Observations of a year as a Bioconductor pipeline package developer

Dr. Rory Stark
Principal Bioinformatics Analyst
Cancer Research UK – Cambridge Research Institute

14 December, 2012

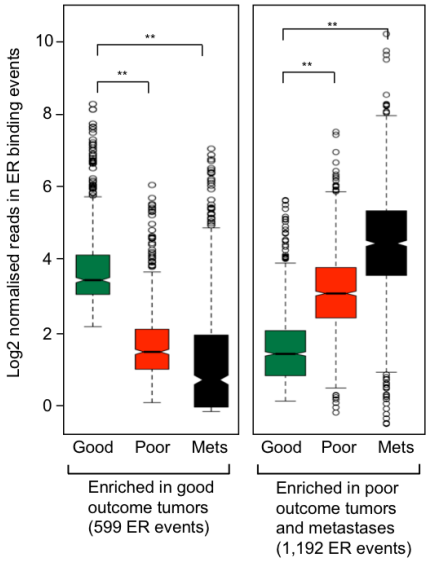
Agenda

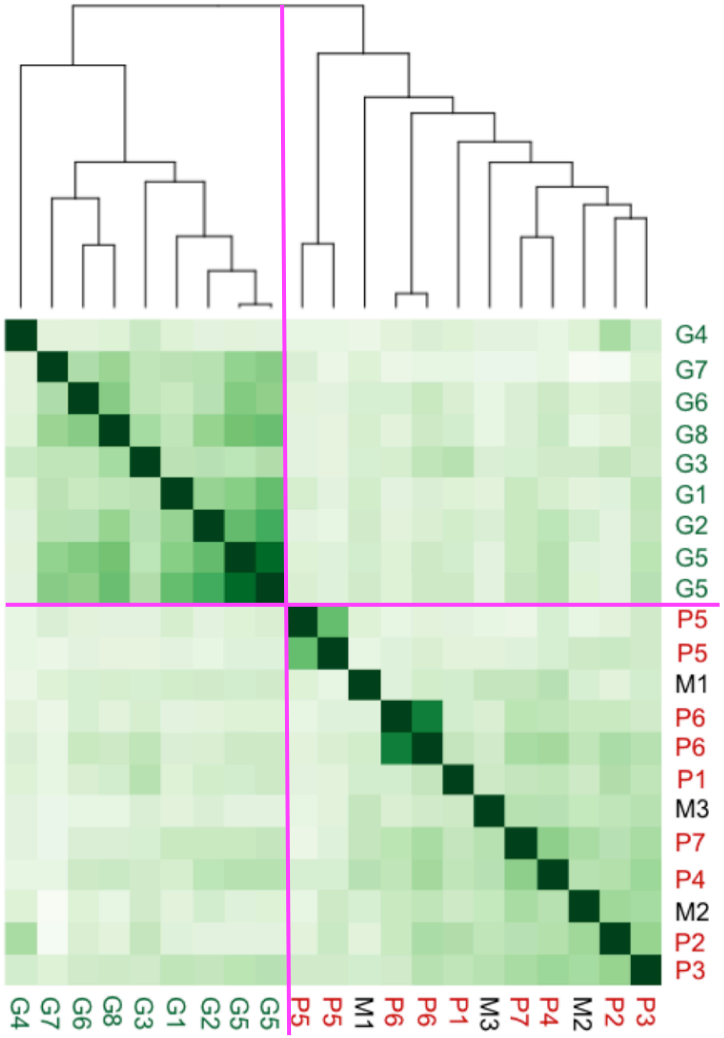
- **A year in the life of a Bioconductor developer**
  - DiffBind overview
  - Work since initial release
- **Discussion issues**
  - Are “pipeline” packages a thing?
  - Stepping on toes (references/overlaps)
  - Different developer and user audiences: software developers/biologists/bioinformatics analysts/statisticians
  - Technical issues

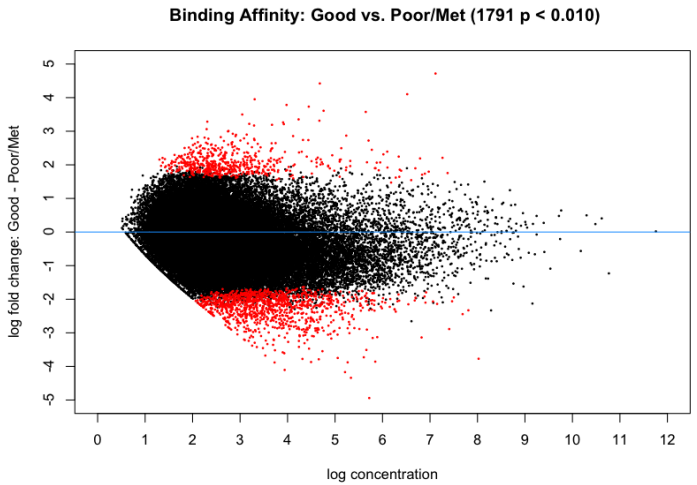
Differentially bound ChIP-seq sites separate tumours by prognosis

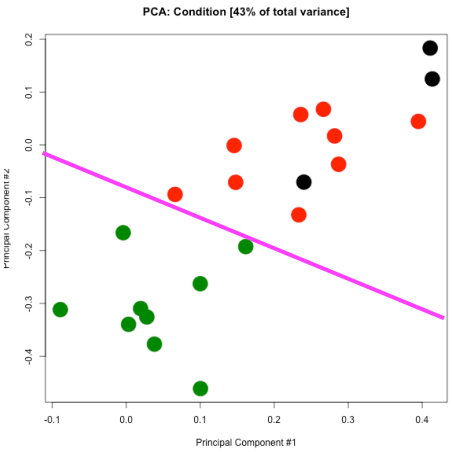
1,791 ER binding sites identified as differentially bound between good and poor prognosis

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



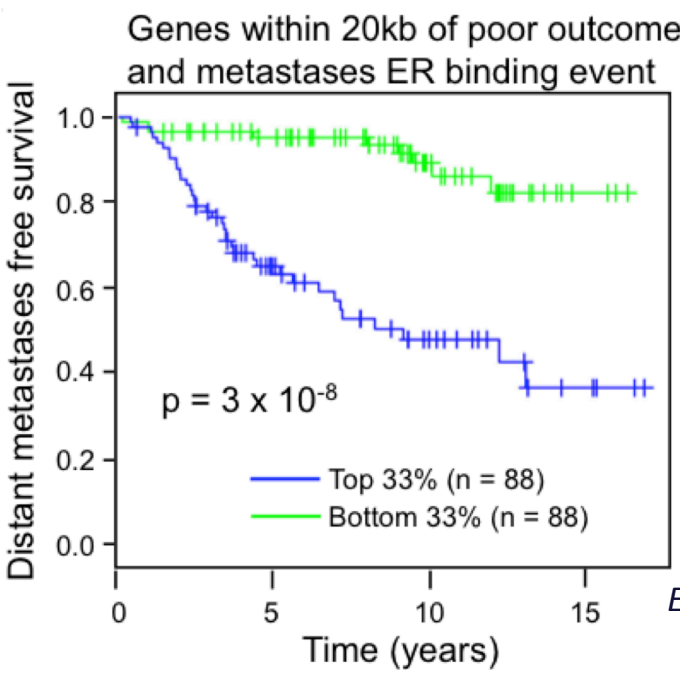




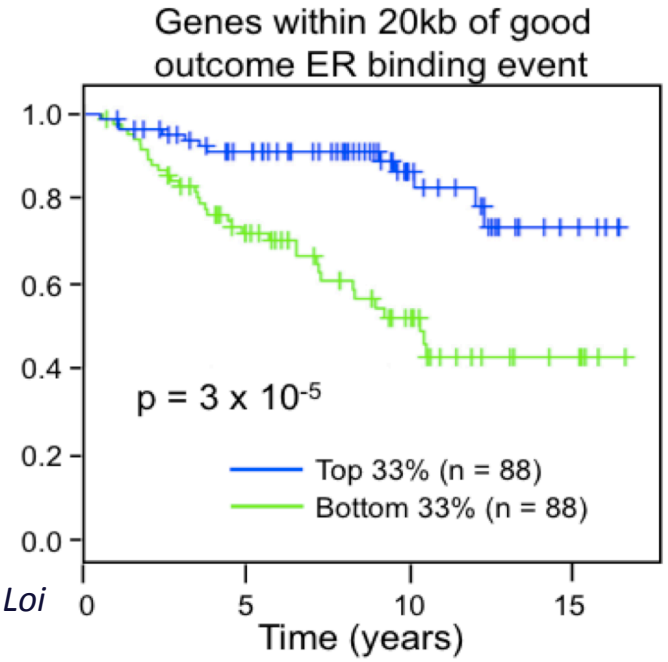


CANCER RESEARCH UK

Genes near DB sites form prognostic gene signatures



Expression data: Loi et al.

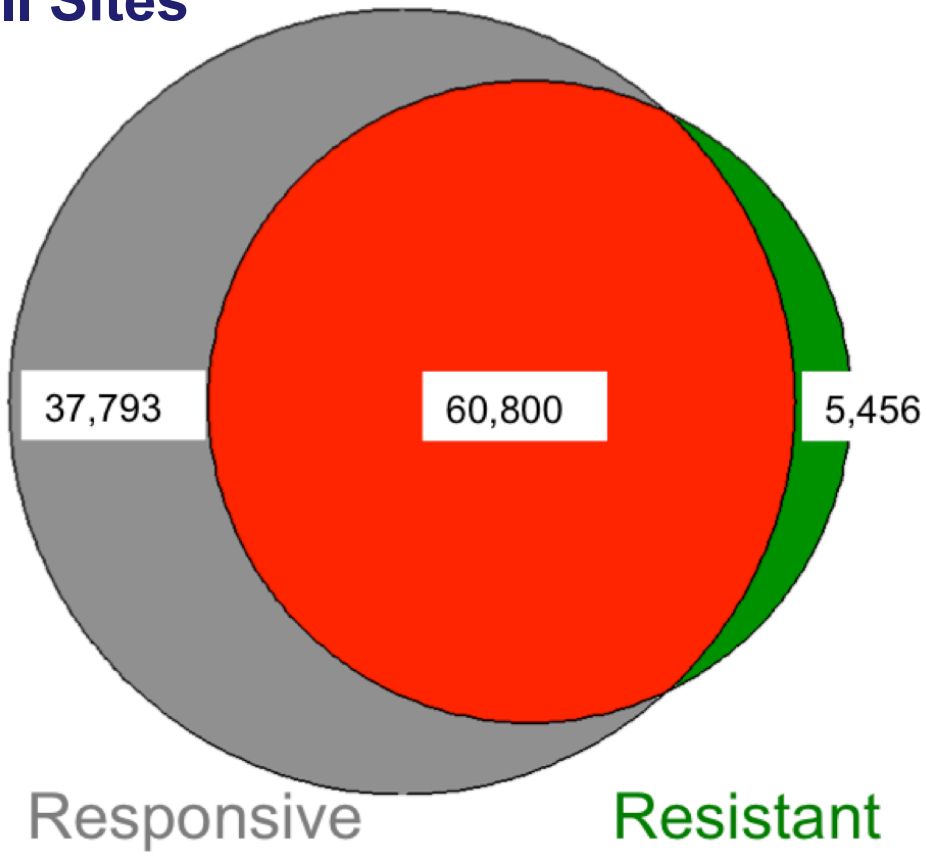


- Signature composed of genes within 20kb of DB sites
  - **265** genes in Poor outcome signature
  - **109** genes in Good outcome signature
- Classifier based on up/down regulation in mRNA expression sets
- Validated in 7 publicly available BC expression datasets

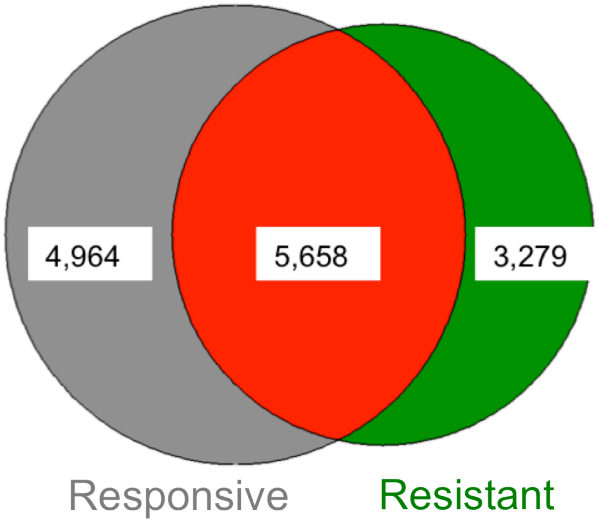
CANCER RESEARCH UK

Differential binding analysis: Occupancy vs. Affinity

All Sites



Differentially Bound Sites



CANCER RESEARCH UK

DiffBind Workflow

1. Reading in peaksets
   - Sample sheets
   - Metadata
   - Peaksets from peak callers
2. Occupancy analysis
   - Overlap venns (2-,3-,4-way)
   - Overlap rate
   - Consensus peaksets
3. Read counting
   - BAM/SAM/BED
   - Scores (RPKM)
   - Filtering
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)
5. Plotting and reporting
   - Retrieving DB sites, stats, counts
   - MA plots
   - Heatmaps (correlation, affinity)
   - PCA
   - Boxplots

DiffBind Summary

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

DiffBind efforts during 2012

- Why isn’t my NEWs page right?
- Analyses...
- Keeping up with a moving target
- Stuff I learned from users
- Fun stuff! (new features)
- The ever expanding todo list…

Keeping up with a moving target: developing in a changing environment

- **edgeR and DESeq**
  - Interface changes
  - GLMs
- **R/Bioconductor**
  - Multicore/parallel
  - GRanges (now default per MM request)

Stuff I learned from users: changes in response to support requests

- **Mailing list vs. direct mail?**
- **Bugs**
- **Efficiency (esp. memory)**
- **Warning and error messages**
- **Documentation and examples**
- **Features for common tasks**
    - Sets of consensus peaks
    - Plot using results of a different analysis

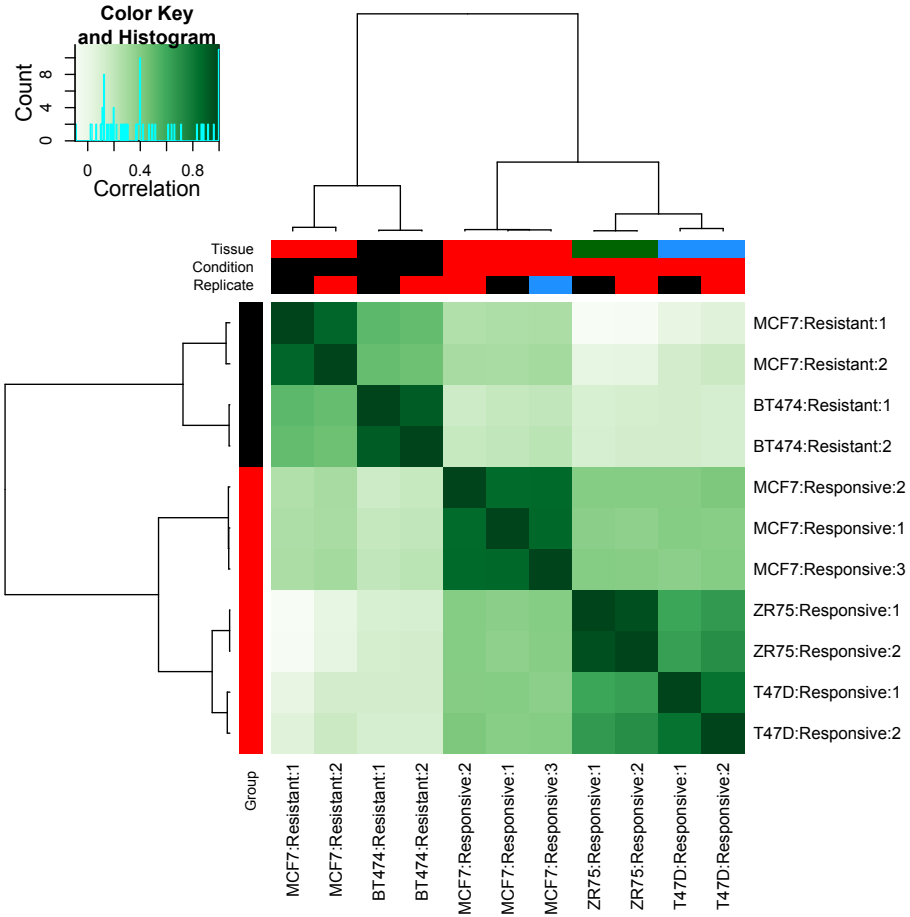
CANCER RESEARCH UK

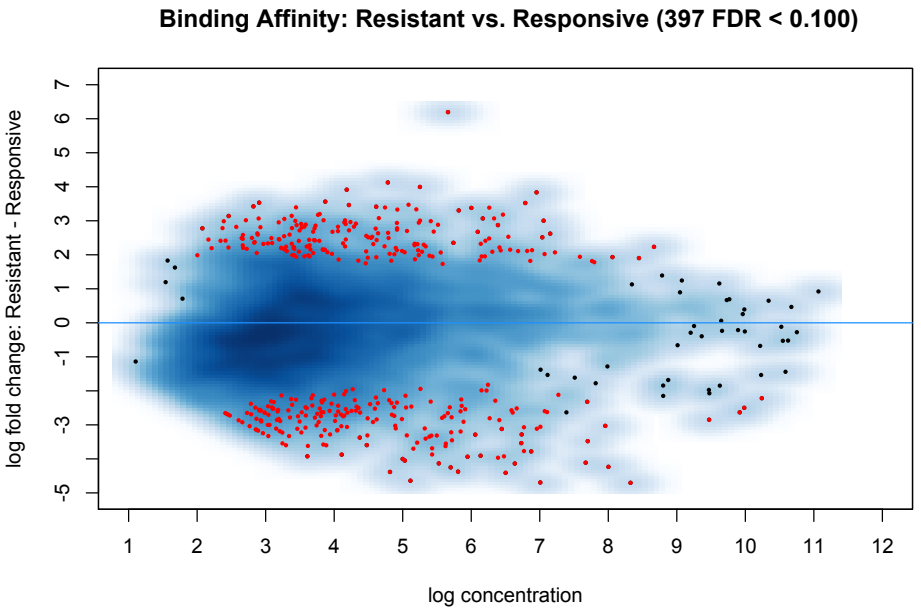
Fun stuff: new features!

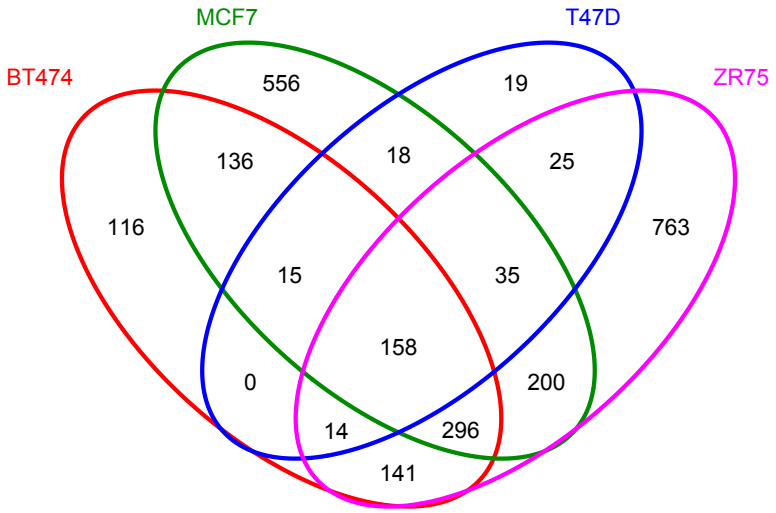
- **Data formats (esp. peaks)**
- **Metadata**
- **Counting**
  - Scaling control libraries
  - Scoring re-vamp
  - Normalization options
- **Multi-factor analysis**
  - Blocking
  - Matched
- **Reporting/plotting**
  - Thresholding (fold)

CANCER RESEARCH UK

New features: Plotting







Thanks to Thomas Girke

CANCER RESEARCH UK

The never-ending todo list…

- ABCD-DNA
- Summits
- Profiles
- Peak centering
- Cluster parallelization
- RNA-seq
- ChIPseqQC
- …

CANCER RESEARCH UK

Discussion: Are workflow/pipeline packages a thing?

- **DiffBind, easyRNA, QuasR, ArrayExpressHTS, HTSeqGenie etc.**
  - We're working on ChIPseqQC and beadarrayPipeline etc.
- **Not focused on implementing statistical methods, but on organizing project data and moving it through processing workflow using existing statistical packages**
  - Reference/citation/overlap issues
- **Generally as "automatic" as possible, using defaults in multiple places**
  - Reproducibility +/-
- **Users: biologists and support bioinformaticians**
- **Authors: bioinformaticians and software developers**

CANCER RESEARCH UK

Stepping on toes: references, citations, functional overlap

- **e.g. DiffBind publications**
- **Reference/citation issues**
  - edgeR (Smyth)
    - Package attribution
    - Specific statistical method attribution (normalization, exact test)
  - DESeq?
- **Functional overlap**
  - ABCD-DNA (Robinson/Repitools)
  - MMSeq (Schweikert)

CANCER RESEARCH UK

Discussion: Bioconductor package authors and package users

- **Statisticians**
  - "DiffBind doesn't *do* anything!" – Yes, exactly!
    - 7500 lines of R (150 functions), C++
  - MM: Well, *I* wouldn't use DiffBind…
- **Bioinformatics analysts**
- **Biologists**
- **Software developers**

CANCER RESEARCH UK

Discussion: Technical issues

- **New standard types for pipelines?**
  - Example: sample sheets with metadata and filenames
- **Function conventions**
  - Namespaces (dba.)
  - General vs. specific
  - Overloading parameters
- **External (16) vs Internal (133) functions**
  - Do I have an obligation to document 100+ internal functions?

CANCER RESEARCH UK

Acknowledgements

- **Gordon Brown**
- CRI Bioinformatics Core
  - **Matthew Eldridge**
  - Suraj Menon
  - Tom Carroll
- **Jason Carroll** and his laboratory
  - **Caryn Ross-Innes**
  - Vasiliki Therodorou
- Co-Authors
  - Carlos Caldas
  - Andrew Teschendorff