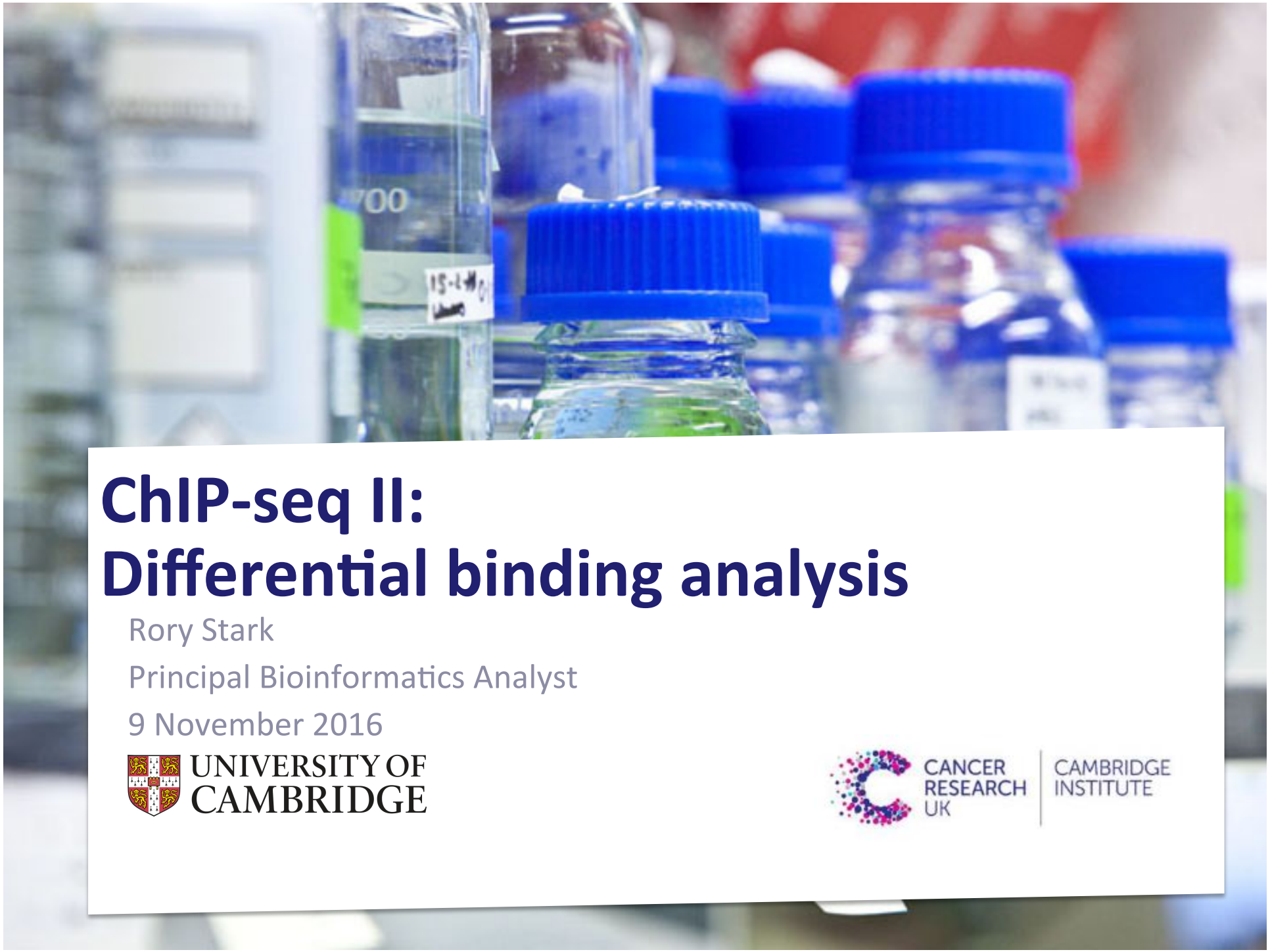


ChIP-seq II: Differential binding analysis

Rory Stark

Principal Bioinformatics Analyst

9 November 2016

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

ChIP-seq for functional genomics

Most ChIP-Seq studies to date have focused on **mapping**, not **function** (cf ENCODE)

- Comparisons limited to peak overlaps (co-occupancy)
- Limited quantitative analysis

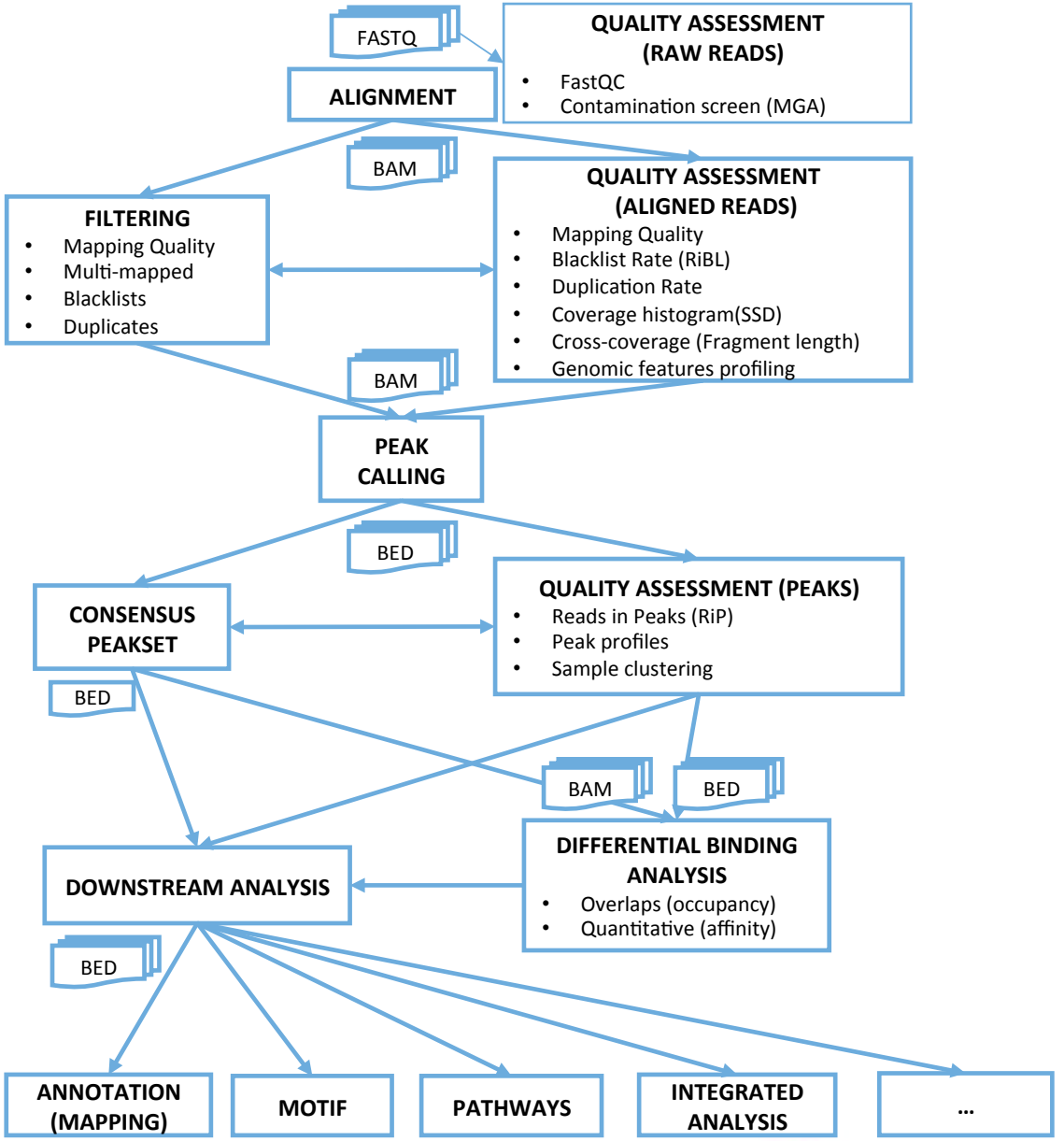
Most **functional** studies to date have focused on RNA levels

- Well established design/analysis
- Unable to directly distinguish driver/upstream from passenger/ downstream changes
- Regulatory schema **inferred** (knockouts, modelling)

Can we use ChIP-Seq to more directly **observe** regulatory events?

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

ChIP-seq analysis workflow

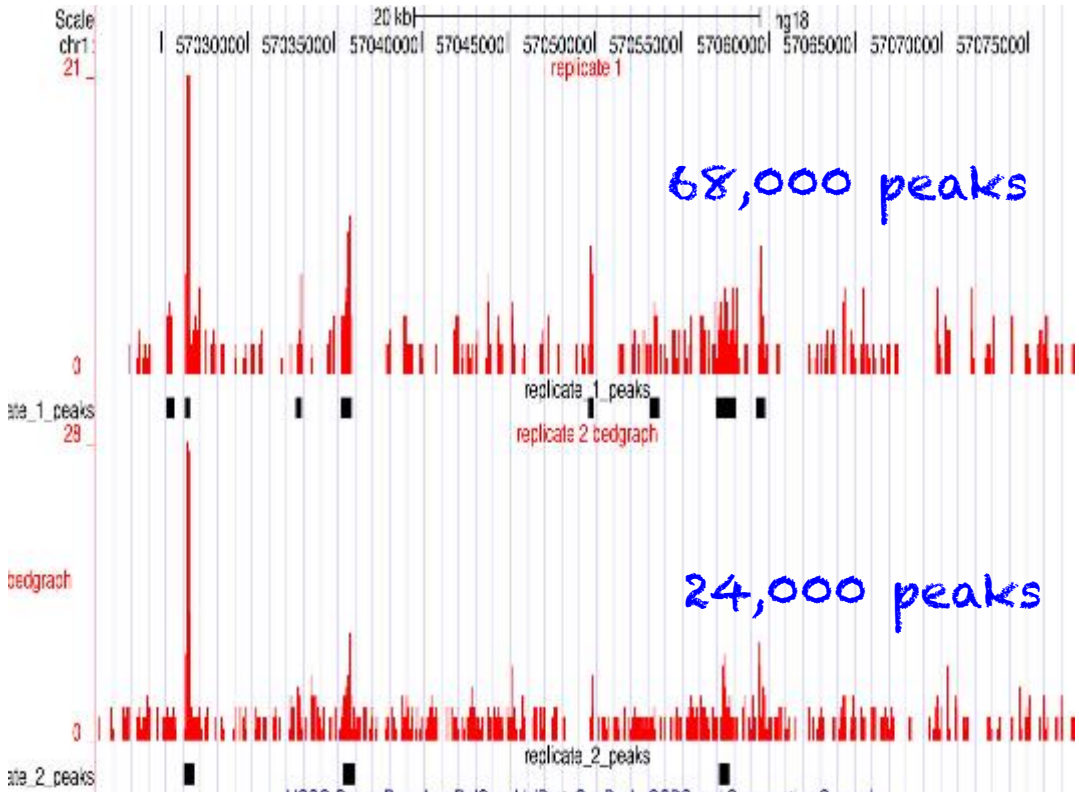


CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

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 affinity)
  - Binding site count density (ChipDiff, DiffBind)
  - Binding profile (MMDiff)
  - Sliding windows (csaw)
  - Etc.

Case Study: ER binding in breast cancer

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

NATURE | LETTER

previous article next article

Differential oestrogen receptor binding is associated with clinical outcome in breast cancer

Caryn S. Ross-Innes, Rory Stark, Andrew E. Teschendorff, Kelly A. Holmes, H. Raza Ali, Mark J. Dunning, Gordon D. Brown, Ondrej Gojis, Ian O. Ellis, Andrew R. Green, Simak Ali, Suet-Feung Chin, Carlo Palmieri, Carlos Caldas & Jason S. Carroll

Affiliations | Contributions | Corresponding authors

Nature **481**, 389–393 (19 January 2012) | doi:10.1038/nature10730

Received 19 May 2011 | Accepted 23 November 2011 | Published online 04 January 2012

Functional genomics of breast cancer

- Tumors cluster into subtypes based on gene expression
- 70% of tumors over-express primary prognostic marker ER
- ER+ tumors respond to hormone and/or tamoxifen treatment
- Two secondary prognostic markers: PR and HER2
- Prognostic gene expression signatures readily derivable from expression data

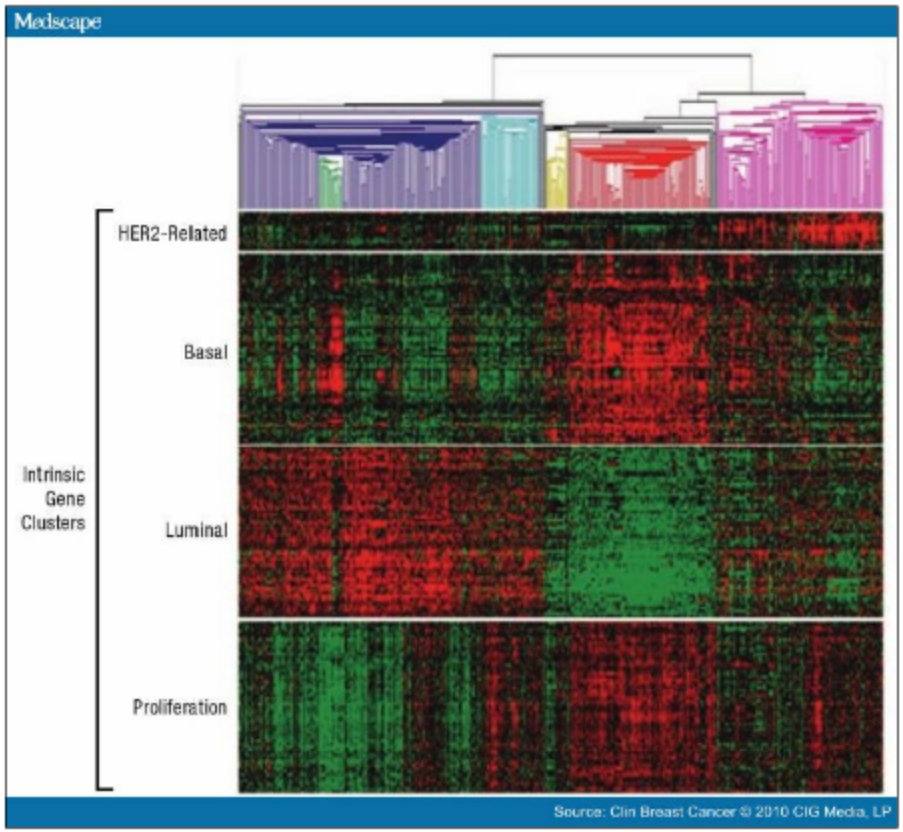


Figure 1.

Semi-Unsupervised Gene Expression Array Analysis of a Cohortof Breast Cancers Identifies Several Intrinsic SubtypesFigure

Shown are luminal A (outlined in dark blue), luminal B (pale blue), HER2-enriched (pink), basal-like (red), claudin-low (yellow), and normal-like (green) tumors. Heat map courtesy of CM Perou.

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

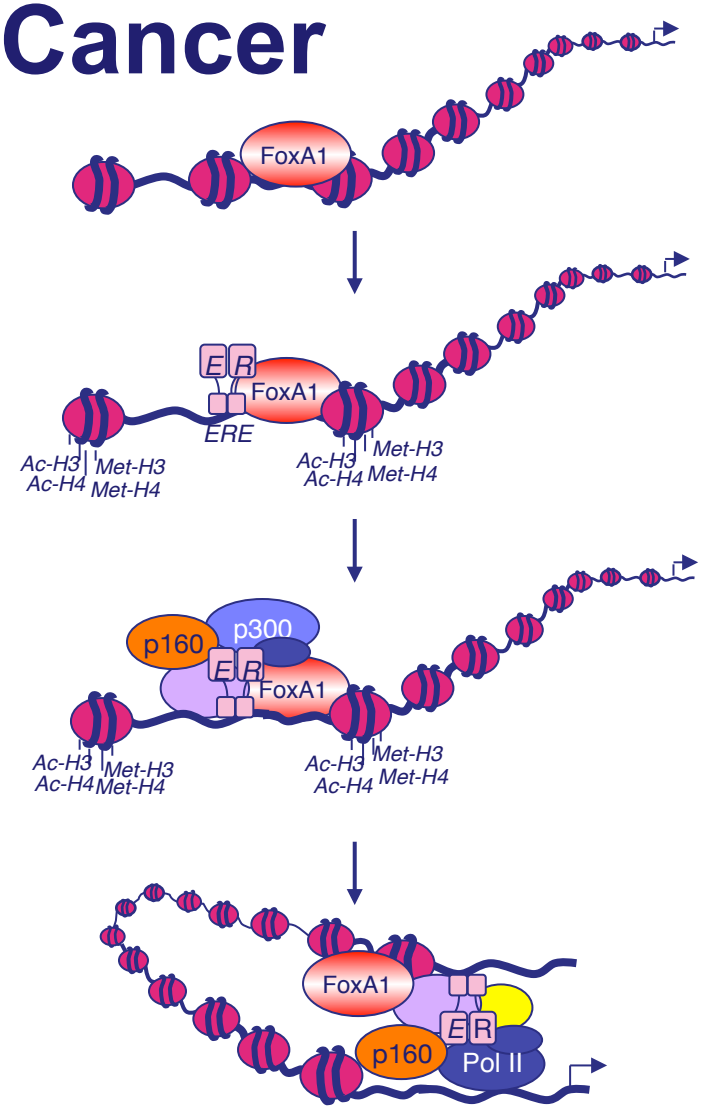
ER Binding in Breast Cancer

ER part of ERE (estrogen response element) regulatory complex with estrogen (E2) and factors

E2 binds ER, then ER-E2 complexes dimerize, bind to DNA at ERE after pioneer factor (e.g. FoxA1)

Most ER binding is intergenic

Many other co-factors in different combinations



CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Example : tamoxifen resistance in BC cell lines

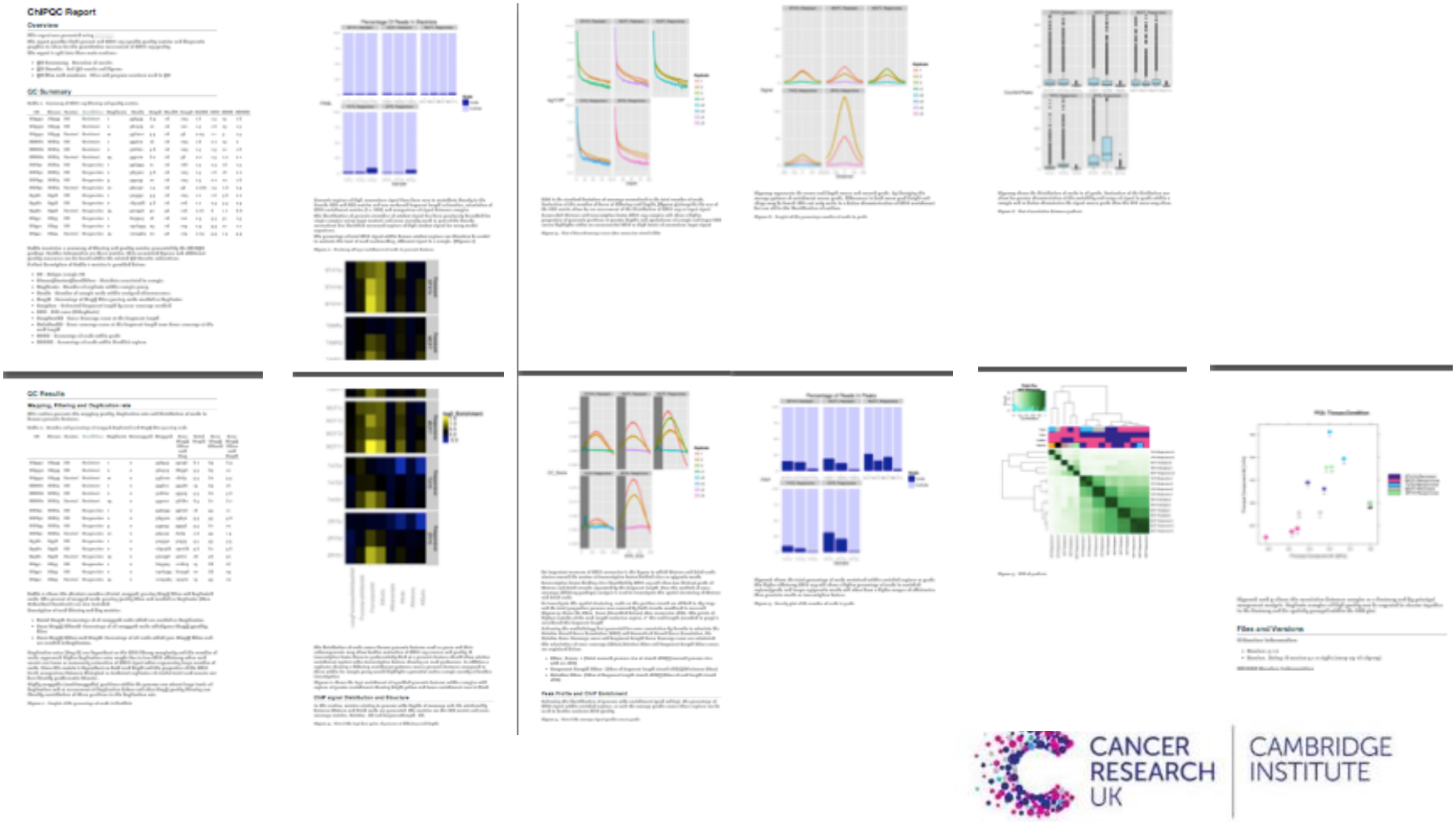
Example: identifying ER binding sites likely to be functional in tamoxifen resistance

Sample	Tissue	Factor	Status	Rep#	Peaks
MCF71	MCF7	ERα	Responsive	1	74,029
MCF72	MCF7	ERα	Responsive	2	49,075
MCF73	MCF7	ERα	Responsive	3	67,130
T47D1	T47D	ERα	Responsive	1	28,713
T47D1	T47D	ERα	Responsive	2	23,575
ZR751	ZR75	ERα	Responsive	1	74,971
ZR752	ZR75	ERα	Responsive	2	70,560
MCF7r1	MCF7	ERα	Resistant	1	47,034
MCF7r2	MCF7	ERα	Resistant	2	52,517
BT4741	BT474	ERα	Resistant	1	41,924
BT4742	BT474	ERα	Resistant	2	40,783

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

ChIPQC report for example dataset

http://starkhome.com/ChIPQC/Reports/tamoxifen/ChIPQC.html

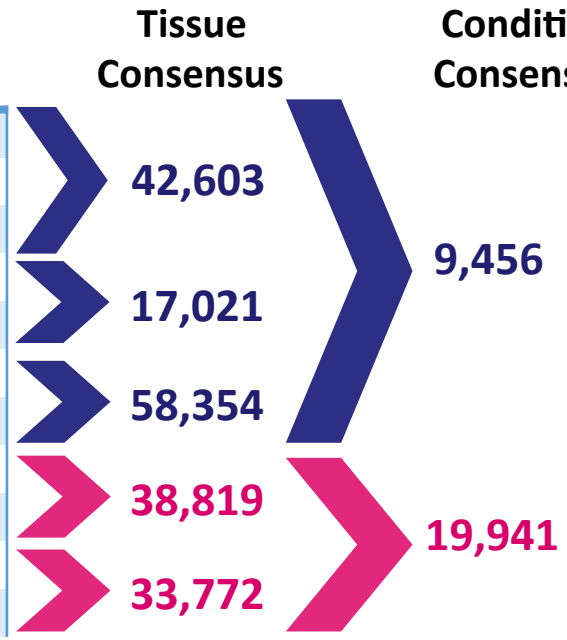


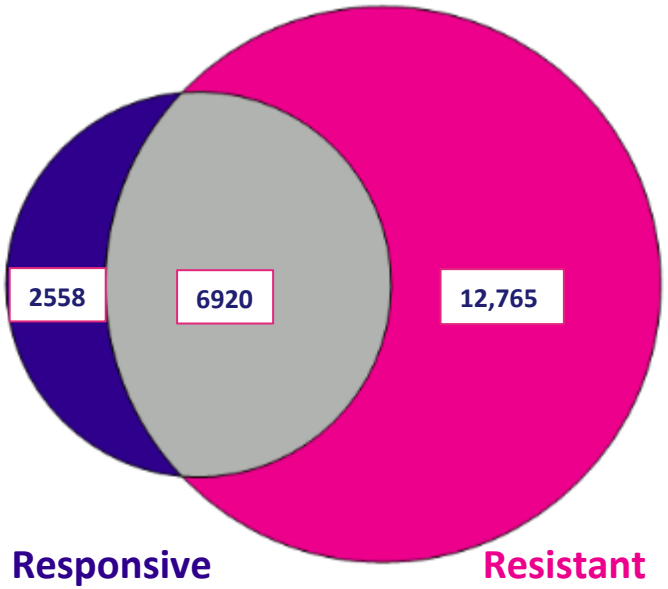
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Occupancy Analysis

Occupancy Analysis: Strict Consensus Peaks

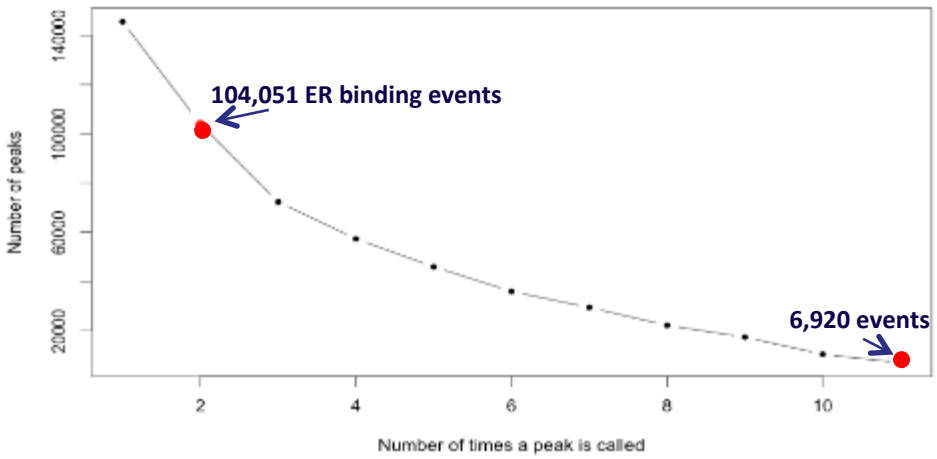
Tissue	Status	Rep#	Peaks
MCF7	Responsive	1	74,029
MCF7	Responsive	2	49,075
MCF7	Responsive	3	67,130
T47D	Responsive	1	28,713
T47D	Responsive	2	23,575
ZR75	Responsive	1	74,971
ZR75	Responsive	2	70,560
MCF7	Resistant	1	47,034
MCF7	Resistant	2	52,517
BT474	Resistant	1	41,924
BT474	Resistant	2	40,783

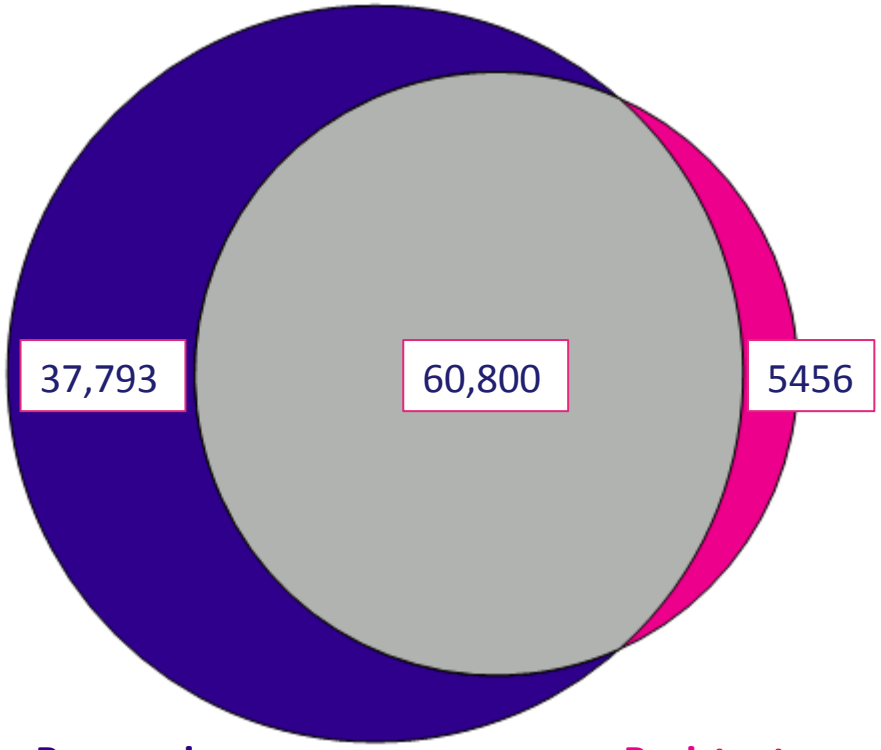




CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Occupancy Analysis: Lenient Consensus Peaks



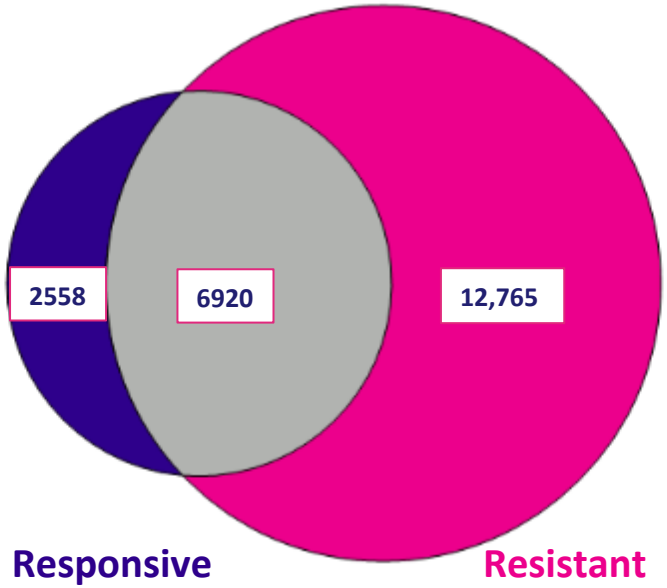


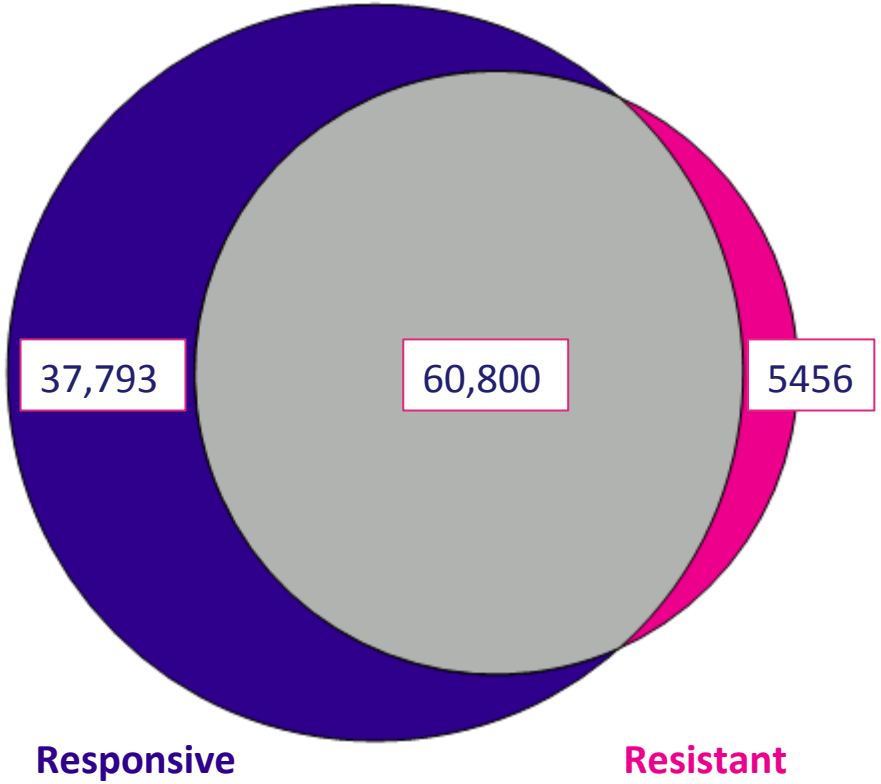
104,051 peaks identified in at least 2 samples

- **Responsive** only:
  - >= 2 Responsive samples
  - <2 Resistant samples
- **Resistant** only:
  - >= 2 Resistant samples
  - <2 Responsive samples

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

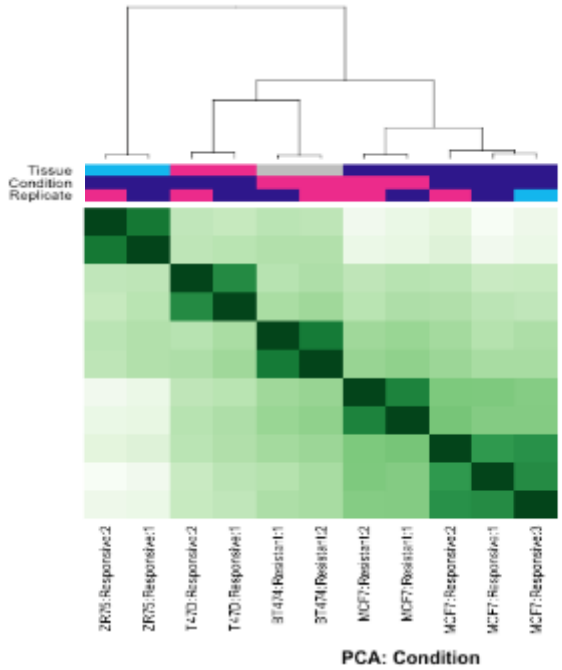
Occupancy Analysis:
Strict vs Lenient Consensus Peaks

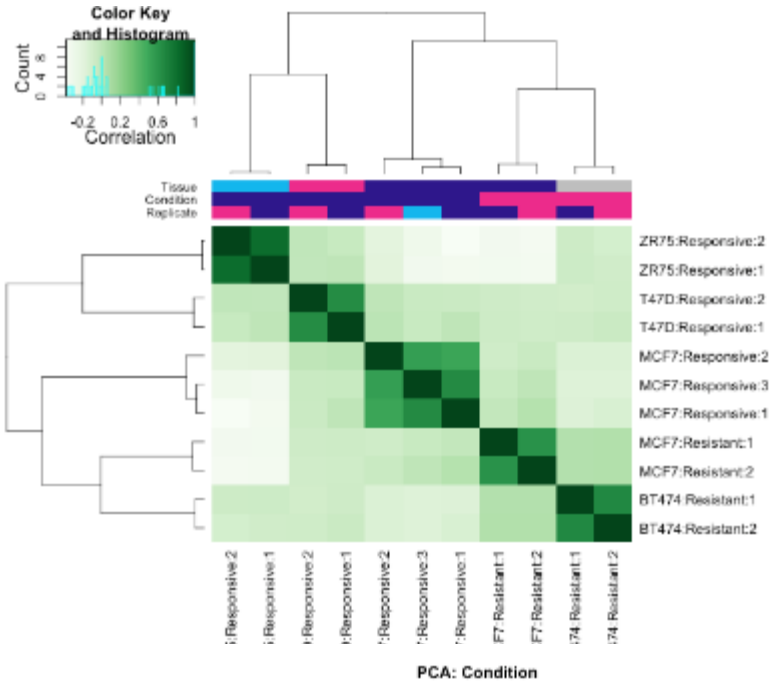


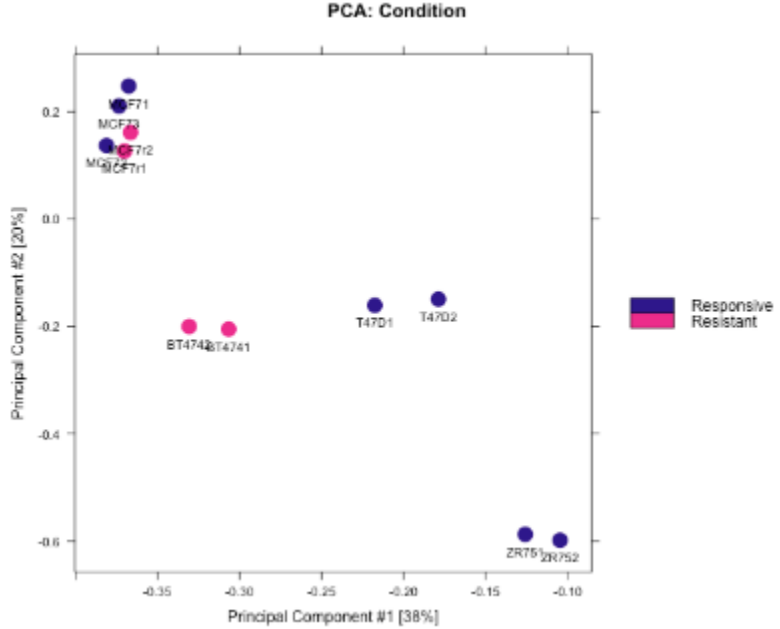


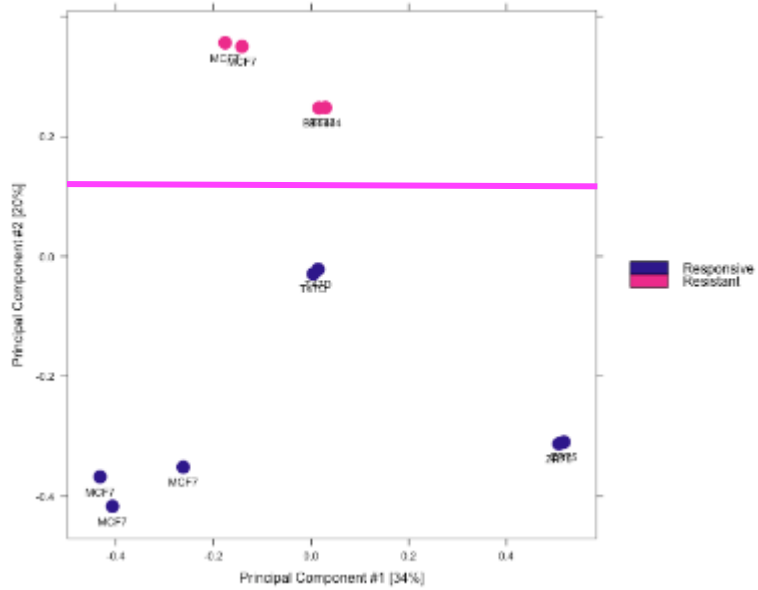
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Occupancy Clustering









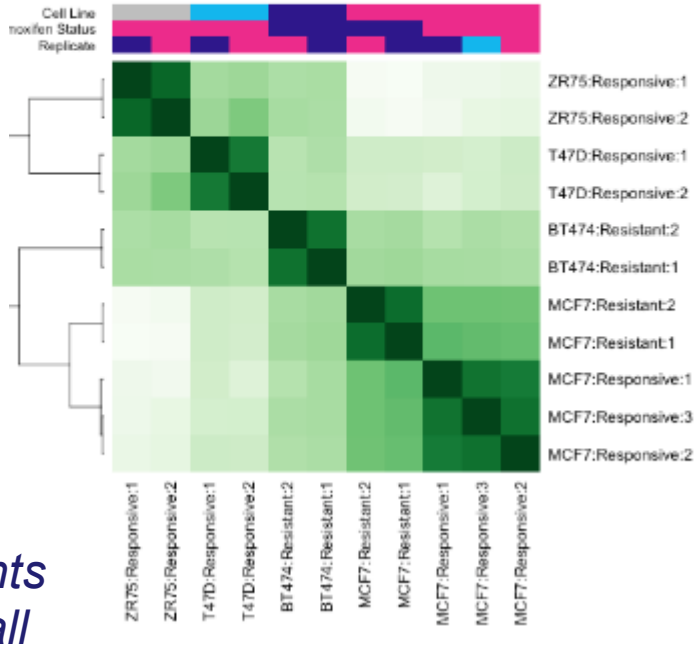
Quantitative Analysis

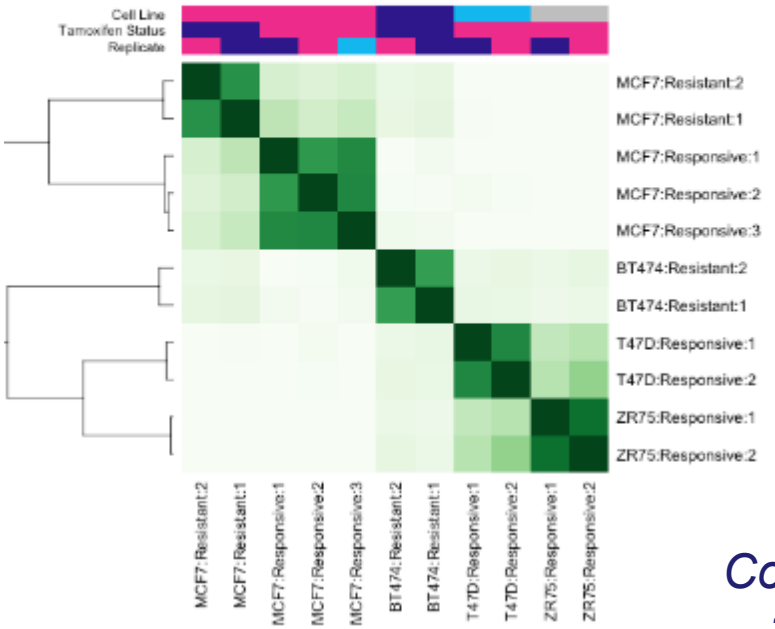
Binding affinity matrix

- **Rows**: decide interval (binding site) "universe"
  - Peak callers -> occupancy/overlap consensus
    - 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?
  - Normalisation?

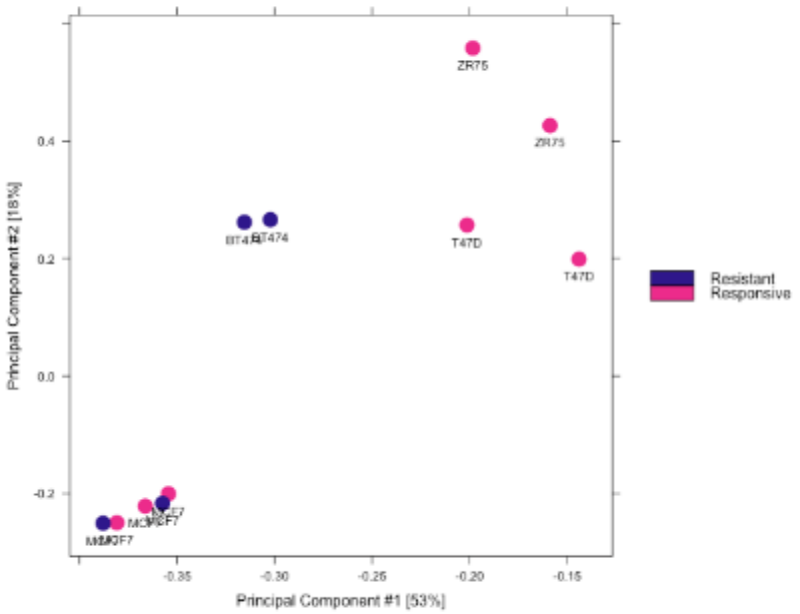
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Affinity (count) clustering

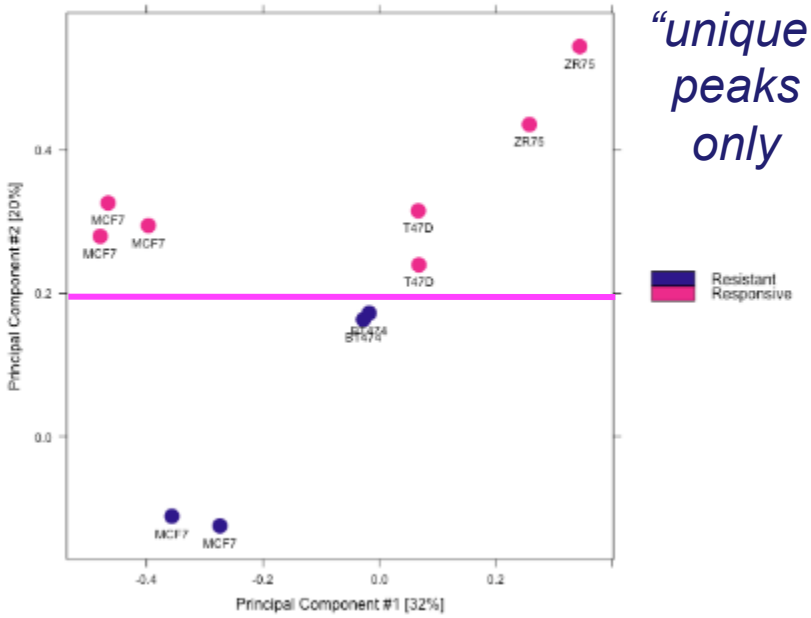




Counts for all peaks

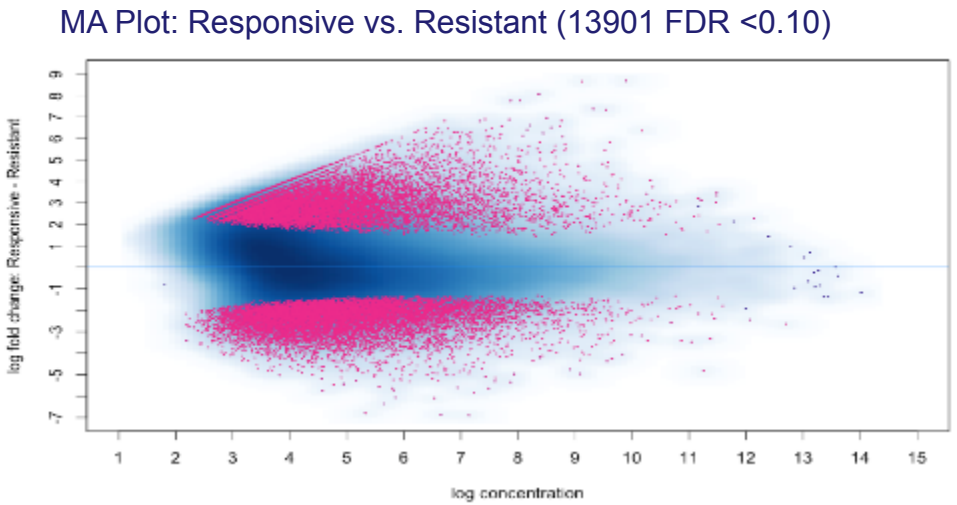


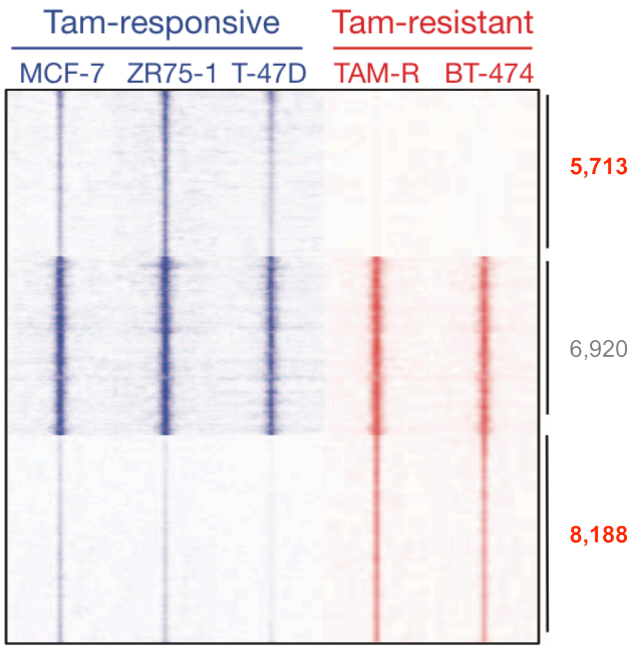
Counts for "unique" peaks only



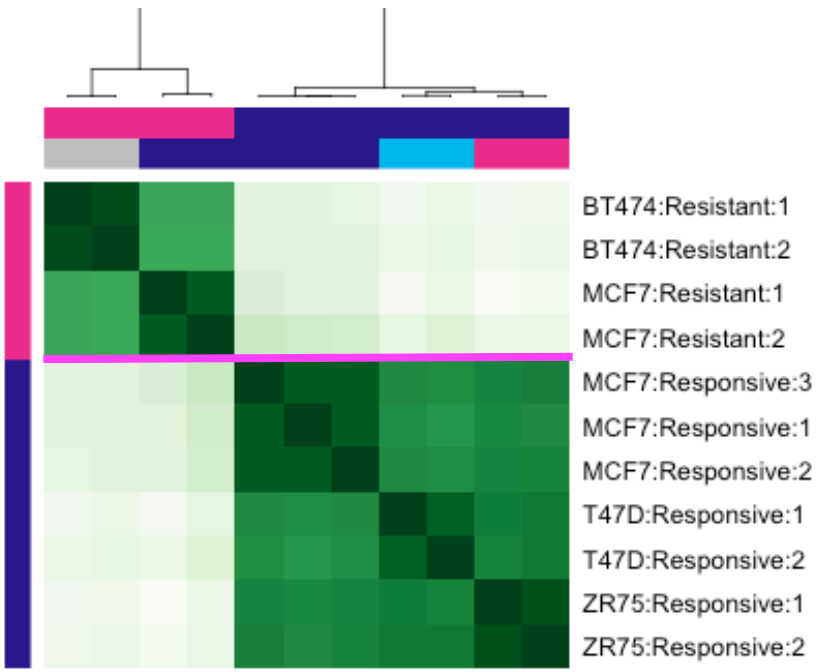
Differential binding analysis

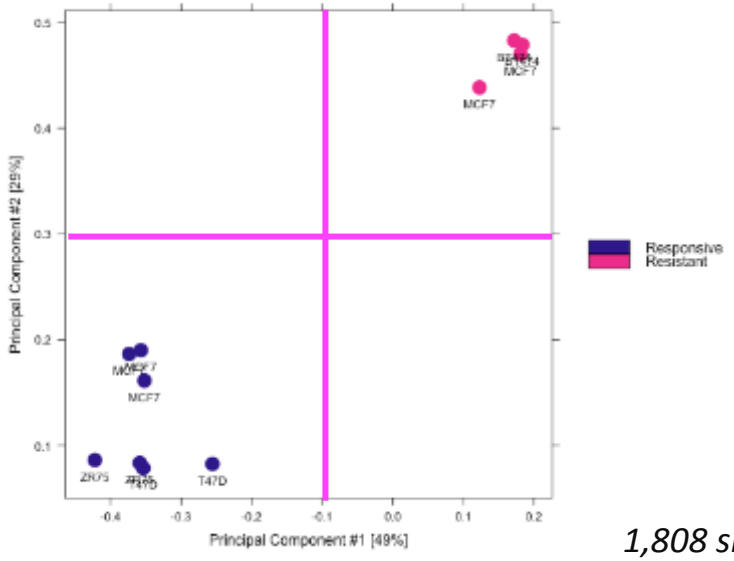
- Determine contrasts
  - Single-factor
  - Multi-factor (GLM/blocking)
    - Matched tumour-normal
    - Common tissue
    - Replicate groups (batch)
- Run count-based RNA-Seq DE package
  - edgeR, DESeq2, etc.
  - Fit negative binomial distribution
  - Exact test
  - Multiple testing correction (B&H FDR)

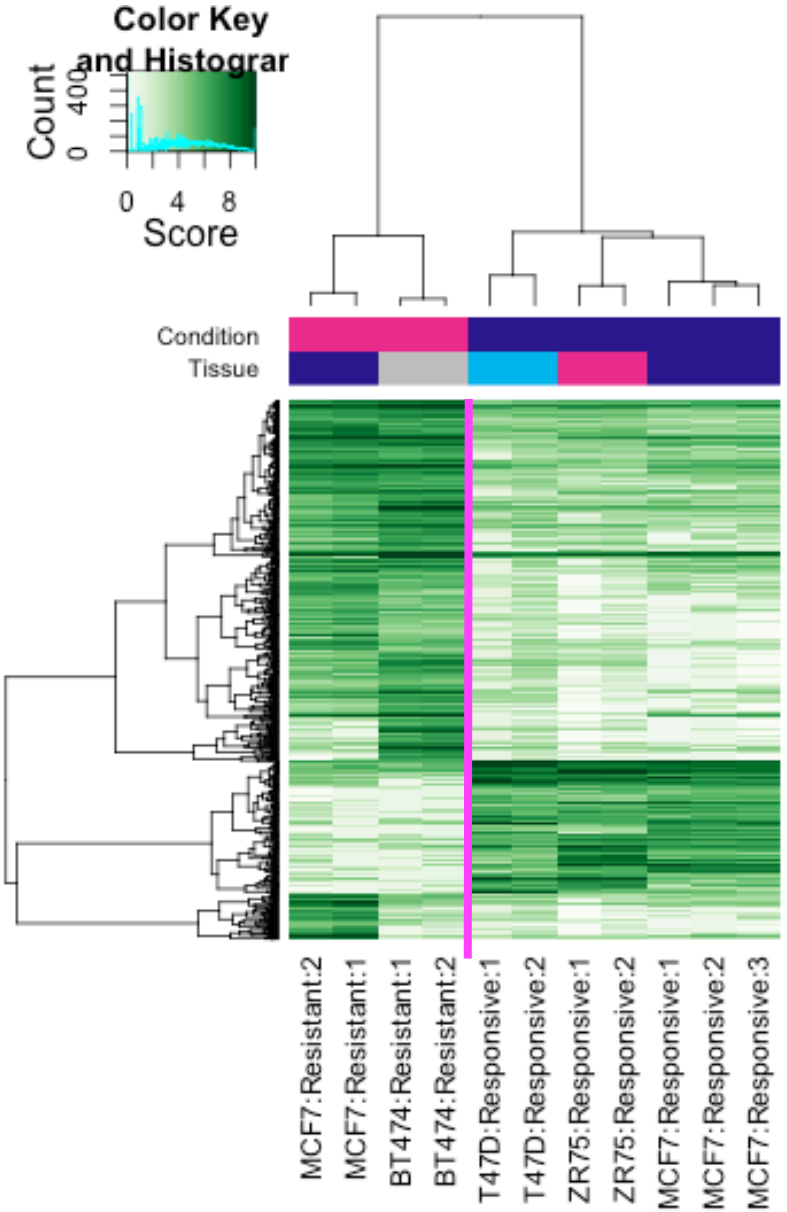




CANCER RESEARCH UK | CAMBRIDGE INSTITUTE



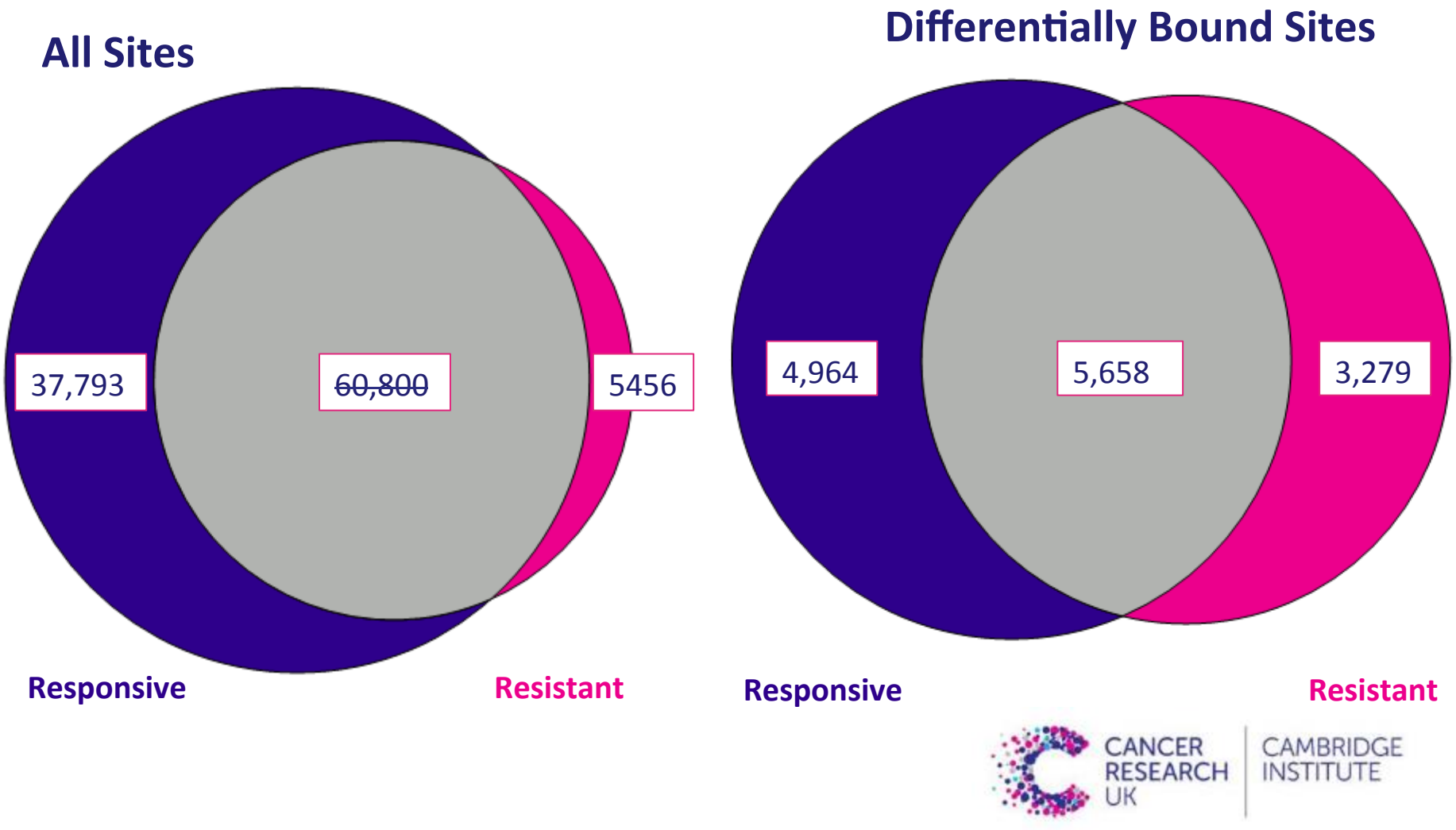




1,808 sites differentially bound at FDR <= 0.005

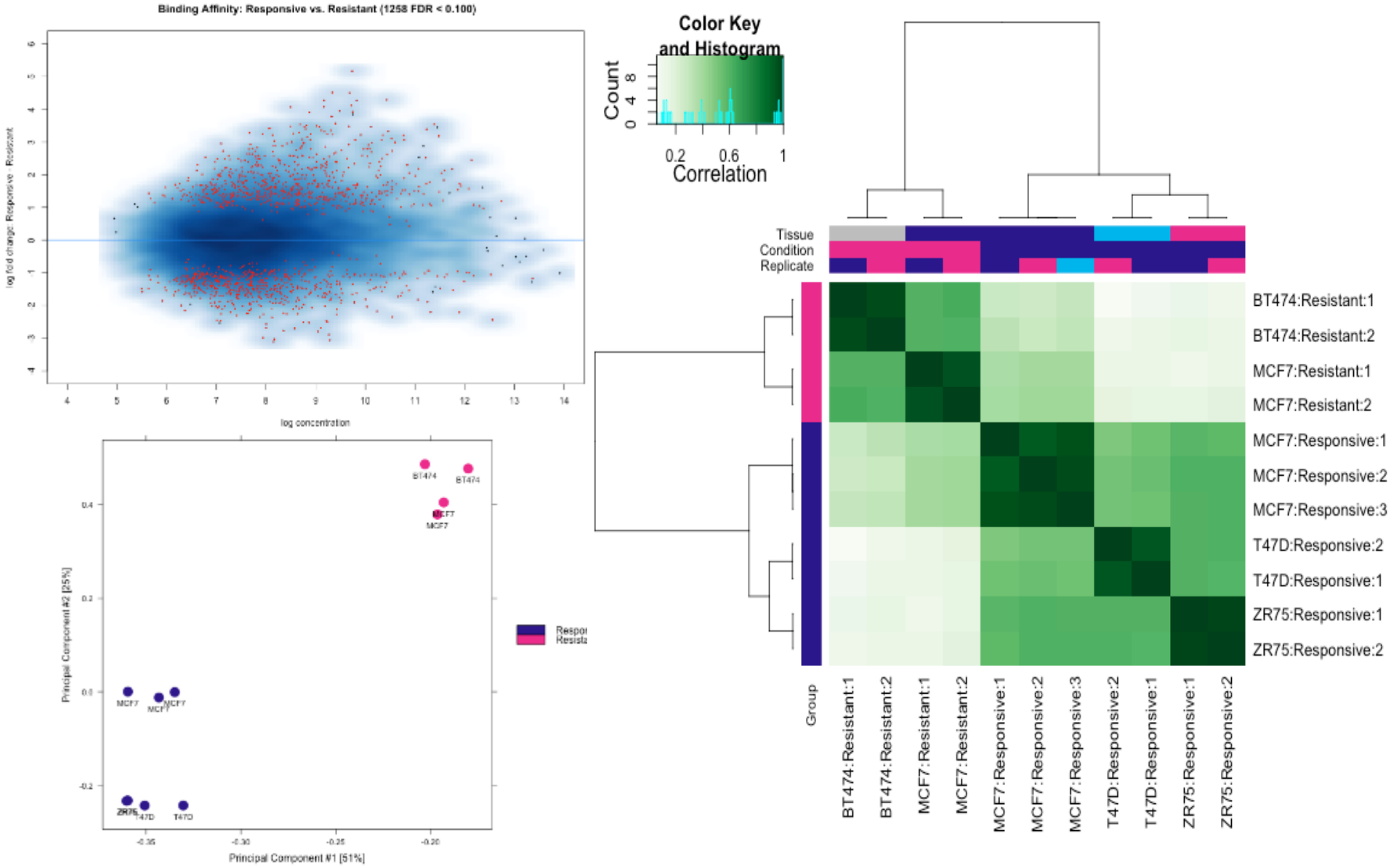
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Differential binding analysis: Occupancy vs. Affinity



CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Differential binding signature isolated from 6920 sites common to all samples



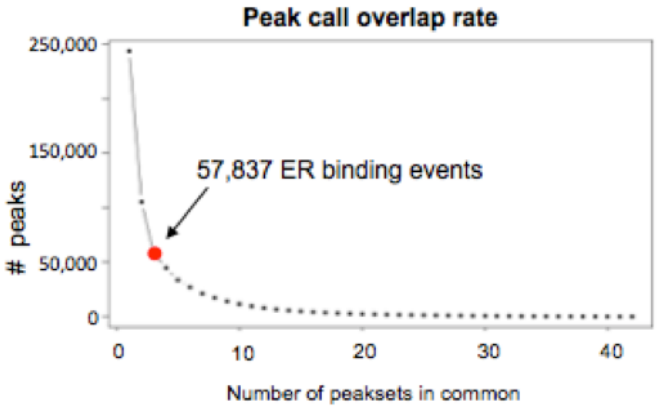
Result: Predicting outcome from tumour samples

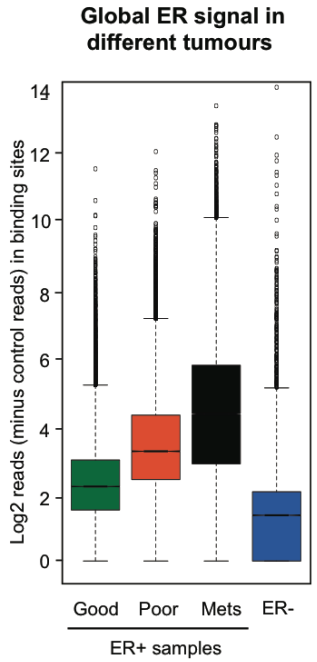
UNIVERSITY OF CAMBRIDGE

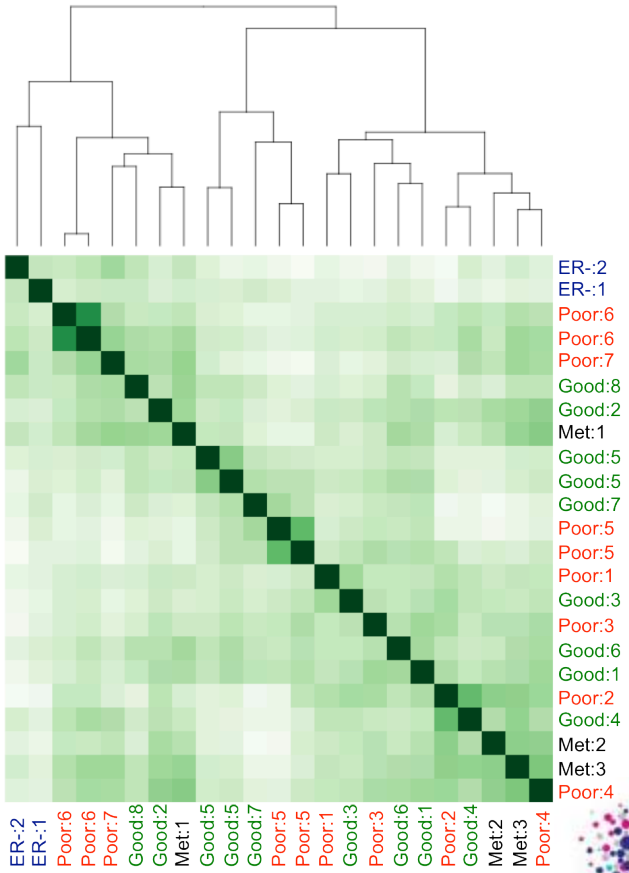
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

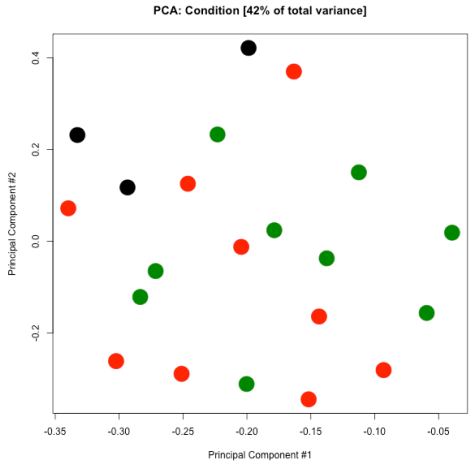
ER ChIP-seq in clinical samples

- 20 BC tumours
  - 18 ER+, 2 **ER-**
  - 15 primary, 3 metastases
  - 3 sampled in replicate
  - Additional controls: 3 normal breast, 2 normal liver
- Two peak callers
  - 42 peaksets
- **Good**/**poor** prognosis based on PR/HER2 status







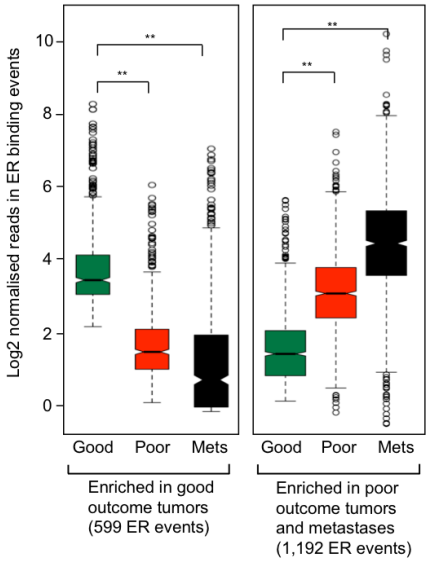


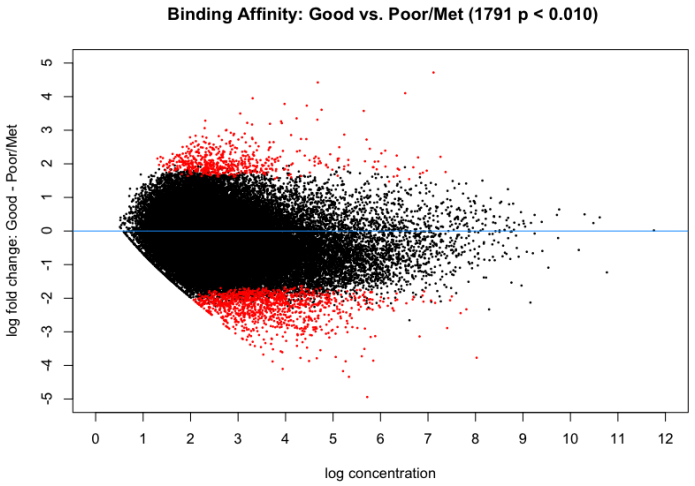
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

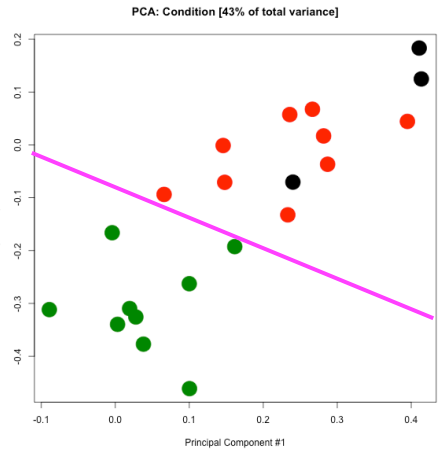
Differentially bound sites separate tumours by prognosis

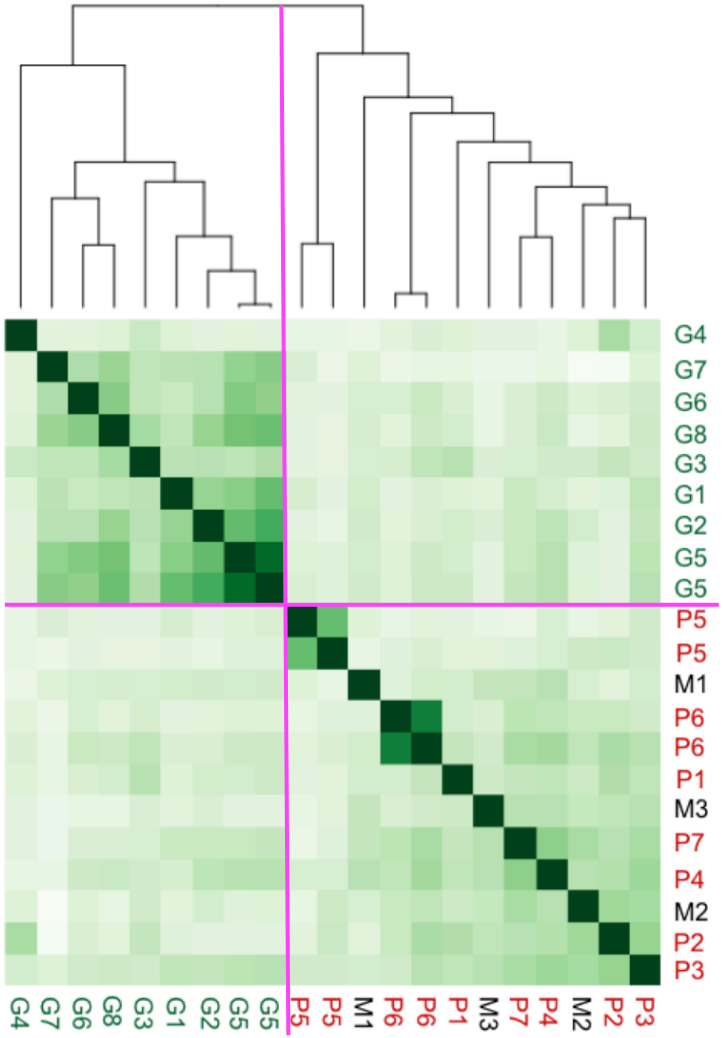
1,791 sites identified as differentially bound between good and poor prognosis

- **599** enriched in good prognosis
- **1,192** enriched in poor pronosis



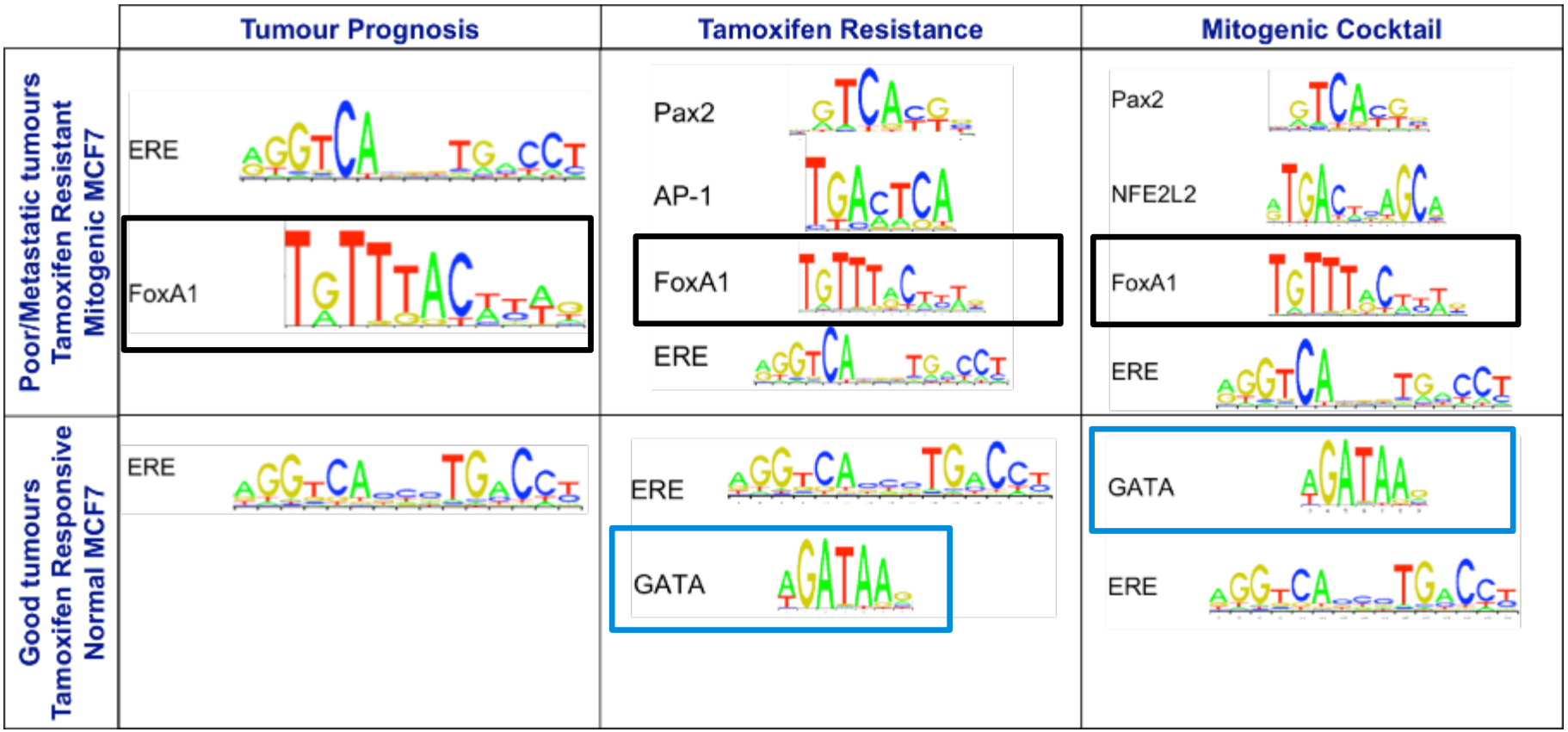






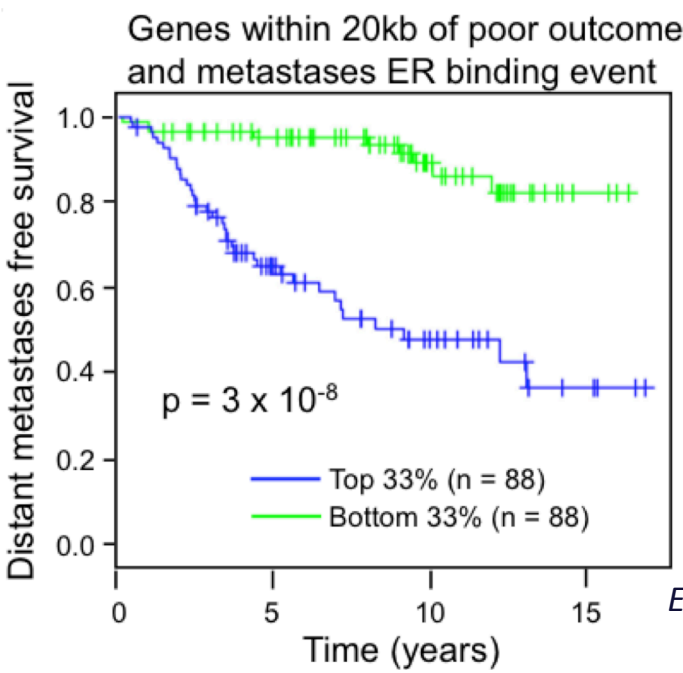
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Differentially enriched co-factor motifs

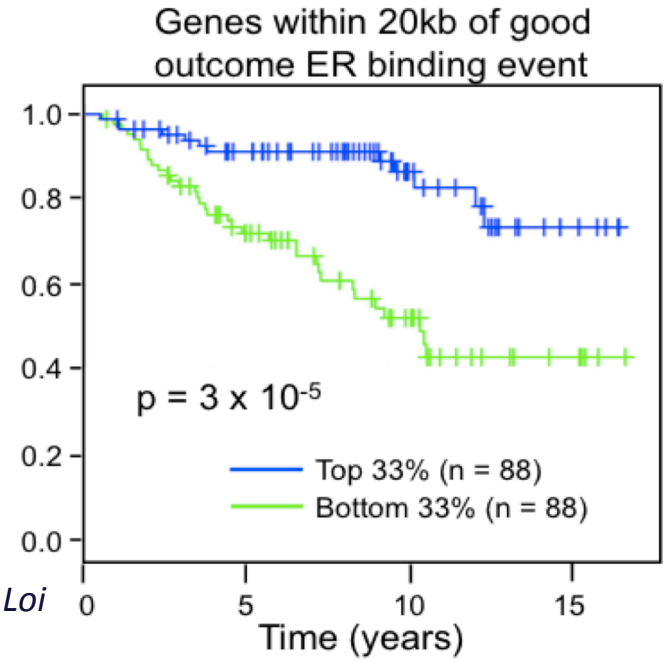


CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Genes near DB sites form prognostic gene signatures



Expression data: Loi et al.



- Signature composed of genes within 20kb of DB sites
  - **265** genes in Poor outcome signature
  - **109** genes in Good outcome signature
- Classifier based on up/down regulation in mRNA expression sets
- Validated in 7 publicly available BC expression datasets

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

DiffBind


UNIVERSITY OF
CAMBRIDGE


CANCER
RESEARCH
UK
CAMBRIDGE
INSTITUTE

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home | Install

Home » Bioconductor 3.3 » Software Packages » DiffBind

DiffBind

platforms all | downloads top 5% | posts 28 / 1 / 1 / 4 | in Bioc 5 years
build ok | commits 2.50 | test coverage 8%

Differential Binding Analysis of ChIP-Seq peak data

```
> library(DiffBind)
> tamoxifen <- dba(sampleSheet="tamoxifen.csv")
> tamoxifen <- dba.count(tamoxifen, summits=250)
> tamoxifen <- dba.contrast(tamoxifen, categories=DBA_CONDITION)
> tamoxifen <- dba.analyze(tamoxifen)
> tamoxifen.DB <- dba.report(tamoxifen)
```

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

Conclusion

Functional analysis of genome-scale regulatory data from enrichment assays

- Focus primarily on differential *expression* limits ability to identify upstream/driver genes
- Direct study of differential *regulation* should result in gene signatures enriched for upstream events
- Categorization of differentially regulated genes helps identify co-regulators
- These analysis techniques can be applied to epigenomic regulatory data in general

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

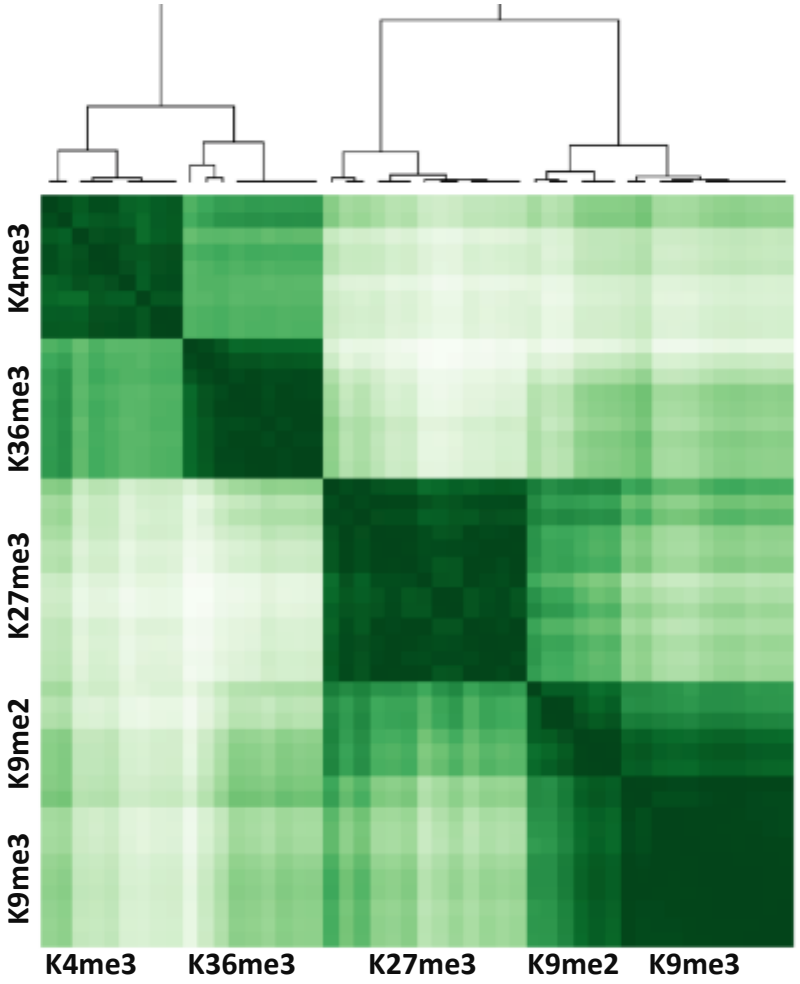
Acknowledgements

- **CRUK-CI Bioinformatics Core**
  - Matthew Eldridge
  - Suraj Menon
  - Thomas Carroll (**ChipQC**, now MRC Clinical Sciences Centre)
- **Jason Carroll lab**
  - Caryn Ross-Innes
  - Vasiliki Therodorou
  - **Gordon Brown (DiffBind)**
- **Masashi Narita lab**
  - Tamir Chandra (now Babraham)

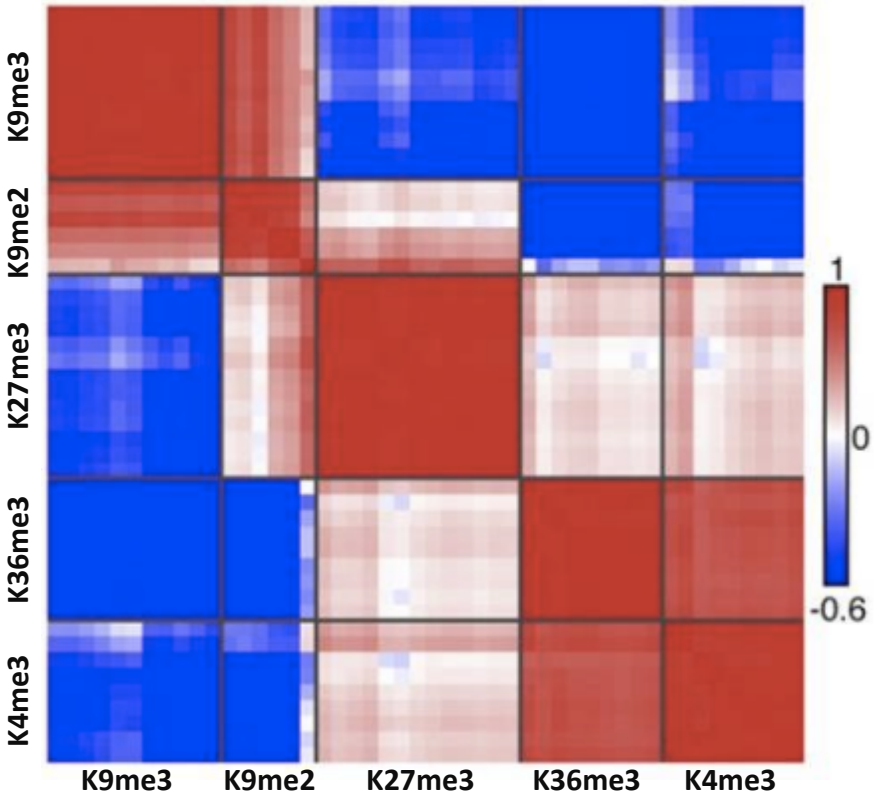
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Backup slides

Count-based correlation clustering



5kb windows around TSS



1Mb windows across genome

Data from Chandra et al Molecular Cell 2012

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE