

An aerial, top-down view of a large, diverse group of people of various ages, ethnicities, and clothing styles, scattered across a white background. The people are looking in different directions, some standing, some sitting, creating a sense of a busy, multi-cultural gathering.

Associating differential ER binding with clinical outcome in breast cancer

RORY STARK
18 JULY 2013

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?

Agenda

- Differential ER binding in breast cancer: Overview of results
 - Identification of differentially bound sites
 - Performance of prognostic signature
 - Downstream analysis (differential co-factor motifs)
- Method: Differential binding analysis
 - Occupancy analysis
 - Quantitative analysis
 - Bioconductor package: DiffBind



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

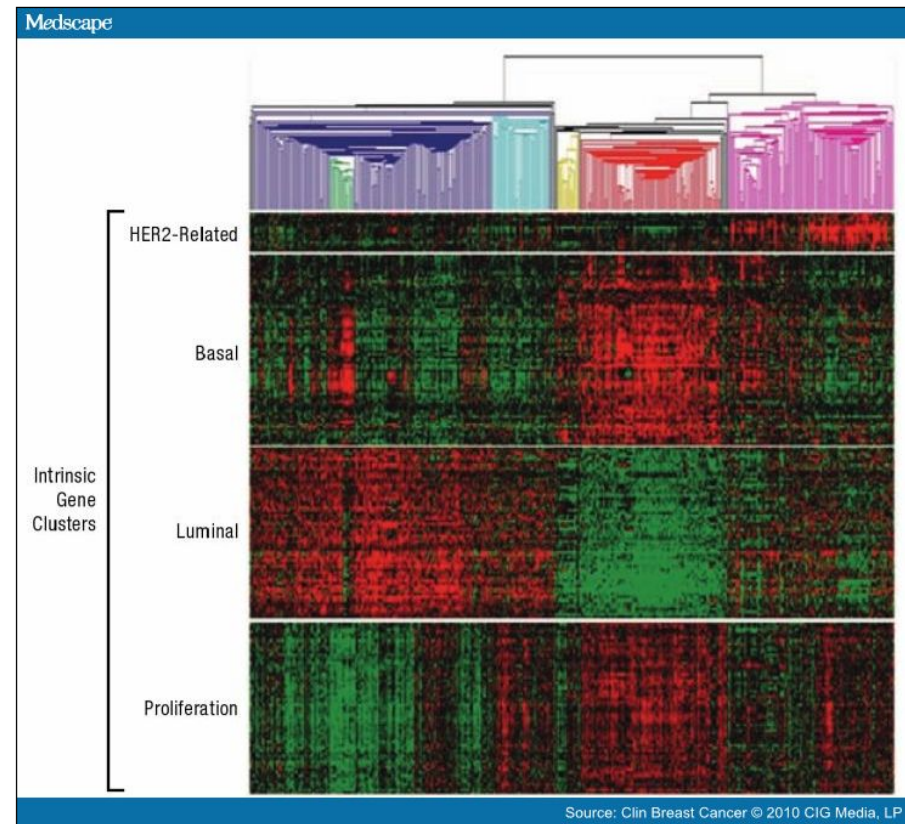


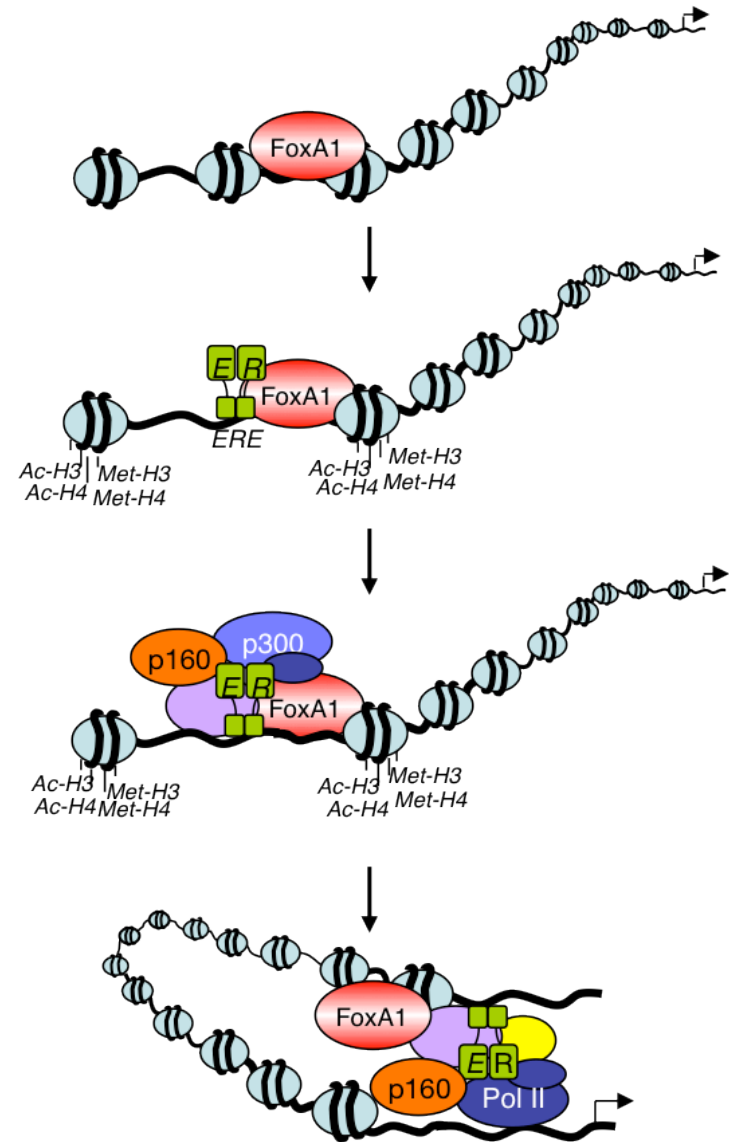
Figure 1.

Semi-Unsupervised Gene Expression Array Analysis of a Cohort of Breast Cancers Identifies Several Intrinsic Subtypes

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.

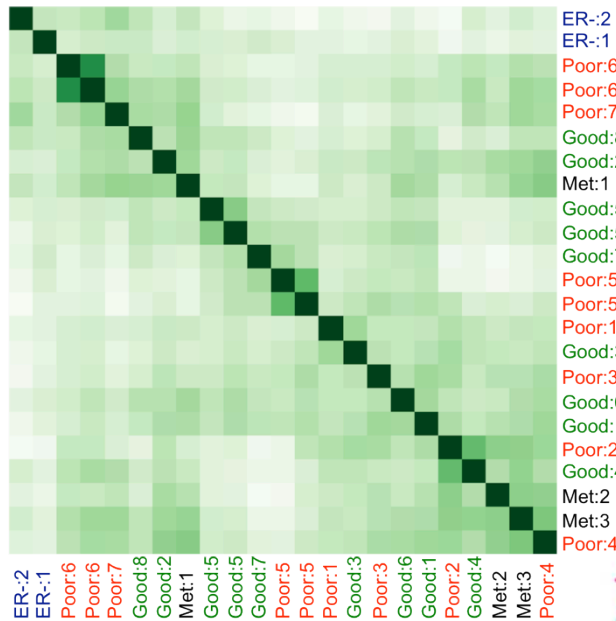
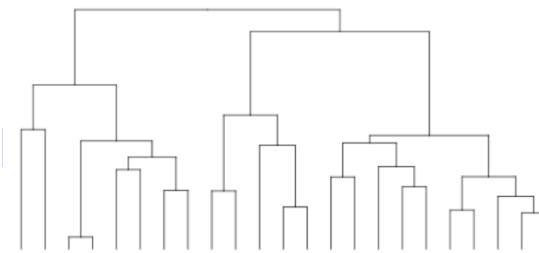
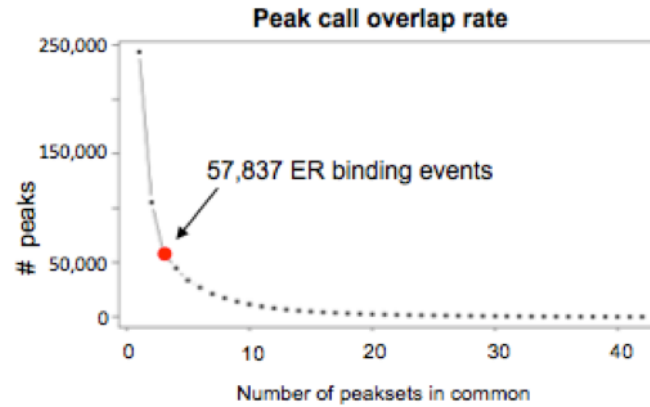
ER binding in breast cancer

- ER is a transcription factor
- ERE (estrogen response element) regulatory complex
 - E2 binds ER
 - ER-E2 complexes dimerize
 - Pioneer factor (e.g. FoxA1) opens chromatin
 - ER-E2 dimers bind to DNA at ERE
 - Other TF factors co-bind at ERE
- Most ER binding is intergenic (enhancers, not promoters)
- Evidence of DNA looping
- Previously, all genomic ER binding data derived from a single cell line (MCF-7)

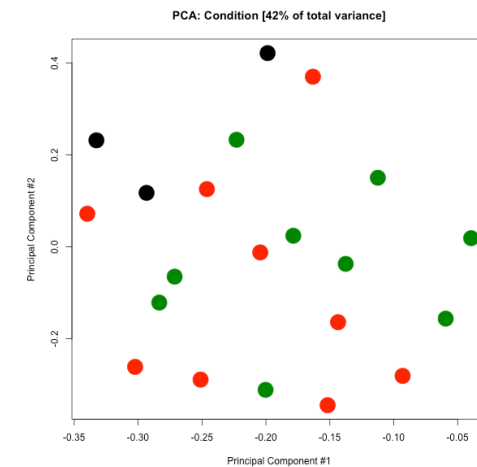
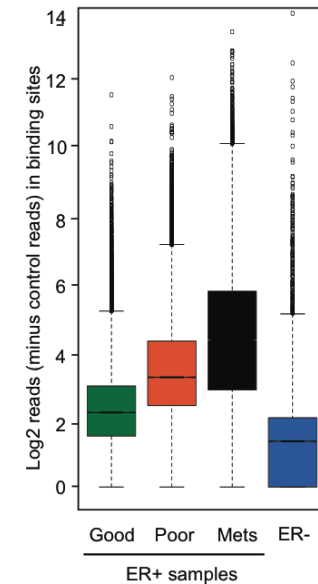


ER ChIP-seq in clinical samples

- 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 MACS/SWEMBL (42 peaksets)
- **Good/poor** prognosis based on PR/HER2 status

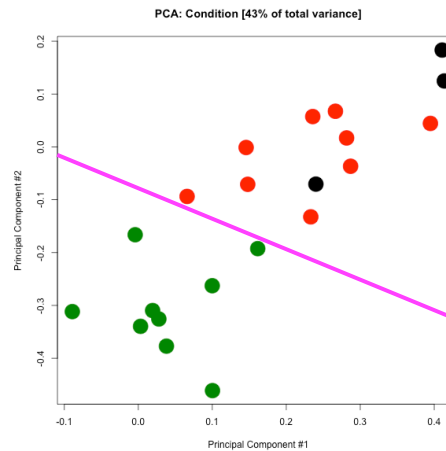
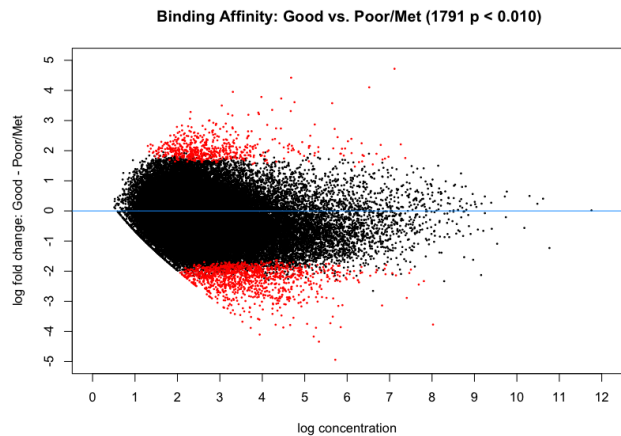
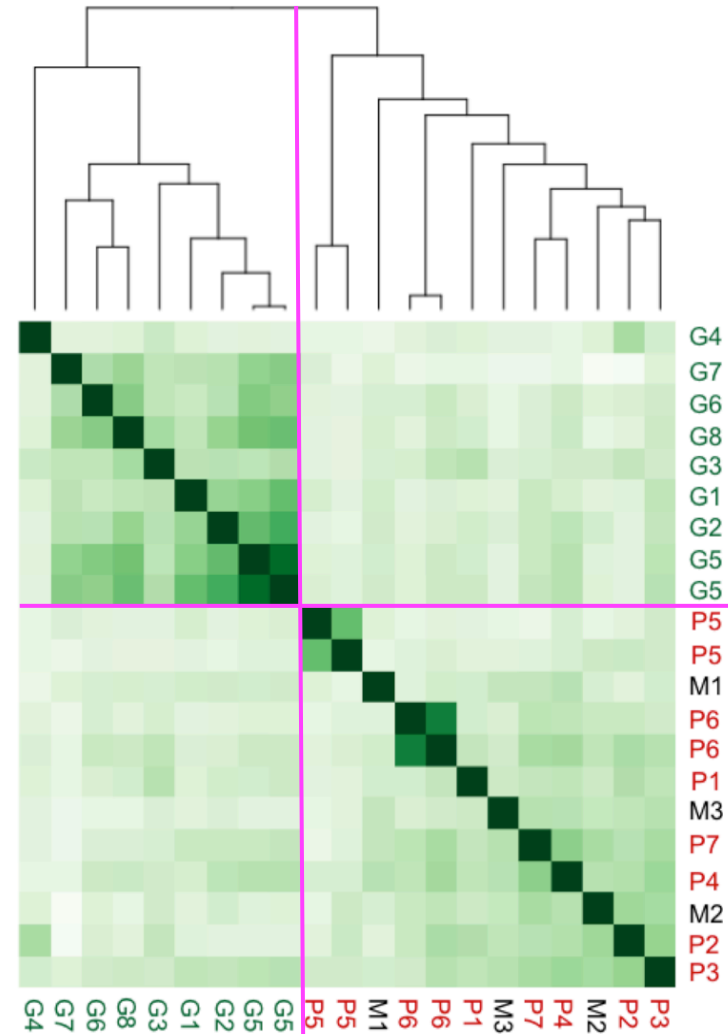
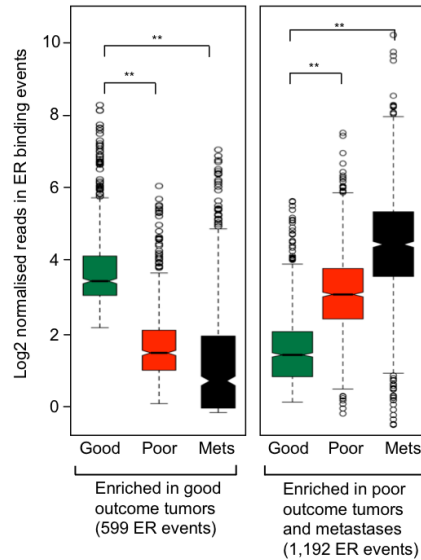


Global ER signal in different tumours

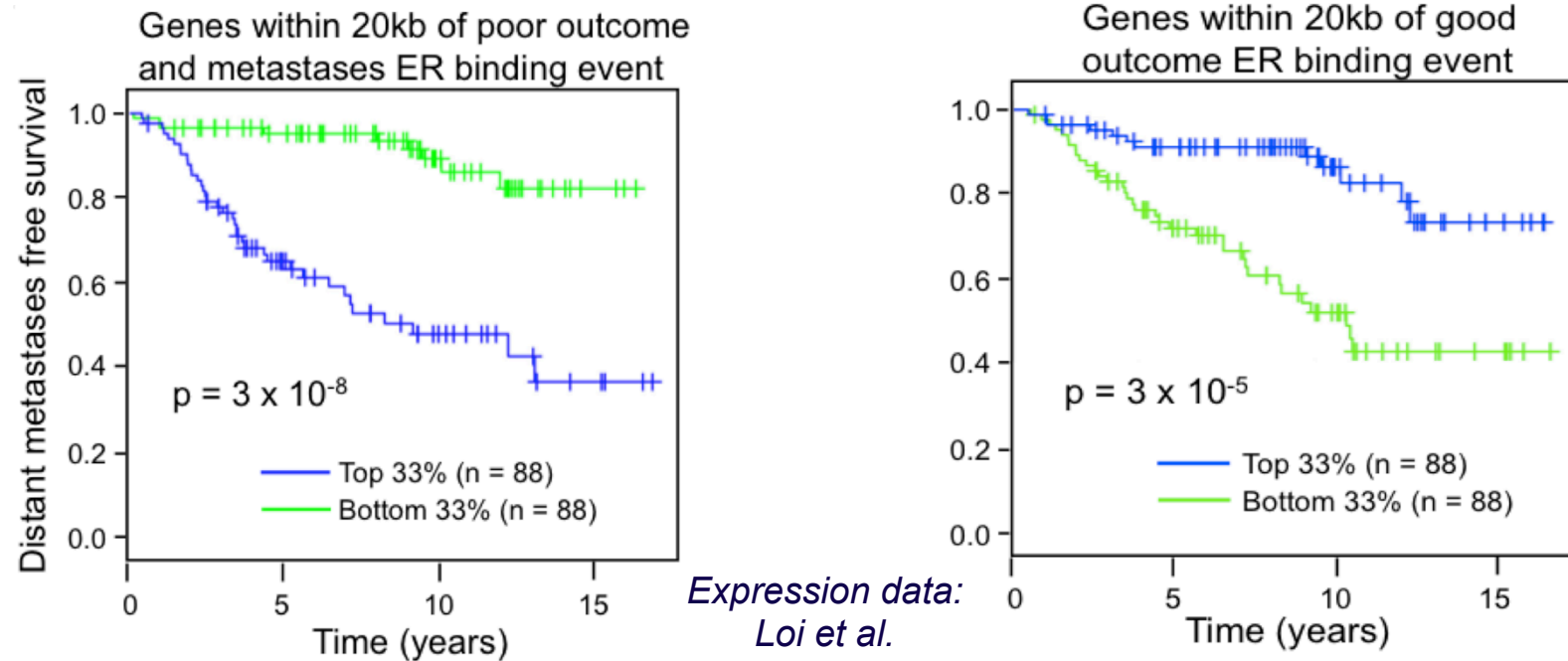


Differentially bound sites separate tumours by prognosis

- **1,791** sites identified as differentially bound between good and poor prognosis
 - **599** enhanced in good prognosis
 - **1,192** enhanced in poor prognosis/metastases

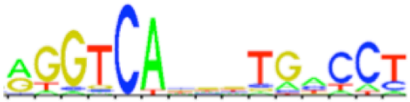


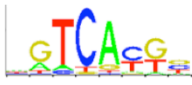
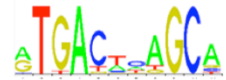

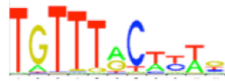

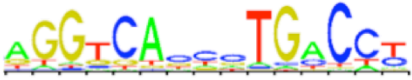
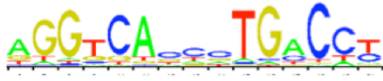
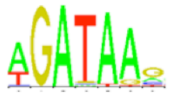
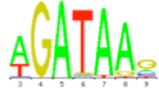
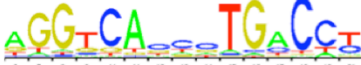


Genes near DB sites form prognostic gene signatures



- Signature composed of genes within 20k bases 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

Differentially enriched co-factor motifs

	Tumour Prognosis	Tamoxifen Resistance	Mitogenic Cocktail
Poor/Metastatic tumours Tamoxifen Resistant Mitogenic MCF7	ERE 	Pax2  AP-1 	Pax2  NFE2L2 
	FoxA1 	FoxA1 	FoxA1 
Good tumours Tamoxifen Responsive Normal MCF7	ERE 	ERE  GATA 	GATA  ERE 

Differential Binding Analysis

Differential binding analysis: Observations

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



Differential binding analysis: Goals

Be robust to noise

- Noisy experiments
- Noisy peak calling

Determine DB without defining global binding maps for each ChIP

Exploit quantitative **affinity** (read scores) beyond binary **occupancy** (peak calls)

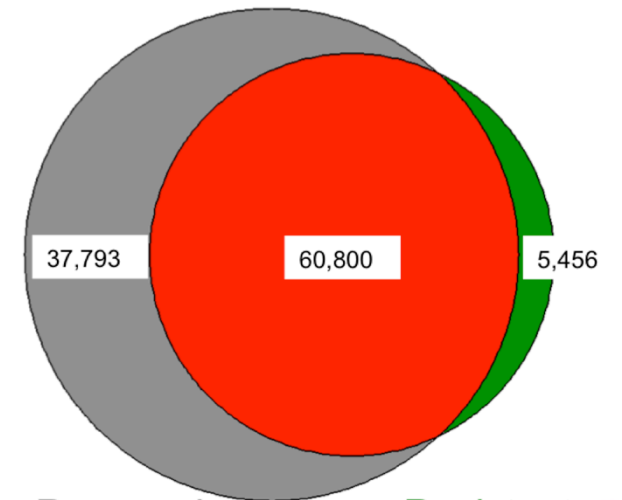
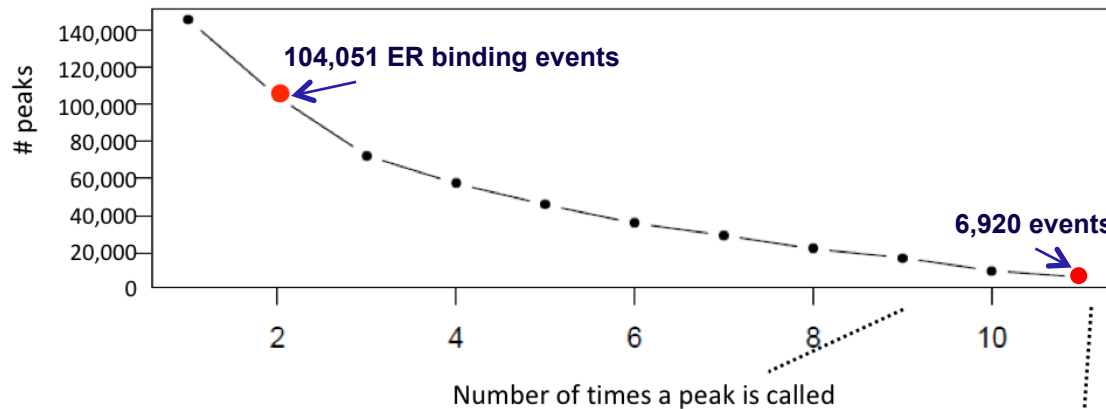
Link differential regulatory events (DB) with differential mRNA levels (DE)



Example: Occupancy (peak) analysis

11 Samples, 145586 sites in matrix:

	ID	Tissue	Factor	Condition	Replicate	Intervals
1	JC398	MCF7	ER	Responsive	1	74029
2	JC430	MCF7	ER	Responsive	2	49075
3	JC448	MCF7	ER	Responsive	3	67130
4	JC432	T47D	ER	Responsive	1	28713
5	JC439	T47D	ER	Responsive	2	23575
6	JC431	ZR75	ER	Responsive	1	74971
7	JC438	ZR75	ER	Responsive	2	70560
8	JC403	BT474	ER	Resistant	1	41924
9	JC381	BT474	ER	Resistant	2	40783
10	JC511	TAMR	ER	Resistant	1	47023
11	JC510	TAMR	ER	Resistant	2	52517



Binding affinity matrix

1. Rows: decide interval (binding site) “universe”

- Peak callers -> occupancy/overlaps
 - High-confidence sites (stringent)
 - All potential sites (lenient)
- Genomic intervals
 - Promoters
 - Windows

2. Columns: count and normalize reads for all samples in all intervals

- Duplicate reads
- Controls
- Normalization

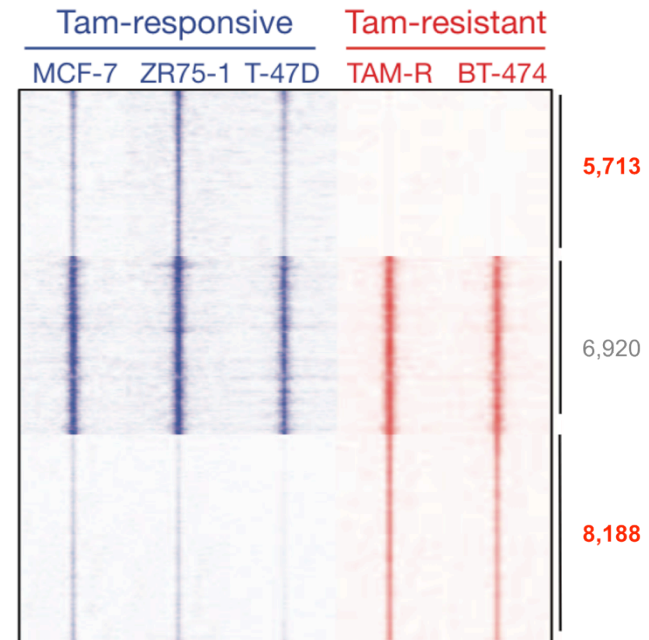
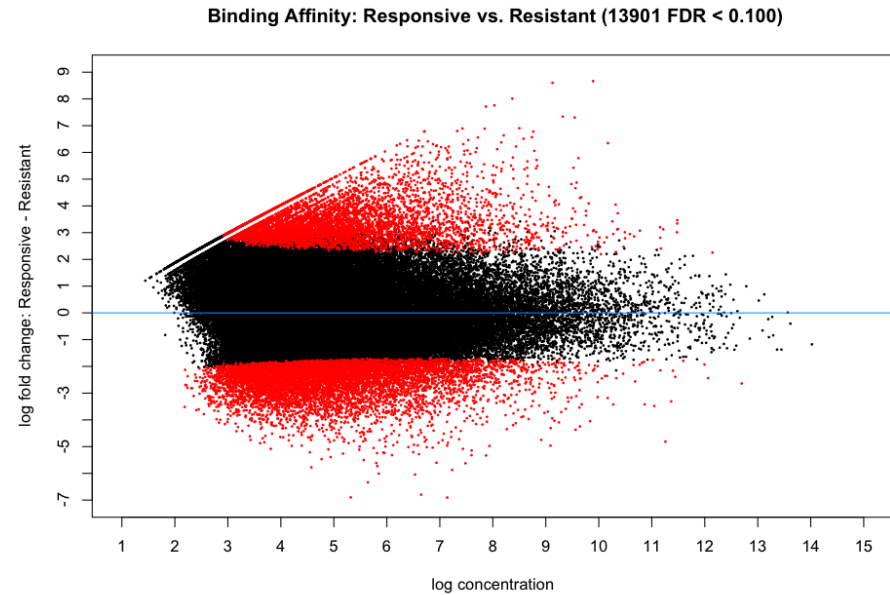
Differential binding analysis

1. Determine contrasts

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

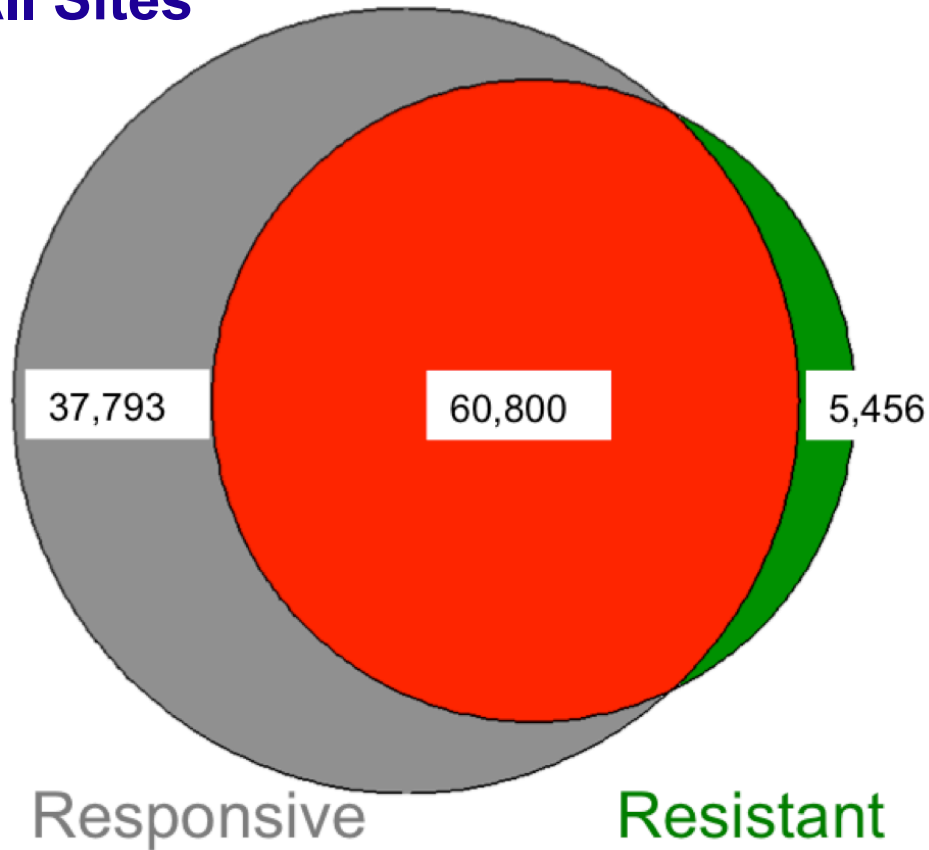
2. Run RNA-Seq DE package

- edgeR, DESeq, etc.
- Fit negative binomial distribution
- Exact test
- Multiple testing correction (B&H FDR)

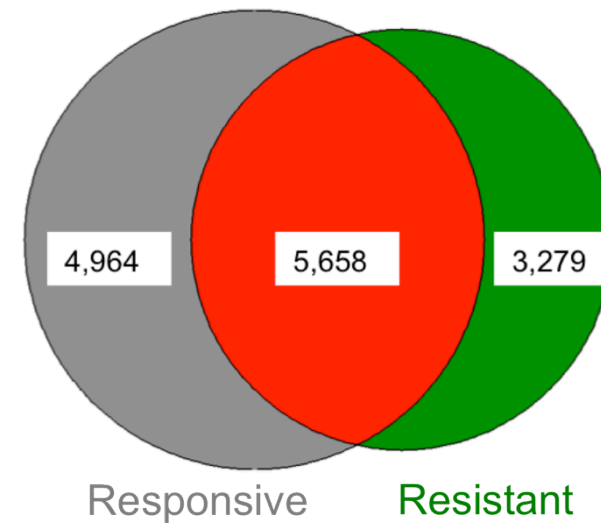


Differential binding analysis: Occupancy vs. Affinity

All Sites



Differentially Bound Sites



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)
```


Functional analysis of genome-scale regulatory data

- 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



Hands-on workshop Friday @ 1PM

Thanks to:

- Cancer Research UK (CRUK)
- **Gordon Brown**
- CRI Bioinformatics Core
 - Matthew Eldridge
 - Thomas Carroll
- **Jason Carroll** and his laboratory
 - Caryn Ross-Innes
 - Vasiliki Therodorou



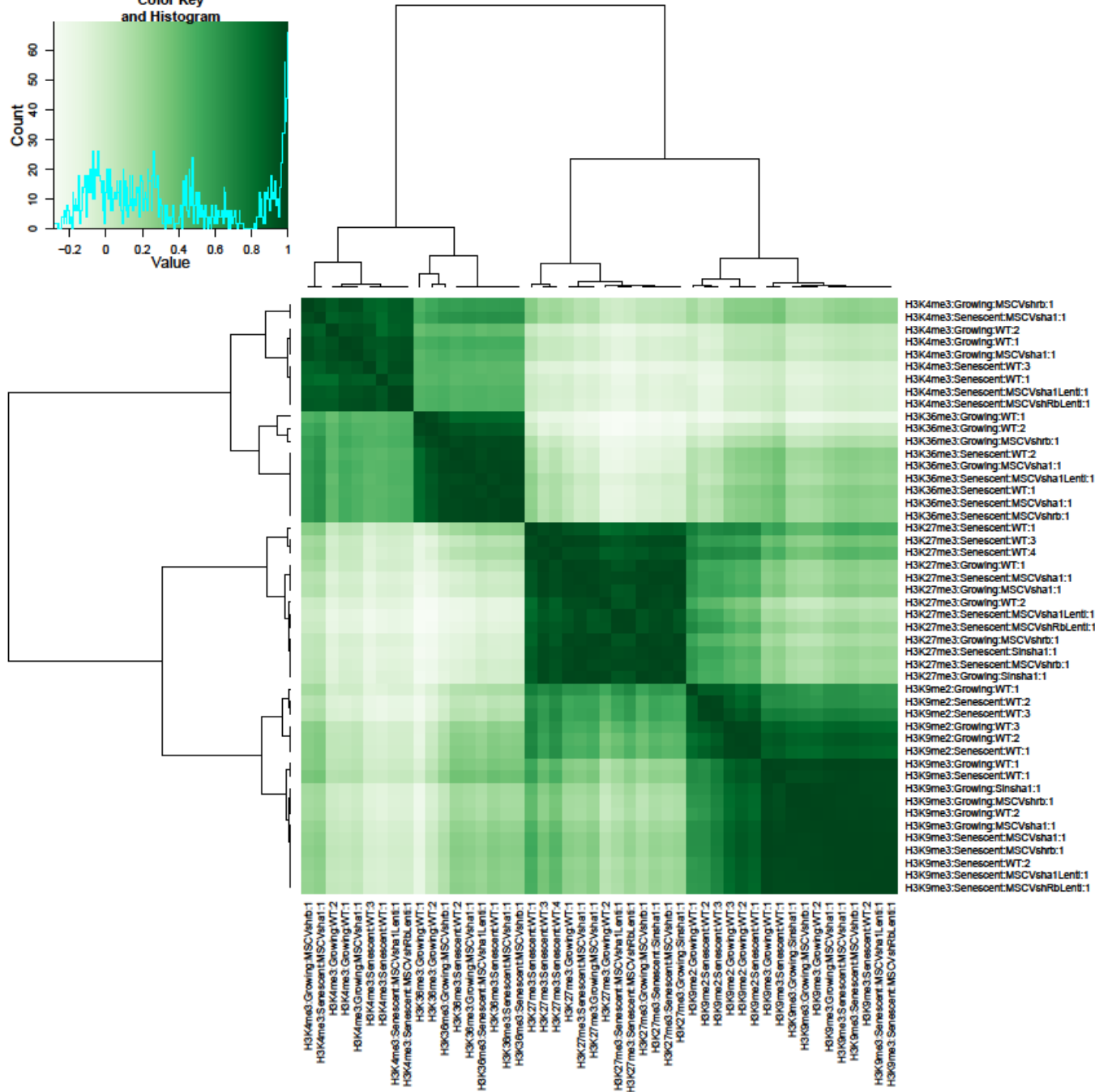
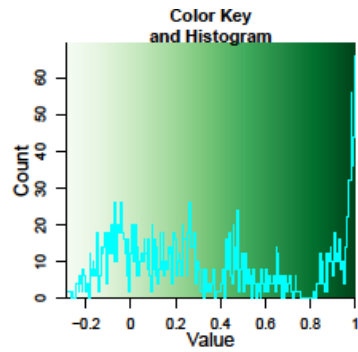
UNIVERSITY OF
CAMBRIDGE

Rory Stark
18 July 2013



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE



- **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