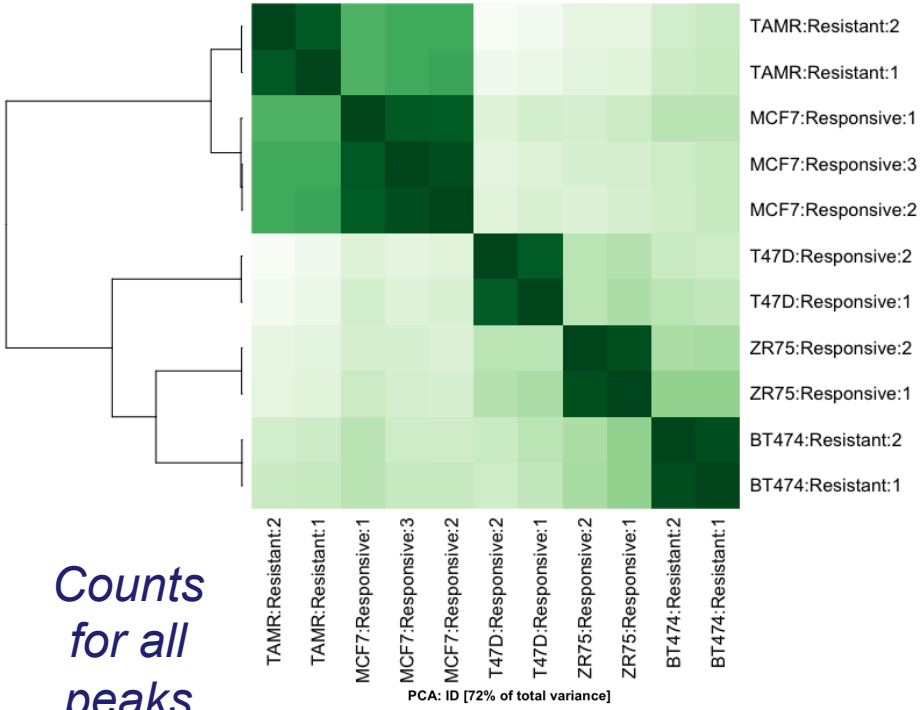
# Quantitative Analysis


# Binding affinity matrix

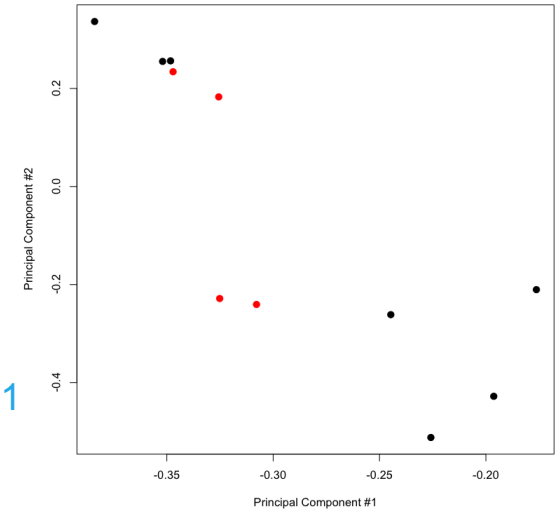
- **Rows**: decide interval (binding site) "universe"
  - Peak callers -> occupancy/overlaps
    - High-confidence sites (stringent)
    - All potential sites (lenient)
  - Genomic intervals
    - Promoters
    - Windows
- **Columns**: count and normalize reads for all samples in all intervals
  - Duplicate reads
  - Controls
  - Normalization

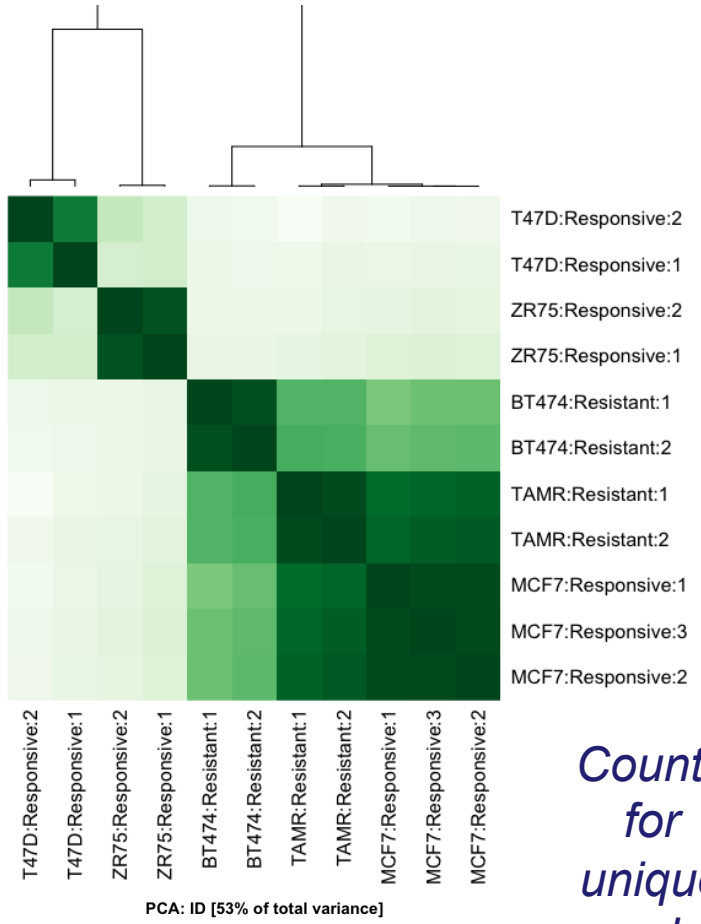
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

# Affinity (count) analysis

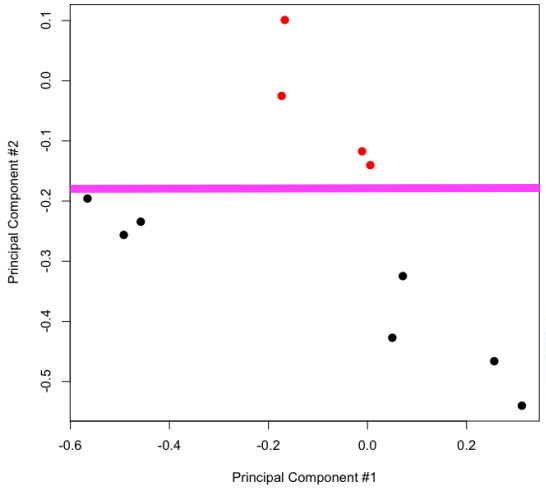



*Counts for all peaks*



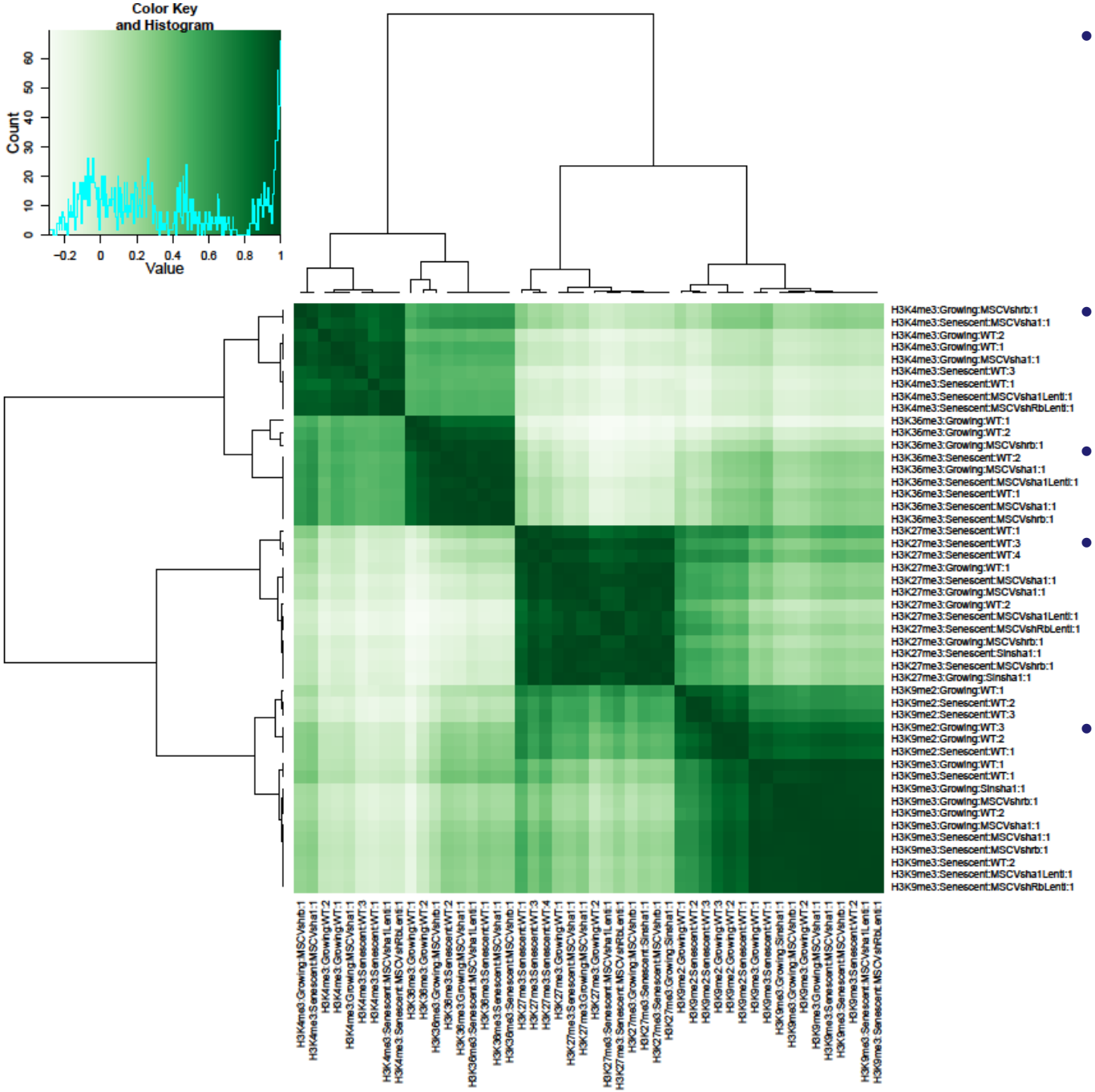

*Counts for unique peaks only*




1

CAMBRIDGE INSTITUTE

Data from Chandra et al Molecular Cell 2012



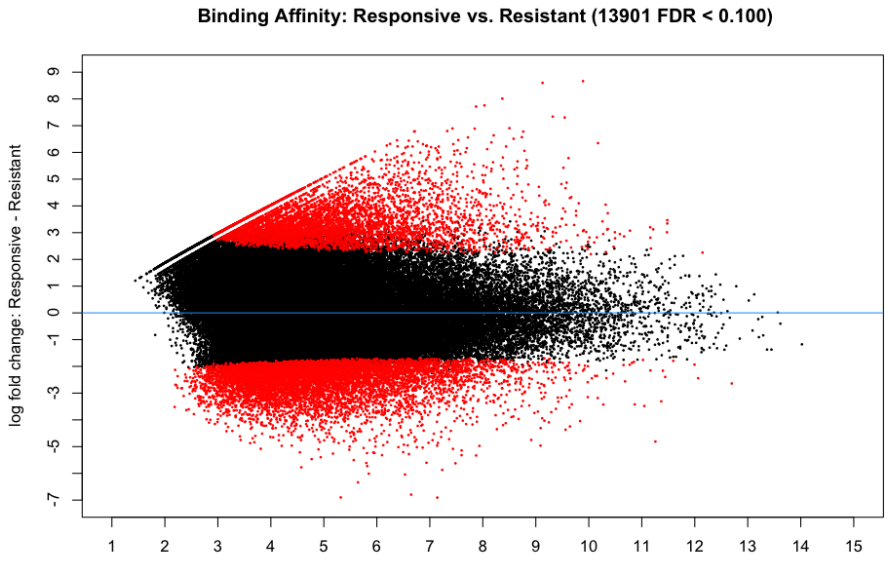

- **Histone marks**
  - *H3K4me3*
  - *H3K36me3*
  - *H3K9me2*
  - *H3K9me3*
  - *H3K27me3*
- **Conditions:**
  - Growing vs. Senescent
- ***Treatment:***
  - *WT vs. treated*
- **Replicates:**
  - 1-3 for each mark/condition/ treatment
- ***"Peaks":***
  - *Windows around TSSs (-1000, +4000)*

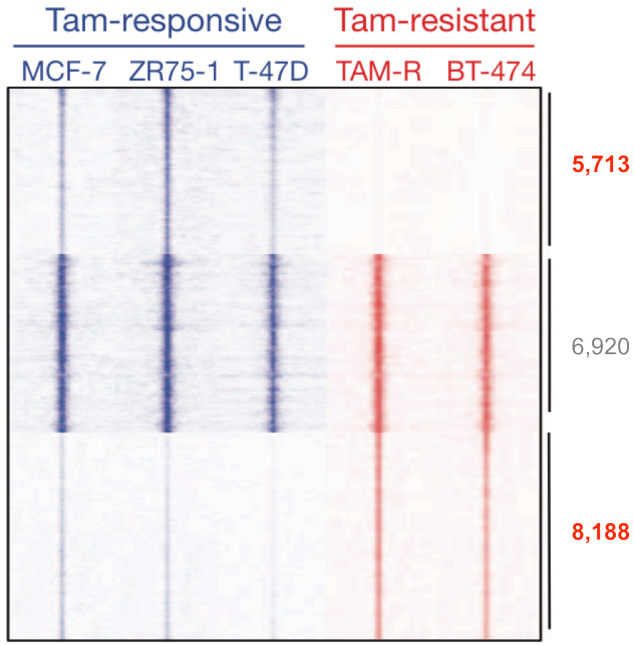
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

# Differential binding analysis

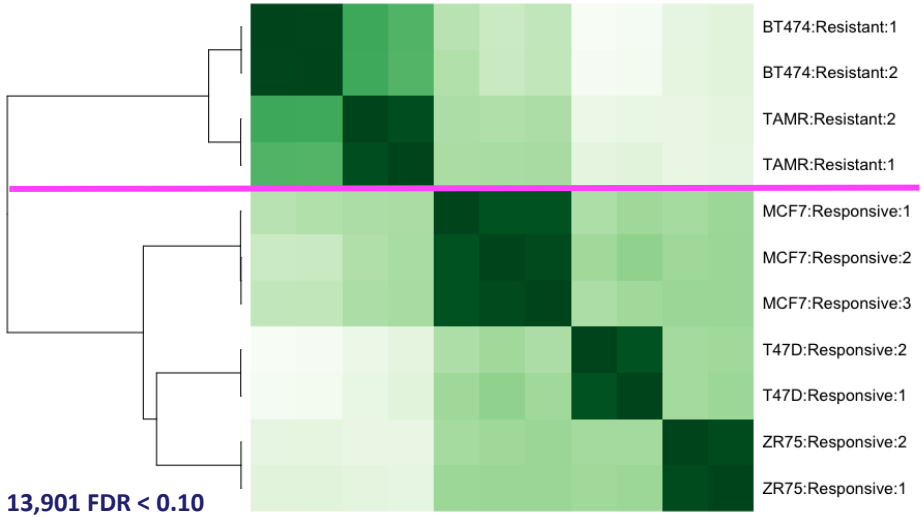
- **Determine contrasts**
  - Single-factor
  - Multi-factor (GLM/blocking)
    - Matched tumour-normal
    - Common tissue
    - Replicate groups (batch)

- **Run RNA-Seq DE package**
  - edgeR, DESeq, etc.
  - Fit negative binomial distribution
  - Exact test
  - Multiple testing correction (B&H FDR)

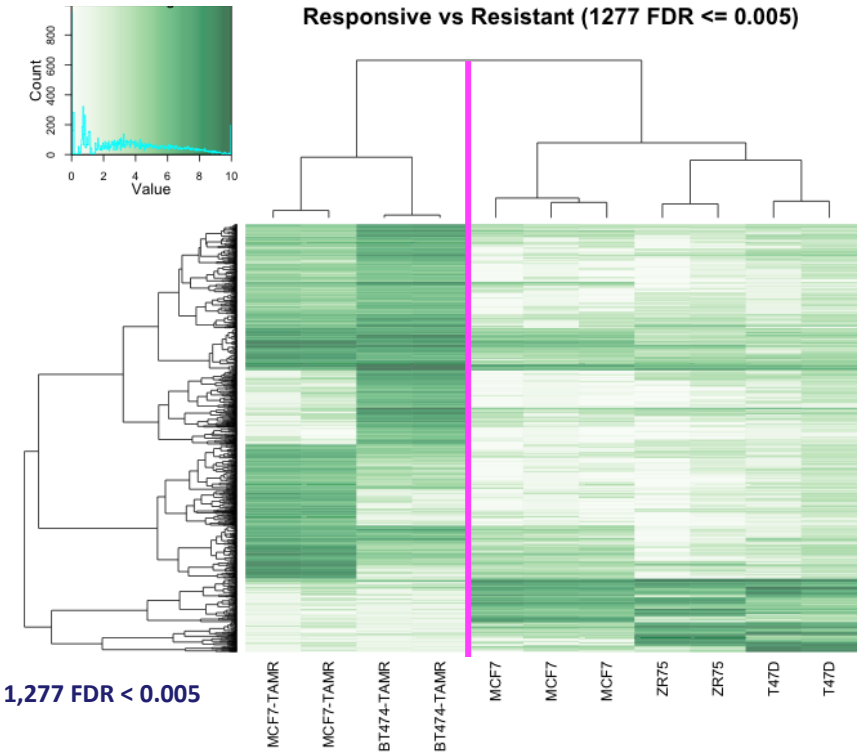


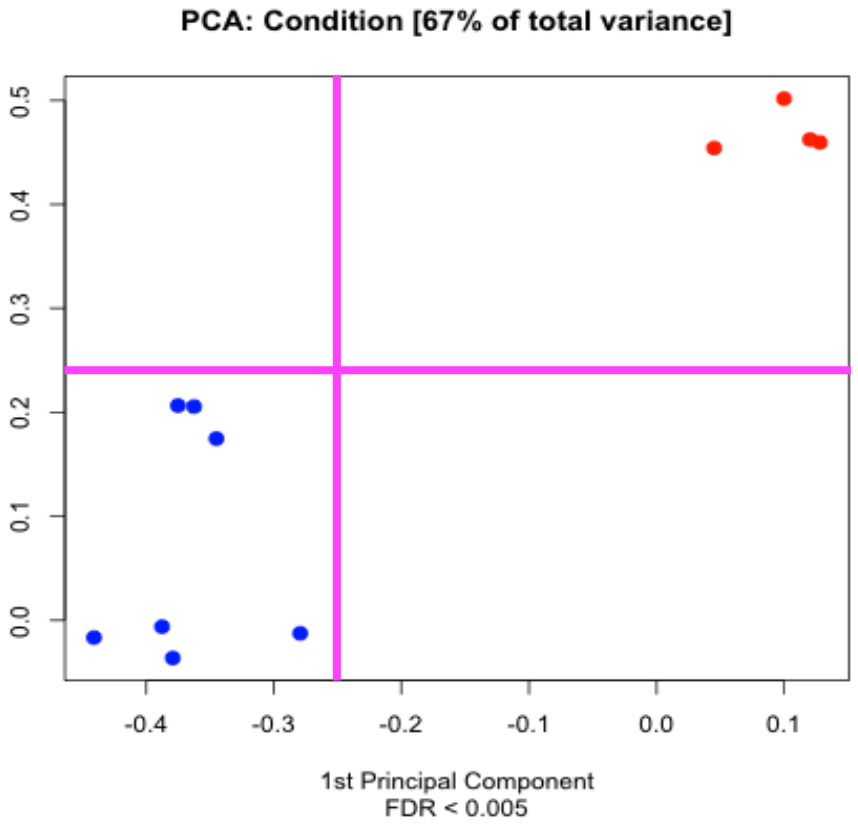

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE




**13,901 FDR < 0.10**



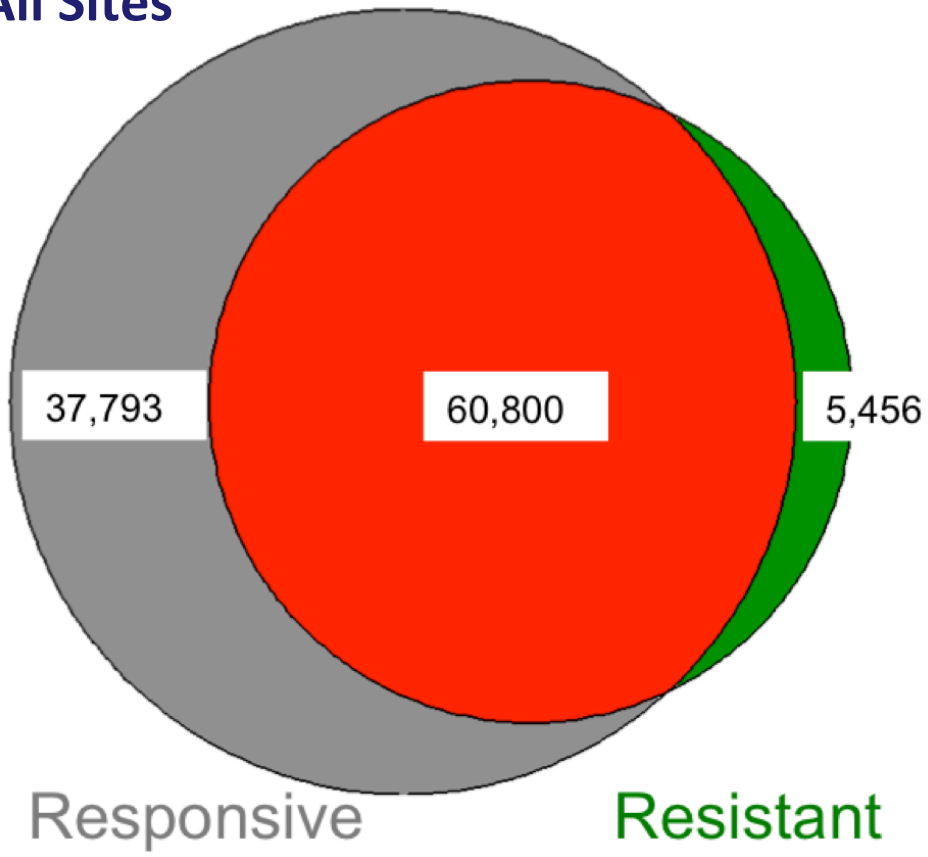

**1,277 FDR < 0.005**



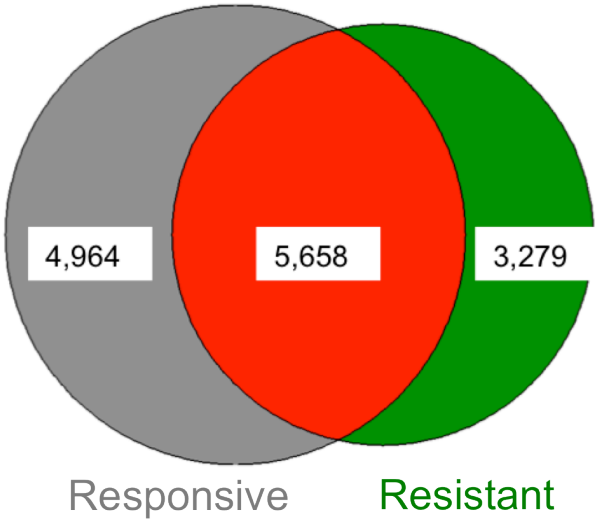

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

# Differential binding analysis: Occupancy vs. Affinity

## All Sites

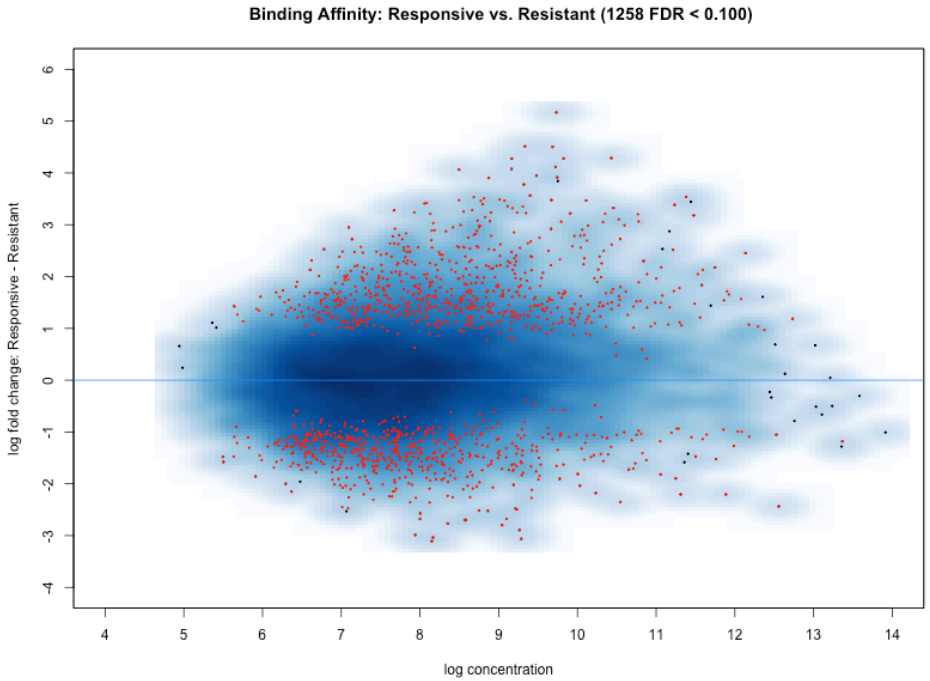


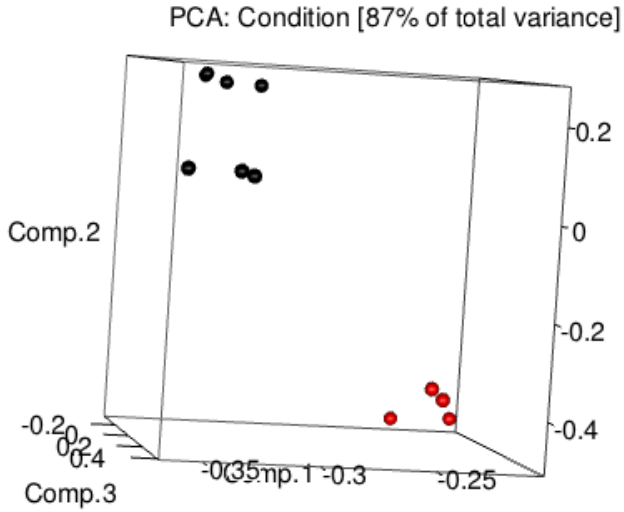

## Differentially Bound Sites

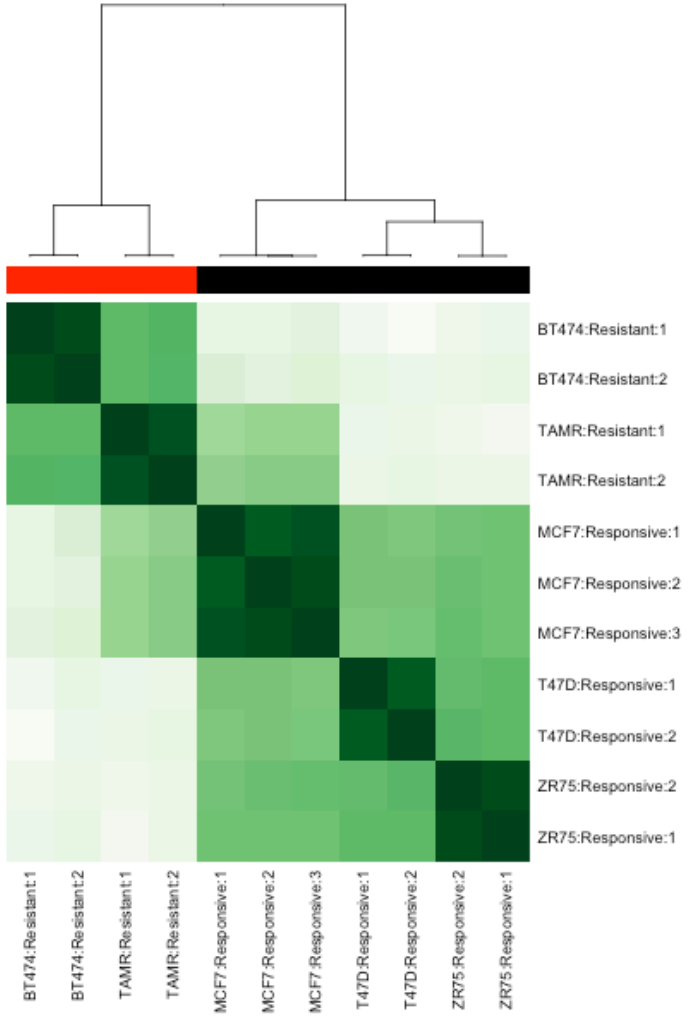



CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

# Differential binding signature can be isolated even amongst sites common to all samples




CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

# DiffBind


UNIVERSITY OF
CAMBRIDGE


CANCER
RESEARCH
UK
CAMBRIDGE
INSTITUTE

# R/Bioconductor package -- DiffBind

| | |
|---|---|
| **dba** | Construct a DBA object |
| **dba.peakset** | Add a peakset to a DBA object |
| | |
| **dba.overlap** | Compute binding site overlaps |
| **dba.count** | Count reads in binding sites |
| | |
| **dba.contrast** | Establish contrast(s) for analysis |
| **dba.analyze** | Execute differential binding analysis |
| **dba.report** | Generate report for a contrast analysis |
| | |
| **dba.plotHeatmap** | Heatmap plots (correlation/affinity) |
| **dba.plotPCA** | Principal Components Analysis plot |
| **dba.plotMA** | MA/scatter plot |
| **dba.plotBox** | Boxplot |
| **dba.plotVenn** | Venn diagram plot of overlaps |

```
> 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)
```

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

# DiffBind Workflow

1. Reading in peaksets
   - Sample sheets
   - Metadata
   - Peaksets from peak callers
   - `data(tamoxifen_peaks)`
2. Occupancy analysis
   - Overlap venns
   - Overlap rate
   - Consensus peaksets
3. Read counting
   - BAM/SAM/BED
   - Scores (RPKM)
   - Filtering
   - `data(tamoxifen_counts)`
4. DBA
   - Contrasts
     - GLMs
     - Multi-factor designs (paired, blocking)
   - Normalisation
     - Subtract control reads
     - Library size: full vs. effective
     - e.g. TMM (edgeR)
   - DE Method (edgeR, DESeq)
   - `data(tamoxifen_analysis)`
5. Plotting and reporting
   - Retrieving DB sites, stats, counts
   - MA plots
   - Heatmaps (correlation, affinity), PCA, boxplots

# Acknowledgements


- CRUK-CI Bioinformatics Core
  - Matthew Eldridge
  - Tom Carroll
  - Suraj Menon
- **Gordon Brown (DiffBind)**
- **Jason Carroll** and his laboratory
  - Caryn Ross-Innes
  - Vasiliki Therodorou


UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

# NOW LET'S GET OUR HANDS DIRTY….

www.cruk.cam.ac.uk

Rory Stark

**Rory.stark@cruk.cam.ac.uk**

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE