ROCK: a breast cancer functional genomics resource

David Sims, Borisas Bursteinas, Qiong Gao, Ekta Jain, Alan Mackay, Costas Mitsopoulos, Marketa Zvelebil

ROCK: a Breast Cancer Functional Genomics Resource

David Sims, Borisas Bursteinas, Qiong Gao, Ekta Jain, Alan MacKay, Costas Mitsopoulos and Marketa Zvelebil*

* To whom correspondence should be addressed. Tel: +44 207 153 5350; Fax: +44 207 153 5016; Email: marketa@icr.ac.uk

Breakthrough Breast Cancer Research Centre, Institute of Cancer Research, Chester Beatty Laboratories, 237 Fulham Road, London, SW3 6JB.

Abstract

The clinical and pathological heterogeneity of breast cancer has instigated efforts to stratify breast cancer sub-types according to molecular profiles. These profiling efforts are now being augmented by large-scale functional screening of breast tumour cell lines, using approaches such as RNA interference. We have developed ROCK (rock.icr.ac.uk) to provide a unique, publicly accessible resource for the integration of breast cancer functional and molecular profiling datasets. ROCK provides a simple online interface for the navigation and cross-correlation of gene expression, aCGH and RNAi screen data. It enables the interrogation of gene lists in the context of statistically analysed functional genomic datasets, interaction networks, pathways, GO terms, mutations and drug targets. The interface also provides interactive visualisations of datasets and interaction networks. ROCK collates data from a wealth of breast cancer molecular profiling and functional screening studies into a single portal, where analysed and annotated results can be accessed at the level of a gene, sample or study. We believe that portals such as ROCK will not only afford researchers rapid access to profiling data, but also aid the integration of different data types, thus enhancing the discovery of novel targets and biomarkers for breast cancer.

Keywords:

Breast cancer, functional genomics, integration, public resource, networks, gene-specific data.

Introduction

Breast cancer rates have increased by more than 50% over the last twenty years and breast cancer is now the most common cancer in the UK, with more than 46,000 women diagnosed each year, and more than a million cases worldwide [1]. Breast cancer is a heterogeneous disease, comprising multiple entities with distinct risk factors, biological features, histopathological characteristics and response to therapies. In the last decade, a wide range of molecular profiling approaches have been employed in an attempt to identify sub-types of breast cancer [2-5] with the aim of better understanding the disease process and identifying novel therapeutic approaches. More recently, functional genomic methods such as RNA interference (RNAi) screens have aimed to identify new molecular targets, such as genes that are recurrently amplified, over-expressed and functionally required for viability in a range of cancer cell lines [6]. Indeed, use of these and other functional genomic approaches is set to grow as they offer the potential to identify novel therapeutic targets in breast cancer sub-types that are unresponsive to existing treatment regimes.

Gene-specific data from large-scale profiling studies are often difficult to access from supplementary material in manuscripts or microarray data repositories [7, 8]. This is largely due to different data formats, gene identifiers and sample annotations being used in different studies. Furthermore, many functional genomic studies generate large-scale datasets that can be mined in a number of different ways from those originally published. To address these issues, we have developed ROCK (rock.icr.ac.uk) to provide a unique, publicly accessible resource for the analysis and integration of breast cancer functional genomic datasets.

ROCK provides an intuitive online interface for the navigation, cross-linking and cross-correlation of microarray gene expression, DNA copy number (array competitive genomic hybridisation) and RNAi screen data from breast cancer cell lines and tumour samples. The interface was designed to provide researchers with access to the results of statistical analysis of key molecular and functional profiling datasets at the level of a single gene, gene list, sample or study. Thus, ROCK enables breast cancer researchers to quickly assess the potential significance of genes of interest in the context of prior knowledge in the field.

Furthermore, we believe that integration of molecular and functional profiling datasets in resources such as ROCK can aid classification of breast cancer sub-types leading to improved prediction of prognosis and response to treatment.

The Database

The ROCK resource is based on a number of key breast cancer functional genomics and somatic mutation datasets collated from literature and online resources [7-10]. Public gene expression, aCGH and RNAi datasets have been systematically analysed using well-established methodologies to enable users to mine the most useful data quickly and easily.

In order to facilitate the analysis and integration of breast tumour datasets, annotations in various formats were translated into a consistent ontology detailing tumour pathology [11], tumour macroscopy [12], histological grading [13], tumour staging [14] and molecular markers (including oestrogen receptor (ER), progesterone receptor (PgR), androgen receptor (AR), HER2 and Ki67 staining, as well as *p53*, *BRCA1* and *BRCA2* mutation status).

In the case of microarray gene expression studies, four key analyses were performed on each dataset. Firstly, the breast cancer subtype for each sample in each study was predicted using centroid correlation to the PAM50 classifier [5, 15]. Secondly, significance analysis of microarrays (SAM) [16] was performed on each dataset using predicted breast cancer subtype as well as common classes of functional annotations such as oestrogen receptor (ER) status and tumour grade to delineate sample groups. Thirdly, groups of co-expressed genes were identified using the Pearson correlation ($r>0.5$, adjusted p-value<0.01). Finally, survival analysis (survival package in R) was performed for each gene in each study (where survival and censoring data were available) to identity genes where expression had potential clinical predictive or prognostic value.

For aCGH datasets, normalised data from individual studies were smoothed using circular binary segmentation [17] to translate noisy intensity measurements into

regions of equal copy number, then scaled to the genome median absolute deviation. Recurrent gains, losses, amplifications and deletions (>10% samples) were called from the smoothed and scaled data using consistent thresholds. For several studies where data were available the correlation between DNA copy number and gene expression level was assessed for each gene that could be matched across platforms. Firstly, the Pearson correlation between normalised gene expression level and smoothed CGH value was calculated for each gene in each study. Secondly, raw and adjusted [18] Wilcoxon p-values for gain, loss, amplification and deletion were computed wherever relevant.

RNAi screen data were curated from literature. Where raw screen data were available analysis was performed using the cellHTS2 package in R [19], and hits were called using screen-specific thresholds. Where screen hit or phenotypes annotations were provided by authors, these were also added to the database.

Interaction networks were constructed by amalgamating data from a range of different online resources [20-27], and supplemented with data curated in-house. Similarly, pathway datasets were parsed from public resources [23, 28-30] and assembled into a single dataset.

User interface

Within ROCK users can browse through lists of breast cancer specific gene expression and DNA copy number (aCGH) studies, RNAi screens, genetic mutations [9] and existing therapeutics [31]. For each curated gene expression study users can view graphical summaries of the available sample annotation and the various analyses already performed on the dataset. For each aCGH study users can view copy number aberration frequency plots (Figure 1a), and focus in on individual samples, gaining access to sample plots (Figure 1b), which can be zoomed in to individual chromosome level. Users can also access details of specific recurrent amplicons and view lists of genes within then. Similarly, For RNAi screens users can focus upon lists of individual hits and their associated phenotypes. Furthermore, complete normalised datasets and associated standardised functional annotations are available for download, to enable bioinformaticians to pursue more detailed or specific analyses.

Alternatively, users can employ text-based searches to navigate within the database. For example, users can begin their search with a single gene of interest to identify the corresponding gene expression pattern, survival data, copy number aberrations or RNAi phenotype for that gene in any study. ROCK contains gene identifiers from Ensembl [32], Entrez gene [33], UniGene [34], HUGO [35], RefSeq [36] and CCDS [37], along with protein data from UniProt [38] and microarray probe identifiers from Affymetrix, Illumina, Agilent and several other cDNA microarray platforms, to maximise the likelihood of identifying genes of interest. Users can search for genes from mouse [39], yeast [40], fly [27] or worm [41] as well as man. If a search is initiated using a non-human gene, lists of putative human orthologues will be returned from the three different homology datasets contained in the database: HomoloGene [36], InParanoid [42] and Ensembl Compara [32]. Once a particular human gene has been selected, the user is directed to the ROCK entry for that gene. This page is divided into a set of tabs relating to different types of gene-specific information.

By navigating down from the gene expression, aCGH, methylation or RNAi summary tab for each gene, users can access detailed information about the significance of that gene in each study (Figure 2). For example, in the case of RNAi data, details of the cell line and assay are supplied, along with phenotypic data. For expression data, users can navigate down from the initial summary to view evidence of differential regulation (Figure 3a), or lists of genes co-expressed with the gene of interest in particular studies (Figure 3b). Where data are available, users can view Kaplan Meier survival curves for individuals expressing different levels of a given gene (Figure 3c), or assess the correlation between gene expression level and DNA copy number (Figure 3d).

From the primary gene page, users can also quickly access Gene Ontology [43], InterPro[44], homology [32, 36, 42], somatic mutation[9], small molecule [31], protein interaction [20-27], and pathway [23, 28, 29] information for that gene.

Batch searches

ROCK has been designed to allow users to analyse the results of high-throughput experiments in the context of other large-scale datasets. Thus, searches can be carried out using lists of genes, microarray probes, GO terms, protein motifs and small molecule inhibitors. Gene lists generated within the database can be saved and combined in a number of ways by registered users in the 'Account' section. All gene lists, whether from batch searches or ROCK analysis results, can be used in a variety of batch searches using the 'Link' function. Linking enables users to cross-reference datasets. For example, lists of genes showing significantly different expression between ER+ and ER- breast cancer samples (as established by SAM analysis) can be saved. Intersection of such lists from several studies can then be used to derive a single list of consistently differentially expressed genes. This gene list can be linked to pathway analysis, to identify statistically over-represented pathways, to network analysis, to visualise gene connectivity (Figure 4), or interrogated for gene annotation such as cytoband, or known breast cancer somatic and germ line mutations.

Discussion and Future Prospects

ROCK was created to meet a clear need in the breast cancer research community for a public, integrated functional genomics resource. High-throughput experiments have formed a core part of cancer research for over a decade now, and of the solid malignancies, breast cancer has been the most extensively analysed with high throughput methods. However, until now there has been no means for the bench scientist to easily and quickly access the vast amounts of data contained in these studies. Microarray data repositories such as GEO [7] and ArrayExpress [8] enable sharing of gene expression datasets in a standard format, but not all important cancer studies are included in these resources and they do not offer statistical analysis of the data. Oncomine[45] was developed to meet this shortfall, but it covers a much wider range of cancer types and as a result lacks a breast cancer relevant interpretation of the gene expression datasets ROCK contains. Furthermore, Oncomine[45] does not include aCGH or RNAi screen data. While expression array studies have helped to unravel molecular subgroups of breast cancer and to develop prognostic and predictive signatures, their role in the identification of novel therapeutic targets has so

far been limited. The promise of high throughput methods is likely to be realised only by the integration of multiple sources of data, including those from functional RNA interference studies.

ROCK was developed to complement existing breast cancer databases such as the Breast Cancer Information Core (http://research.nhgri.nih.gov/bic/) and the Breast Cancer Database (http://www.breastcancerdatabase.org/), which are focussed on curating literature on somatic and germline mutations in breast cancer, rather than high-throughput functional genomic studies.

The disease-specific focus of ROCK will continue to assure that it differs in its function, scope and content from other online resources. In the future we plan to add not only new datasets for our existing range of supported experimental platforms, but also to extend the database infrastructure to support SNP genotyping data, DNA methylation arrays, shRNA barcode screens and next generation sequencing (NGS) data. Furthermore, we aim to expand our in-house curated pathways to give complete and up-to-date coverage of key breast cancer relevant pathways, particularly those involved in DNA repair and oestrogen signalling.

Acknowledgements

The authors would like to thank Dr. Amar Sabri Ahmad, Dr. Jorge Reis-Filho, Dr. Chris Lord and Prof. Alan Ashworth of the Institute for comments on the manuscript. We acknowledge funding from Breakthrough Breast Cancer and NHS funding to the NIHR Biomedical Research Centre. Funding to pay the Open Access publication charges for this article was provided by Breakthrough Breast Cancer. There are no conflicts of interest.

References

1. Harris, J.R., Morrow,M., Lippman, M.E., Osborne,C.K., *Diseases of the Breast*. Fourth ed. 2009: Lippincott Williams & Wilkins. 1408.
2. Hicks, J., et al., *Novel patterns of genome rearrangement and their association with survival in breast cancer.* Genome Res, 2006. **16**(12): p. 1465-79.
3. Hu, Z., et al., *The molecular portraits of breast tumors are conserved across microarray platforms.* BMC Genomics, 2006. **7**: p. 96.
4. Natrajan, R., et al., *Tiling path genomic profiling of grade 3 invasive ductal breast cancers.* Clin Cancer Res, 2009. **15**(8): p. 2711-22.
5. Parker, J.S., et al., *Supervised risk predictor of breast cancer based on intrinsic subtypes.* J Clin Oncol, 2009. **27**(8): p. 1160-7.
6. Iorns, E., et al., *Integrated functional, gene expression and genomic analysis for the identification of cancer targets.* PLoS One, 2009. **4**(4): p. e5120.
7. Barrett, T., et al., *NCBI GEO: mining tens of millions of expression profiles--database and tools update.* Nucleic Acids Res, 2007. **35**(Database issue): p. D760-5.
8. Parkinson, H., et al., *ArrayExpress--a public database of microarray experiments and gene expression profiles.* Nucleic Acids Res, 2007. **35**(Database issue): p. D747-50.
9. Bamford, S., et al., *The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website.* Br J Cancer, 2004. **91**(2): p. 355-8.
10. Gollub, J., et al., *The Stanford Microarray Database: data access and quality assessment tools.* Nucleic Acids Res, 2003. **31**(1): p. 94-6.
11. Tavassoli, A.F., Devilee, P., *Tumours of the breast and female genital organs, WHO classification of tumours*. 2003.
12. NHS Breast Screening Programme, *Pathology Reporting of Breast Disease*. Vol. 58. 2005: NHS.
13. Elston, C.W. and I.O. Ellis, *Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up.* Histopathology, 1991. **19**(5): p. 403-10.
14. Singletary, S.E., et al., *Revision of the American Joint Committee on Cancer staging system for breast cancer.* J Clin Oncol, 2002. **20**(17): p. 3628-36.
15. Tibshirani, R., et al., *Diagnosis of multiple cancer types by shrunken centroids of gene expression.* Proc Natl Acad Sci U S A, 2002. **99**(10): p. 6567-72.
16. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.
17. Venkatraman, E.S. and A.B. Olshen, *A faster circular binary segmentation algorithm for the analysis of array CGH data.* Bioinformatics, 2007. **23**(6): p. 657-63.
18. Benjamini, Y., Hochberg, Y., *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of Royal Statistical Society Series, 1995. **B 57**(1): p. 289-300.
19. Boutros, M., L.P. Bras, and W. Huber, *Analysis of cell-based RNAi screens.* Genome Biol, 2006. **7**(7): p. R66.
20. Alfarano, C., et al., *The Biomolecular Interaction Network Database and related tools 2005 update.* Nucleic Acids Res, 2005. **33**(Database issue): p. D418-24.

21. Chatr-aryamontri, A., et al., *MINT: the Molecular INTeraction database.* Nucleic Acids Res, 2007. **35**(Database issue): p. D572-4.
22. Kerrien, S., et al., *IntAct--open source resource for molecular interaction data.* Nucleic Acids Res, 2007. **35**(Database issue): p. D561-5.
23. Matthews, L., et al., *Reactome knowledgebase of human biological pathways and processes.* Nucleic Acids Res, 2009. **37**(Database issue): p. D619-22.
24. Pagel, P., et al., *The MIPS mammalian protein-protein interaction database.* Bioinformatics, 2005. **21**(6): p. 832-4.
25. Peri, S., et al., *Human protein reference database as a discovery resource for proteomics.* Nucleic Acids Res, 2004. **32**(Database issue): p. D497-501.
26. Stark, C., et al., *BioGRID: a general repository for interaction datasets.* Nucleic Acids Res, 2006. **34**(Database issue): p. D535-9.
27. Tweedie, S., et al., *FlyBase: enhancing Drosophila Gene Ontology annotations.* Nucleic Acids Res, 2009. **37**(Database issue): p. D555-9.
28. Okuda, S., et al., *KEGG Atlas mapping for global analysis of metabolic pathways.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W423-6.
29. Schaefer, C.F., et al., *PID: the Pathway Interaction Database.* Nucleic Acids Res, 2009. **37**(Database issue): p. D674-9.
30. Zhao, F., et al., *TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies.* Nucleic Acids Res, 2005. **33**(Database issue): p. D103-7.
31. Overington, J.P., B. Al-Lazikani, and A.L. Hopkins, *How many drug targets are there?* Nat Rev Drug Discov, 2006. **5**(12): p. 993-6.
32. Hubbard, T.J., et al., *Ensembl 2009.* Nucleic Acids Res, 2009. **37**(Database issue): p. D690-7.
33. Maglott, D., et al., *Entrez Gene: gene-centered information at NCBI.* Nucleic Acids Res, 2005. **33**(Database issue): p. D54-8.
34. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.* Nucleic Acids Res, 2005. **33**(Database issue): p. D501-4.
35. Bruford, E.A., et al., *The HGNC Database in 2008: a resource for the human genome.* Nucleic Acids Res, 2008. **36**(Database issue): p. D445-8.
36. Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology.* Nucleic Acids Res, 2003. **31**(1): p. 28-33.
37. Pruitt, K.D., et al., *The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes.* Genome Res, 2009. **19**(7): p. 1316-23.
38. Consortium, T.U., *The Universal Protein Resource (UniProt) 2009.* Nucleic Acids Res, 2009. **37**(Database issue): p. D169-74.
39. Bult, C.J., et al., *The Mouse Genome Database (MGD): mouse biology and model systems.* Nucleic Acids Res, 2008. **36**(Database issue): p. D724-8.
40. project., S., *"Saccharomyces Genome Database"*.
41. Chen, N., et al., *WormBase: a comprehensive data resource for Caenorhabditis biology and genomics.* Nucleic Acids Res, 2005. **33**(Database issue): p. D383-9.
42. O'Brien, K.P., M. Remm, and E.L. Sonnhammer, *Inparanoid: a comprehensive database of eukaryotic orthologs.* Nucleic Acids Res, 2005. **33**(Database issue): p. D476-80.

43. Barrell, D., et al., *The GOA database in 2009--an integrated Gene Ontology Annotation resource.* Nucleic Acids Res, 2009. **37**(Database issue): p. D396-403.
44. Hunter, S., et al., *InterPro: the integrative protein signature database.* Nucleic Acids Res, 2009. **37**(Database issue): p. D211-5.
45. Rhodes, D.R., et al., *Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles.* Neoplasia, 2007. **9**(2): p. 166-80.
46. Chin, K., et al., *Genomic and transcriptional aberrations linked to breast cancer pathophysiologies.* Cancer Cell, 2006. **10**(6): p. 529-41.
47. Miller, L.D., et al., *An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival.* Proc Natl Acad Sci U S A, 2005. **102**(38): p. 13550-5.
48. Mackay, A., et al., *A high-resolution integrated analysis of genetic and expression profiles of breast cancer cell lines.* Breast Cancer Res Treat, 2009.
49. Sotiriou, C., et al., *Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis.* J Natl Cancer Inst, 2006. **98**(4): p. 262-72.
50. Neve, R.M., et al., *A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes.* Cancer Cell, 2006. **10**(6): p. 515-27.

Figure Legends

Figure 1. Visualisation of DNA copy number data in ROCK. a) Plot summarising the frequency of amplification and deletions across genome in 118 breast cancer sample from Chin et al [46]. Amplifications are depicted in green and deletions in red. b) Plot of the log ratio for sample B0165 for all BACs across the genome [46]. Red indicates los or deletion, green indicated gain or amplification, and the purple line indicates the CBS smooth.

Figure 2. Screenshot of the expression tab gene page for ESR1 showing the studies where analysed data is available. Users can access details by clicking on the green arrows or SAM analysis parameters. Batch searches are available from the menu on the left hand side and further gene-specific information can be accessed from the tabs above the gene header.

Figure 3. Visualisation of gene expression data analysis in ROCK. a) Boxplot showing differential expression of BRCA2 between Oestrogen receptor positive and

oestrogen receptor negative breast cancer sample in the Miller *et al* study [47]. b) Heatmap of genes co-expressed with PPM1D in MacKay *et al* [48]. c) Kaplan Meier survival curves for high and low expression of CDC2 in Sotiriou *et al* [49]. d) Correlation of gene expression with DNA copy number for ERBB2 in Neve *et al* [50].

Figure 4. Network analysis of genes consistently differentially expressed between oestrogen receptor positive and negative tumours. Nodes coloured red represent ER regulated genes, whereas nodes in green are connecting nodes. Black edges indicate physical interactions and yellow edges indicate that nodes form part of the same protein complex. The ER+ genes (left hand side) are linked to the ER- gene-network (right hand side) by GRB2. Many of the ER regulated genes were shown to interact with genes involved in DNA repair and cell cycle (pathway and GO term analysis).

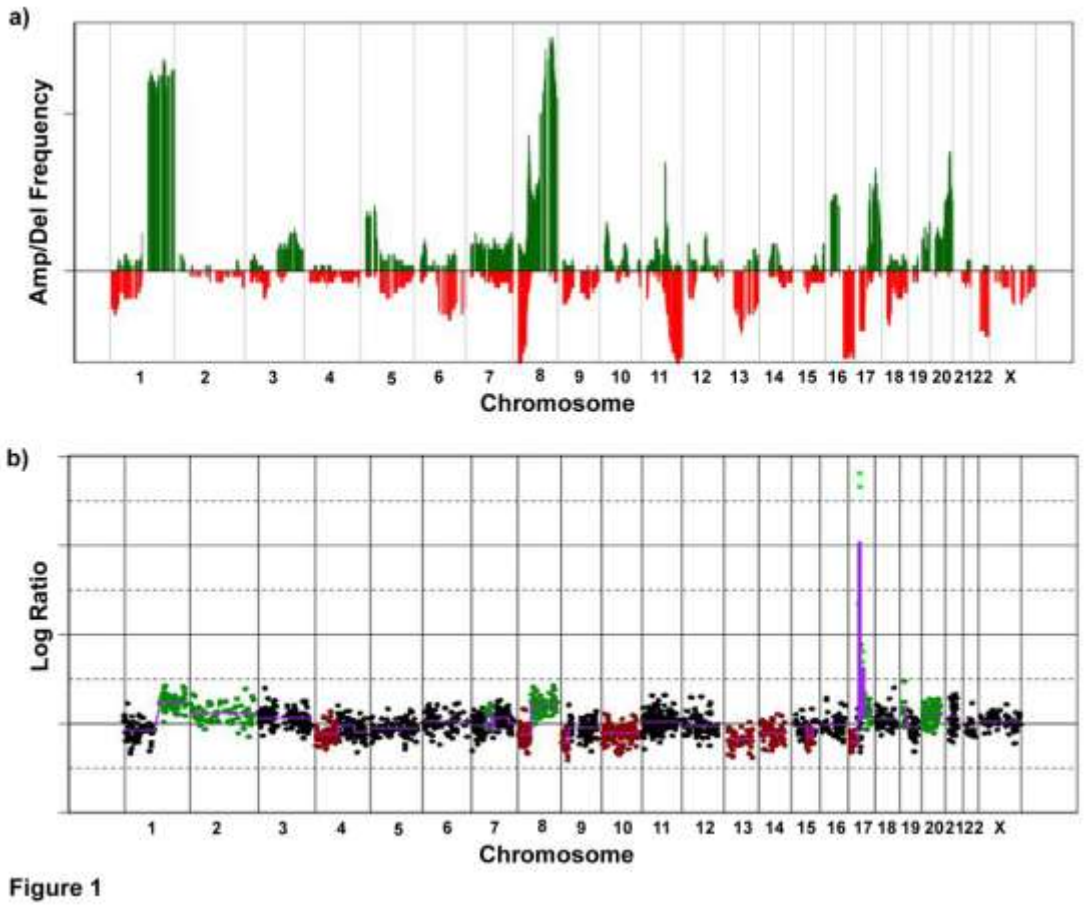


Figure 1

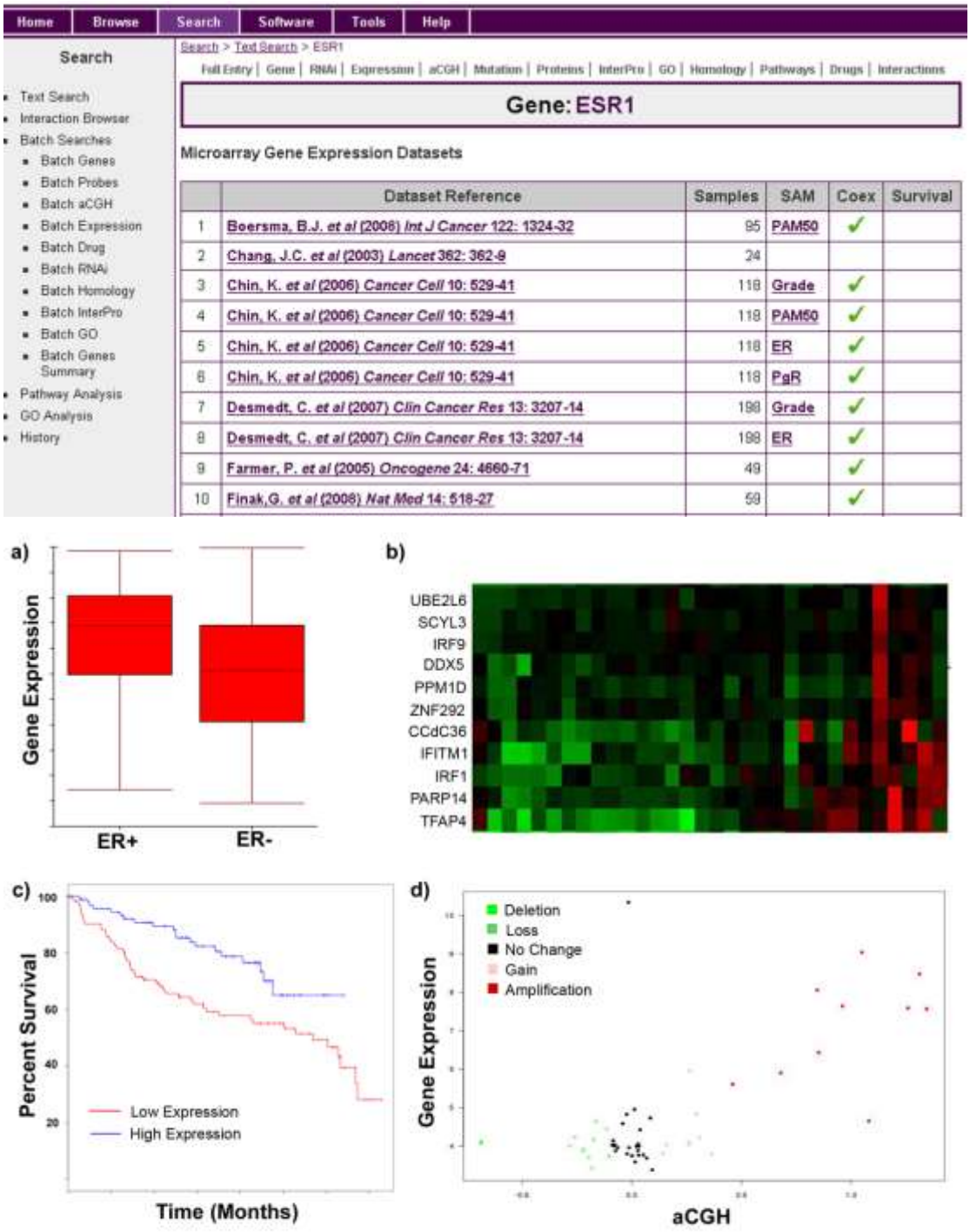
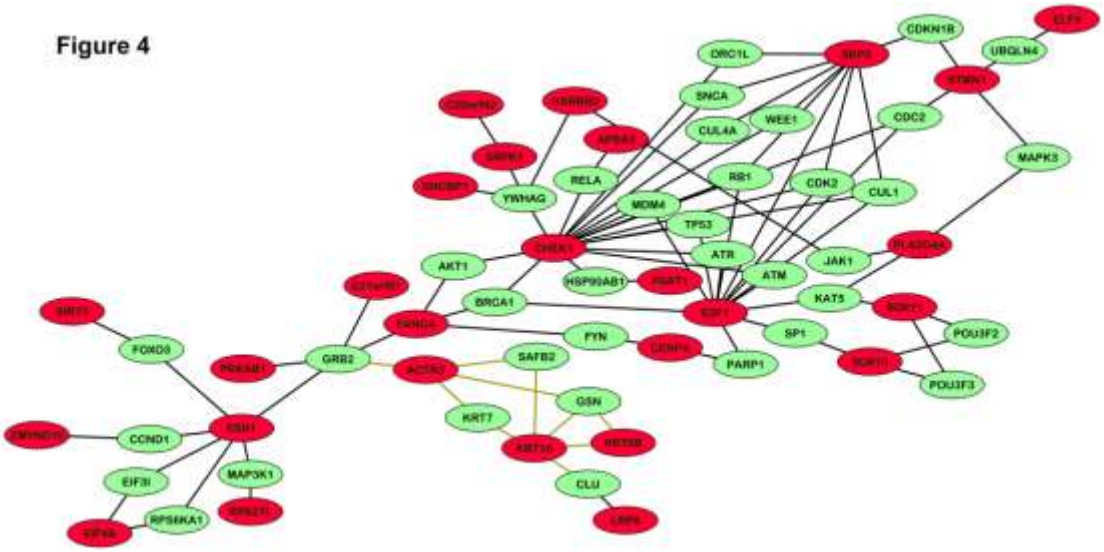

Home
Browse
Search
Software
Tools
Help
Search
Text Search
Interaction Browser
Batch Searches
Batch Genes
Batch Probes
Batch aCGH
Batch Expression
Batch Drug
Batch RNAi
Batch Homology
Batch InterPro
Batch GO
Batch Genes Summary
Pathway Analysis
GO Analysis
History
Search > Text Search > ESR1
Full Entry | Gene | RNAi | Expression | aCGH | Mutation | Proteins | InterPro | GO | Homology | Pathways | Drugs | Interactions
Gene: ESR1
Microarray Gene Expression Datasets
Dataset Reference
Samples
SAM
Coex
Survival
1 Boersma, B.J. et al (2008) Int J Cancer 122: 1324-32 95 PAM50
2 Chang, J.C. et al (2003) Lancet 362: 362-9 24
3 Chin, K. et al (2006) Cancer Cell 10: 529-41 118 Grade
4 Chin, K. et al (2006) Cancer Cell 10: 529-41 118 PAM50
5 Chin, K. et al (2006) Cancer Cell 10: 529-41 118 ER
6 Chin, K. et al (2006) Cancer Cell 10: 529-41 118 PgR
7 Desmedt, C. et al (2007) Clin Cancer Res 13: 3207-14 198 Grade
8 Desmedt, C. et al (2007) Clin Cancer Res 13: 3207-14 198 ER
9 Farmer, P. et al (2005) Oncogene 24: 4660-71 49
10 Finak,G. et al (2008) Nat Med 14: 518-27 59
a)
Gene Expression
ER+
ER-
b)
UBE2L6
SCYL3
IRF9
DDX5
PPM1D
ZNF292
CCdC36
IFITM1
IRF1
PARP14
TFAP4
c)
Percent Survival
100
80
60
40
20
Low Expression
High Expression
Time (Months)
d)
Deletion
Loss
No Change
Gain
Amplification
Gene Expression
aCGH

Figure 3

Figure 4


CDKN1B
UBQLN4
DRC1L
SNCA
WEE1
CUL4A
CDC2
MAPK3
RB1
CDK2
CUL1
RELA
YWHAG
MDM4
TP53
ATR
JAK1
AKT1
HSP90AB1
ATM
BRCA1
KAT5
SP1
POU3F2
FYN
PARP1
POU3F3
FOXO3
GRB2
SAFB2
GSN
KRT7
CCND1
EIF3I
MAP3K1
RPS6KA1