MetaGxData: Breast and Ovarian Clinically Annotated Transcriptomics Datasets

Deena Mohamad Ameen Gendoo [1,2], Natchar Ratanasirigulchai [1], Gregory M Chen[1], Levi Waldron [3 §], Benjamin Haibe-Kains [1,2,4 §]

[1] Bioinformatics and Computational Genomics Laboratory, Princess Margaret Cancer Center, University Health Network, Toronto, Ontario, Canada
[2] Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada
[3] City University of New York School of Medicine, New-York City, New-York, USA
[4] Department of Computer Science, University of Toronto, Toronto, Ontario, Canada

[§] Corresponding authors

ABSTRACT

A wealth of transcriptomic and clinical data on breast and ovarian cancers are under-utilized due to unharmonized data storage and format. We have developed the *MetaGxData* package compendium, which includes manually-curated and standardized clinical, pathological, survival, and treatment metadata across both breast and ovarian cancer microarray data. *MetaGxData* is the largest compendium of breast and ovarian microarray data to date, spanning 65 datasets and encompassing 13,756 samples. Standardization of metadata across the two cancer types promotes the use of their expression datasets in a variety of cross-tumour analyses, including identification of common biomarkers, establishing common patterns of co-expression networks, assessing the validity of prognostic signatures, and the identification of new consensus signatures that reflects upon common biological mechanisms. Here, we present our flexible framework, unified nomenclature, as well as applications that demonstrate the analytical power that is harnessed by combining breast and ovarian cancer datasets.

INTRODUCTION

Ovarian and breast cancers are among the leading causes of cancer deaths among women **(1-3)**, and recent studies have identified biological and molecular commonalities between them. Both cancers are part of hereditary syndromes related to mutations in a number of shared susceptibility genes that contribute to their carcinogenesis, including *BRCA1* and *BRCA2* **(3)**. As evidenced by epidemiological and linkage analysis studies, mutations and allelic loss in the *BRCA1* locus confers susceptibility to ovarian and early-onset breast cancer **(4, 6, 7)**. The *BRCA2* gene appears to account for a proportion of early-onset breast cancer that is roughly equal to that resulting from *BRCA1* **(5, 8)**. *BRCA2*-mutation carriers with mutations within the ovarian cancer cluster region have been observed to exhibit greater risk for ovarian cancer **(5)**. In addition to common susceptibility genes, both tumours may express a variety of common biomarkers that include hormone receptors, epithelial markers (e.g., cytokeratin 7, Ber-EP4), growth factor receptors (Her2/neu) and other surface molecules **(3)**.

Commonalities between breast and ovarian cancer have been observed not only for specific susceptibility genes, but at system-wide levels as well. Comprehensive molecular profiling efforts across transcriptomic profiles, copy-number landscapes, and mutational patterns of both cancers emphasize strong molecular commonalities between basal-like breast tumours with high-grade serous ovarian cancer (HG-SOC) **(2, 9)**. The growing list of parallels between Basal-like breast cancer and HG-SOC include *BRCA1* inactivation, high frequency of *TP53* mutations and *TP53* loss, chromosomal instability and widespread DNA copy number changes, high expression of *AKT3*, *MYC* amplification and high expression, and highly correlated mRNA expression profiles **(2, 9)**. Subtype-specific prognostic signatures also reveals strong similarities between prognostic pathways in basal-like cancer and ovarian cancer, while ER-negative and ER-positive breast cancer subtypes exhibit different prognostic signatures **(10)**. Collectively, these ongoing studies pave future directions for identification of shared prognostic and predictive biomarkers across breast and ovarian cancer molecular subtypes.

Continuous growth of breast and ovarian genome-wide profiling studies necessitates the development of large-scale computational frameworks that can store these complex data types, as well as integrate them for meta-analytical studies. Current bioinformatics initiatives provide extensive data repositories for microarray data retrieval and annotation of specific tumour types **(11-17)**. These resources are advantageous for analysis of single datasets, but fail to provide adaptable methods for integration and standardization across independent studies of single or multiple cancer types. This poses a challenge for meta-analytical investigations that address global patterns across multiple datasets. Some efforts towards alleviating this problem have focused on coupling microarray repositories with graphical user interfaces that address targeted biologic questions on collective transcriptome datasets **(18-20).**

An integrative framework is needed to harness the breadth of breast and ovarian transcriptomic and clinical data, and to serve as an integrative resource for integrative analysis across these aggressive and common women cancers. There are growing efforts towards the development of well-curated, standardized, and clinically relevant microarray repositories for breast cancer **(21-23)** and recently, for ovarian cancer **(24, 25)**. These studies provide a solid foundation towards the development of a controlled language for clinical annotations and standardized transcriptomic data representation across the two cancer types. Here, we have developed the *MetaGxData* package compendium, which includes manually-curated and standardized clinical, pathological, survival, and treatment metadata across both breast and ovarian cancer microarray data. *MetaGxData* is the largest, standardized compendium of breast and ovarian microarray data to date, spanning 61 datasets and encompassing 12,189 samples. Standardization of metadata across the two cancer types promotes the use of their expression and clinical data in a variety of cross-tumour analyses, including identification of common biomarkers, establishing common patterns of co-expression networks, assessing the validity of prognostic signatures, and the identification of new consensus signatures that reflects upon common biological mechanisms. In this paper, we present our flexible framework, unified nomenclature, as well as applications that demonstrate the analytical power that is harnessed by combining breast and ovarian cancer datasets.

METHODS & IMPLEMENTATION

The *MetaGxData* compendium integrates two packages containing fully curated and processed expression datasets for breast (*MetaGxBreast*) as well as ovarian (*MetaGxOvarian*) cancers. Our current framework extends upon the standardized framework we had already generated for curatedOvarianData **(24)**. Our proposed enhancements facilitate rapid and consistent maintenance of our data packages as newer datasets are added, and provides enhanced user-versatility in terms of data rendering across single or multiple datasets **[Figure 1]**.

Breast Cancer Data Acquisition

Breast cancer datasets were extracted from our previous meta-analysis of breast cancer molecular subtypes **(23)**, which includes 35 microarray datasets from a variety of commercially available microarray platforms published from 2002 to 2014. Additional datasets were extracted from the Gene Expression Omnibus (GEO) and manually curated. Gene expression and clinical annotation for Metabric **(27)** were additionally downloaded from EBI ArrayExpress and combined into a dataset of 2,136 samples. The cgdsr R package **(28)** was used to extract 1,098 tumour samples from The Cancer Genome Atlas (TCGA) **(2)**, and matching clinical annotations for these samples (clinical_patient_brca) were downloaded from the TCGA Data Matrix portal (https://tcga-data.nci.nih.gov/tcga/). Combining these studies produced a total of 39 breast cancer microarray expression datasets spanning 10,004 samples.

Ovarian Cancer Data Acquisition

Ovarian microarray expression datasets were obtained from our recent update of the curatedOvarianData data package, onto which we have added 5 expression datasets to the originally published version (24), for a total of 26 microarray datasets spanning 3,752 samples. To obtain these datasets we first used the curatedOvarianData pipeline to generate the "FULLcuratedOvarianData" version of the package, which differs from the public version in that probe sets for the same gene are not merged (https://bitbucket.org/lwaldron/curatedovariandata). Subsequently, the createEsets.R script and patientselection.config scripts were used with default parameters to generate the updated expression datasets of curatedOvarianData.

Curation

Gene and clinical annotations in the source data compendium were standardized within, but not between, breast and ovarian cancers, which could prevent easy integration across cancer types. We therefore developed semi-automatic curation scripts **(24)** to standardize gene and clinical annotations of all our datasets based on the nomenclature used in TCGA **(2)** **[Supplementary Table 1A, Supplementary Table 1B]**. Such annotations include a host of relevant categorical variables that reflect upon tumour histology (stage, grade, primary site, etc), as well as a number of categorical and numerical variables that are crucial for survival analysis and prognostication in breast and ovarian cancers **[Supplementary Figure 1, Supplementary Table 1]**. Most importantly, we have provided a number of comparable clinical variables across breast and ovarian cancer samples, such as age at diagnosis, tumour grade, or vital status **[Figure 3]**. We also provide tumour-specific and critical annotations for each tumour type, including biomarker identification status (HER2, ER, PR), as well as treatment information when available **[Figure 3]**.

Processing of Gene Expression Datasets

The processing of ovarian cancer microarray datasets was previously described **(24)**; breast cancer datasets were processed using in-house R/Bioconductor scripts (see Research Replicability). We used GEO platform descriptions as the primary source of probe and gene annotations when available, otherwise original annotations as published by the authors were used for non-standard gene expression profiling platforms. The full set of gene annotation platforms across all expression sets is provided in **Table 1**.

Gene symbols and Entrez Gene identifiers that matched the probeset ids of a given expression set were subsequently saved as part of the featureData (fData) pertaining to that expression set. For genes with multiple probesets, the collapseRows function of the WGCNA package (version 1.42) was used to identify representative probesets with the highest variance across all dataset, using Entrez Gene identifiers as group labels. Users can select these representative probesets during the probe-gene mapping procedure for their subsequent mapping to either Entrez Gene Ids or gene symbols.

Replicate Identification and Removal

Our pipeline handles fast and flexible generation of fully curated expression datasets while providing users the option of removal of sample replicates. To facilitate quick selection of duplicates for exclusion from the expression datasets, we first generated a pre-computed set of biological and/or technical replicates for a given sample across all datasets. For every tumour type, a merged expressionSet object was first created by combining all the genes and samples across all expression datasets available. Quantile normalization was employed to removal between-platform batch effects, using the *normalizeBetweenArrays* function from the limma package (version 3.22.4). Replicates were identified based on samples that shared correlated expression profiles, selecting replicates sharing a Spearman correlation ≥0.98. The list of replicates generated per samples was saved as pre-computed 'duplicates' column as part of the phenoData (pData) for every sample in every expression set. The development of a pre-computed duplicates list for every sample facilities rapid filtering of these samples from expression sets according to user specifications.

Versatile Generation of Finalized Expression Datasets

To facilitate meta-analyses involving either selected studies or all the datasets presented in the *MetaGxData* compendium packages, selection and filtering of the finalized, curated datasets can be performed according to user specifications. We adapted the *createEsetList.R* script and *patientselection.config* parameter file from curatedOvarianData to expedite sample selection based on properties of patients and datasets. Users can select representative probesets across all samples of a particular dataset prior to probe-gene identifier conversion. Users are also provided options for filtering samples or sample replicates across all datasets based on certain criteria, such as the prevalence of particular survival data. Importantly, we also provide users the ability to specifically select for only primary tumour samples or several tissue types (primary tumours, healthy tissue, etc.).

Collectively, our data compendium, referred to as *MetaGxData*, encompasses 65 processed gene expression datasets, containing in total 13,756 breast and ovarian samples **[Table 1, Figure 2]**. Expression datasets are represented as S4 ExpressionSet objects with attached clinical data (pData), and feature data (fData) for fast and flexible analysis with R/Bioconductor **(29)**.

USAGE AND UTILITY

MetaGxData serves as a large-scale, standardized compendium of genomic data for aggressive subtypes of women's cancers. The *MetaGxData* compendium is a flexible and adaptable resource that promotes selection of individual samples and datasets that match the users' requirements, and facilitates rapid integration of new datasets into the existing framework **[Table 2]**. These combined strengths promote rigorous management of the current datasets at the user and maintainer levels, as well as easy extension of newer studies into the package in future iterations of *MetaGxData*.

Enhancements in Data Assimilation and Annotation within *MetaGxData*

We have extended and standardized the *curatedOvarianData* processing pipeline across both breast and ovarian cancers. These enhancements have ensured that the *MetaGxData* processing framework is consistent across cancer types instead of singular cancer datasets. The processing framework includes the following features:

- Creation and standardization of a probe-gene mapping repository based on platform annotation files and original gene annotation files, across all datasets. Any new dataset that is incorporated into *MetaGxData* is matched against its corresponding annotation platform, facilitating quick mapping of gene annotations against probeset identifiers of the raw data.
- Pre-computed identification and filtering of the 'best probe' for probe-gene mapping of gene expression data. The selected probe is identified and labelled as part of the "fData" slot of the ExpressionSet object for a given dataset. During the probe-gene mapping procedure, the user has the option to select for these probesets and map them against Entrez Gene identifiers or gene symbols.
- Pre-computed and rapid identification of duplicated samples using a correlation-based matrix of the samples across all datasets. Duplicate samples are labelled in a 'duplicates' column within the fData slot of each dataset. When loading the gene expression datasets, the user has the option to remove these duplicate samples.

Box 1: List of salient features offered as part of *MetaGxData* curation and assimilation of datasets.

Analysis of the Prognostic Value of Individual Genes in Breast and Ovarian Cancer

The wealth and breadth of transcriptomic datasets in *MetaGxData* will serve as a solid framework for translational cancer research. As an example of the versatility of our packages, we conducted a meta-analysis of the prognostic value of well-studied and prognostic genes in both breast and ovarian cancers, using the *MetaGxBreast* and *MetaGxOvarian* packages **[Figure 4, Figure 5]**. A total of 13 prognostic genes were tested, including 7 for breast cancer (ESR1, ERBB2, STAT1, CASP3, PLAU, VEGF, and AURKA) **(30, 31)** and 6 for ovarian cancer (PTCH1, TGFBR2, CXCL14, POSTN, FAP, and NUAK1) **(32, 33)**. Datasets and samples were selected for meta-analysis dependent on the availability of survival events. For breast cancer, we used recurrence-free survival as the primary endpoint; when recurrence-free survival was unavailable, we used distant metastasis-free survival. This produced a cohort of 2,749 breast cancer patients from 16 datasets. For ovarian cancer, overall survival was used as the primary endpoint, resulting in a total of 2,630 ovarian cancer patients from 17 datasets. For the selected datasets the expression of each gene was median-dichotomized into high and low expression. Cox proportional hazards analysis was performed using the R package survcomp (version 1.18.0) **(34)** to estimate the prognostic value (hazard ratio) and significance (corresponding p-value) for each dichotomized gene expression. The R package metafor (version 1.9-7) **(35)** was used to produce a random-effects meta-analytical estimate of the log hazard ratio, using the restricted maximum-likelihood estimator.

An assessment of the prognostic value of single genes using the *MetaGxBreast* and *MetaGxOvarian* packages provides both dataset- and gene-centric views towards determining prognostic value of genes in breast and ovarian cancer **[Figure 4, Figure 5]**. Unsurprisingly, higher gene expression levels of the proliferation gene AURKA indicate poorer survival in breast cancer **[Figure 4, Supplementary Table 2]**. This supports our previous findings regarding the importance of this gene in biology-driven signatures of breast cancer, and its comparable prognostic effect with other multi-gene prognostic signatures **(23, 36)**. We have also observed that the NUAK1 genes exhibits worst prognosis in ovarian cancer **[Figure 5, Supplementary Table 3]**. We have previously demonstrated the utility of NUAK1 in the development of a debulking signature that can predict the outcome of cytoreductive surgery **(32)**.

Meta-Analysis of Gene Expression Prognosis Across Breast and Ovarian Cancer

Our single-gene prognostic analysis can be easily extended to a genome-wide meta-analysis. To this end, we determined the prognostic capability of 11,386 genes that are common to both the ovarian and breast cancer datasets **[Figure 6, Supplementary Table 4]**. We identified 58 genes that are significantly prognostic across both tumours (False Discovery Rate [FDR] < 5%). Of these, we identified 15 genes for which elevated expression values indicate worse prognosis in both cancers (HR>1), and 14 genes for which it indicates better prognosis (HR<1). Such findings will be integral in future studies of parallels between breast and ovarian cancer subtypes, for example, comparing basal-like breast cancer and high-grade serous ovarian cancer (HG-SOC)**(9)**.

Research Replicability

All the code required to reproduce the single-gene prognosis analysis, as well as the genome-wide meta-analysis, is publicly available on https://github.com/bhklab/MetaGxData. The procedure to setup the software environment and run our analysis pipeline is also provided on the github repository. This work complies with the guidelines proposed by Robert Gentleman **(37)** in terms of code availability and reproducibility of results.

DISCUSSION

Meta-analysis and data integration across breast and ovarian cancers is an area of intense research supporting common biology between these malignancies**.** We provide an integrative, standardized, and comprehensive platform to facilitate analysis between breast and ovarian cancer types and subtypes. This platform provides a flexible framework for data assimilation and unified nomenclature, with standardized data packages hosting the largest compendia of breast (10,004 samples) and ovarian (3,752 samples) cancer transcriptomic and clinical datasets available to date.

Integration of genomic data into standardized frameworks is challenged by the inconsistency of the clinical curations across datasets and across tumour types. Annotation of clinicopathological variables may vary widely due to different protocols in different laboratories, institutions, and across international boundaries. We have standardized, as much as possible, the catalog of variables within each tumour type. For characteristics pertaining to a specific tumour type, including ER, PGR, and HER2 IHC status in breast cancer samples, we have generated a semantic positive/negative variable to reflect IHC status. This facilitates searching across all patients irrespective of the original assay annotations, some of which may have been binary, or may have been on a 0-3 scale, or may have been qualitative. Similarly, a boolean y/n variable has been assigned to ovarian cancer patients to reflect whether they had been treated with platin, taxol, or neoadjuvant therapy. Many of the annotated variables (for example, those variables representing stage and tumour grade in MetaGxOvarian) have also been standardized to facilitate comparisions across multiple studies, and further analyzes using our previously developed packages (curatedOvarianData) have indicated good consistency across datasets, and ultimately facilitated uniform and consistent investigations on the prognostic effect of biomarkers in ovarian cancer survival **(38,39)**.

The *MetaGxData* processing framework standardized annotations within a specific tumour type and subsequently extended that across tumour types. To this end we have included relevant categorical variables that reflect upon tumour histology in both tumour types (for example, tumour grade). The package also hosts a number of categorical and numerical variables that are crucial for survival analysis and prognostication of breast and ovarian cancer. Our *MetaGxBreast* and *MetaGxOvarian* packages follow a unified framework that facilitates integration of oncogenomic and clinicopatholgical data. We have demonstrated how our packages facilitate easy meta-analysis of gene expression and prognostication in both breast and ovarian cancers. This has the potential to serve as an important resource and a primer towards the future development of cancer-specific compendia.

REFERENCES

1. Integrated genomic analyses of ovarian carcinoma. Nature. 2011;474(7353):609-15.

2. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61-70.

3. Davidson B, Stavnes HT, Holth A, Chen X, Yang Y, Shih Ie M, et al. Gene expression signatures differentiate ovarian/peritoneal serous carcinoma from breast carcinoma in effusions. Journal of cellular and molecular medicine. 2011;15(3):535-44.

4. Futreal PA, Liu Q, Shattuck-Eidens D, Cochran C, Harshman K, Tavtigian S, et al. BRCA1 mutations in primary breast and ovarian carcinomas. Science (New York, NY). 1994;266(5182):120-2.

5. Narod SA. Modifiers of risk of hereditary breast and ovarian cancer. Nature reviews Cancer. 2002;2(2):113-23.

6. Billack B, Monteiro ANA. BRCA1 in breast and ovarian cancer predisposition. Cancer Lett. 2005;227(1):1-7.

7. Ford D, Easton DF. The genetics of breast and ovarian cancer. British journal of cancer. 1995;72(4):805-12.

8. Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science (New York, NY). 1994;266(5182):66-71.

9. Bowtell DD. The genesis and evolution of high-grade serous ovarian cancer. Nature reviews Cancer. 2010;10(11):803-8.

10. Beck AH, Knoblauch NW, Hefti MM, Kaplan J, Schnitt SJ, Culhane AC, et al. Significance Analysis of Prognostic Signatures. PLoS Comput Biol. 2013;9(1):e1002875.

11. Cheng WC, Tsai ML, Chang CW, Huang CL, Chen CR, Shu WY, et al. Microarray meta-analysis database (M(2)DB): a uniformly pre-processed, quality controlled, and manually curated human clinical microarray database. BMC bioinformatics. 2010;11:421.

12. Coletta A, Molter C, Duque R, Steenhoff D, Taminau J, de Schaetzen V, et al. InSilico DB genomic datasets hub: an efficient starting point for analyzing genome-wide studies in GenePattern, Integrative Genomics Viewer, and R/Bioconductor. Genome biology. 2012;13(11):R104.

13. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research. 2002;30(1):207-10.

14. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, et al. ArrayExpress update--simplifying data submissions. Nucleic acids research. 2015;43(Database issue):D1113-6.

15. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. Nat Genet. 2006;38(5):500-1.

16. Wan Q, Dingerdissen H, Fan Y, Gulzar N, Pan Y, Wu TJ, et al. BioXpress: an integrated RNA-seq-derived gene expression database for pan-cancer analysis. Database : the journal of biological databases and curation. 2015;2015.

17. Kannan L, Ramos M, Re A, El-Hachem N, Safikhani Z, Gendoo DM, et al. Public data and open source tools for multi-assay genomic investigation of disease. Brief Bioinform. 2015.

18. Butti MD, Chanfreau H, Martinez D, Garcia D, Lacunza E, Abba MC. BioPlat: a software for human cancer biomarker discovery. Bioinformatics (Oxford, England). 2014;30(12):1782-4.

19. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer discovery. 2012;2(5):401-4.

20. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, et al. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. Neoplasia (New York, NY). 2007;9(2):166-80.

21. Madden SF, Clarke C, Gaule P, Aherne ST, O'Donovan N, Clynes M, et al. BreastMark: an integrated approach to mining publicly available transcriptomic datasets relating to breast cancer outcome. Breast cancer research : BCR. 2013;15(4):R52.

22. Planey CR, Butte AJ. Database integration of 4923 publicly-available samples of breast cancer molecular and clinical data. AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science. 2013;2013:138-42.

23. Haibe-Kains B, Desmedt C, Loi S, Culhane AC, Bontempi G, Quackenbush J, et al. A three-gene model to robustly identify breast cancer molecular subtypes. Journal of the National Cancer Institute. 2012;104(4):311-25.

24. Ganzfried BF, Riester M, Haibe-Kains B, Risch T, Tyekucheva S, Jazic I, et al. curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome. Database : the journal of biological databases and curation. 2013;2013:bat013.

25. Madden SF, Clarke C, Stordal B, Carey MS, Broaddus R, Gallagher WM, et al. OvMark: a user-friendly system for the identification of prognostic biomarkers in publically available ovarian cancer gene expression datasets. Molecular cancer. 2014;13:241.

26. Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, et al. International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. Database : the journal of biological databases and curation. 2011;2011:bar026.

27. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486(7403):346-52.

28. Skanderup AJ. cgdsr: R-Based API for accessing the MSKCC Cancer Genomics Data Server (CGDS). 2015.

29. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome biology. 2004;5(10):R80.

30. Desmedt C, Haibe-Kains B, Wirapati P, Buyse M, Larsimont D, Bontempi G, et al. Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. Clin Cancer Res. 2008;14(16):5158-65.

31. Wirapati P, Sotiriou C, Kunkel S, Farmer P, Pradervand S, Haibe-Kains B, et al. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. Breast cancer research : BCR. 2008;10(4):R65.

32. Riester M, Wei W, Waldron L, Culhane AC, Trippa L, Oliva E, et al. Risk prediction for late-stage ovarian cancer by meta-analysis of 1525 patient samples. Journal of the National Cancer Institute. 2014;106(5).

33. Waldron L, Haibe-Kains B, Culhane AC, Riester M, Ding J, Wang XV, et al. Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. Journal of the National Cancer Institute. 2014;106(5).

34. Schroder MS, Culhane AC, Quackenbush J, Haibe-Kains B. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. Bioinformatics (Oxford, England). 2011;27(22):3206-8.

35. Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. J Stat Softw. 2010;36(3):1-48.

36. Gendoo D, et al. . Genefu: an R/Bioconductor package for computation of gene ex-pression-based signatures in breast cancer. Bioinformatics (Oxford, England). 2015.

37. Gentleman R. Reproducible research: a bioinformatics case study. Stat Appl Genet Mol Biol. 2005;4:Article2.

38. Cheng, X., Lu, W. and Liu, M. Identification of homogeneous and heterogeneous variables in pooled cohort studies. Biometrics 2015;71(2):397-403.

39. Trippa, L., et al. Bayesian nonparametric cross-study validation of prediction methods. 2015:402-428.

FIGURE LEGENDS

Figure 1: Diagrammatic representation of the enhancements in data integration and annotation within the *MetaGxData* framework. The process of downloading a dataset, and subsequent curation, annotation and integration into *MetaGxData* is depicted.

Figure 2: Distribution of the samples and gene expression datasets in the (**A**) *MetaGxBreast* and (**B**) *MetaGxOvarian* packages.

Figure 3: Schematic representation of the clinical variables (pData) that are available across gene expression datasets in both *MetaGxBreast* (**A**) and *MetaGxOvarian* (**B**) packages. Each row represents a dataset, and each column represents a clinical variable. Stacked bar plots indicate the percentage of samples in every dataset annotated with a particular variable designation. Continuous numeric values are represented as bar plots. Clinical variables common to both packages are first represented (left). Different variables relating to Treatment or Histology are highlighted in boxes.

Figure 4: Assessment of the prognostic value of seven key genes in breast cancer, using the *MetaGxBreast* package. (**A**) Heatmap representation hazard ratios for each gene, across 18 gene expression datasets. The estimate is presented as a log ratio, using low gene expression as baseline. Ratios greater than 1 (blue) indicate worse prognosis for elevated expression levels of that gene in the respective datasets. (**B**) Random effects meta-estimates of log hazard ratio, using the restricted maximum-likelihood estimator for each gene, pooled across all gene expression datasets. (**C**) Kaplan-Meier curve of the most prognostic gene, in this case AURKA, indicating survival across patients with low- or high-gene expression levels across all patients.

Figure 5: Assessment of the prognostic value of six key genes in ovarian cancer, using the *MetaGxOvarian* package. (**A**) Heatmap representation hazard ratios for each gene, across 17 gene expression datasets. The estimate is presented as a log ratio, using low gene expression as baseline. Ratios greater than 1 (blue) indicate worse prognosis for elevated expression levels of that gene in the respective datasets. (**B**) Random effects meta-estimates of log hazard ratio, using the restricted maximum-likelihood estimator for each gene, pooled across all gene expression datasets. (**C**) Kaplan-Meier curve of the most prognostic gene, in this case NUAK1, indicating survival across patients with low- or high-gene expression levels across all patients.

Figure 6: Genome-wide assessment of the prognostic value of 11,346 genes common to both the *MetaGxBreast* and *MetaGxOvarian* datasets. A Venn diagram of significant genes (FDR<5%) in each tumour following calculation of the Hazards Ratio is indicated (top). A total of 509 and 1,275 significantly prognostic genes were identified for ovarian

and breast cancer, respectively. Common significant genes between both tumour types (n=58) were further subdivided by their log hazard ratio, for each tumour type. Genes for which elevated expression levels are prognostic (HR>1) across both tumours, or genes for which down-regulated expression is prognostic (HR<1) are indicated.

TABLE LEGENDS

Table 1: List of Datasets that constitute the *MetaGxBreast* **(A)** and *MetaGxOvarian* **(B)** packages. Information on the platforms used and total gene and sample counts are provided.

Table 2: List of salient features offered as part of *MetaGxData* curation and dataset assimilation.

SUPPLEMENTARY FIGURE LEGENDS

Supplementary Figure 1: Heatmap representation of clinical variables availability across gene expression datasets of *MetaGxBreast* and *MetaGxOvarian*. Datasets are represented as rows and clinical variables as columns. The percentage of samples in each dataset that is annotated with a particular variable is represented.

SUPPLEMENTARY TABLE LEGENDS

Supplementary Table 1: Explanation of curated clinical annotations (phenotype data variables) in *MetaGxBreast* **(A)** and *MetaGxOvarian* **(B)**. Common variables to both datasets are highlighted.

Supplementary Table 2: Genome-wide analysis of the prognostic value of 11346 genes in breast and ovarian gene expression datasets. **(A)** List of the computed Hazard Ratio of all genes, using *MetaGxBreast*. **(B)** List of the computed Hazard Ratio of all genes, using *MetaGxOvarian*. **(C)** Hazard ratio of 58 common prognostic genes, using *MetaGxOvarian*. **(D)** Hazard ratio of 58 common prognostic genes, using *MetaGxBreast*.

SUPPLEMENTARY FILES

Supplementary File 1: Forest plot of hazard ratios and survival plot of 7 key prognostic genes across 18 genesets of *MetaGxBreast* package.

Supplementary File 2: Forest plot of hazard ratios and survival plot of 6 key prognostic genes across 17 genesets of the *MetaGxOvarian* package.

Enhancements in Data Assimilation and Annotation within *MetaGxData*

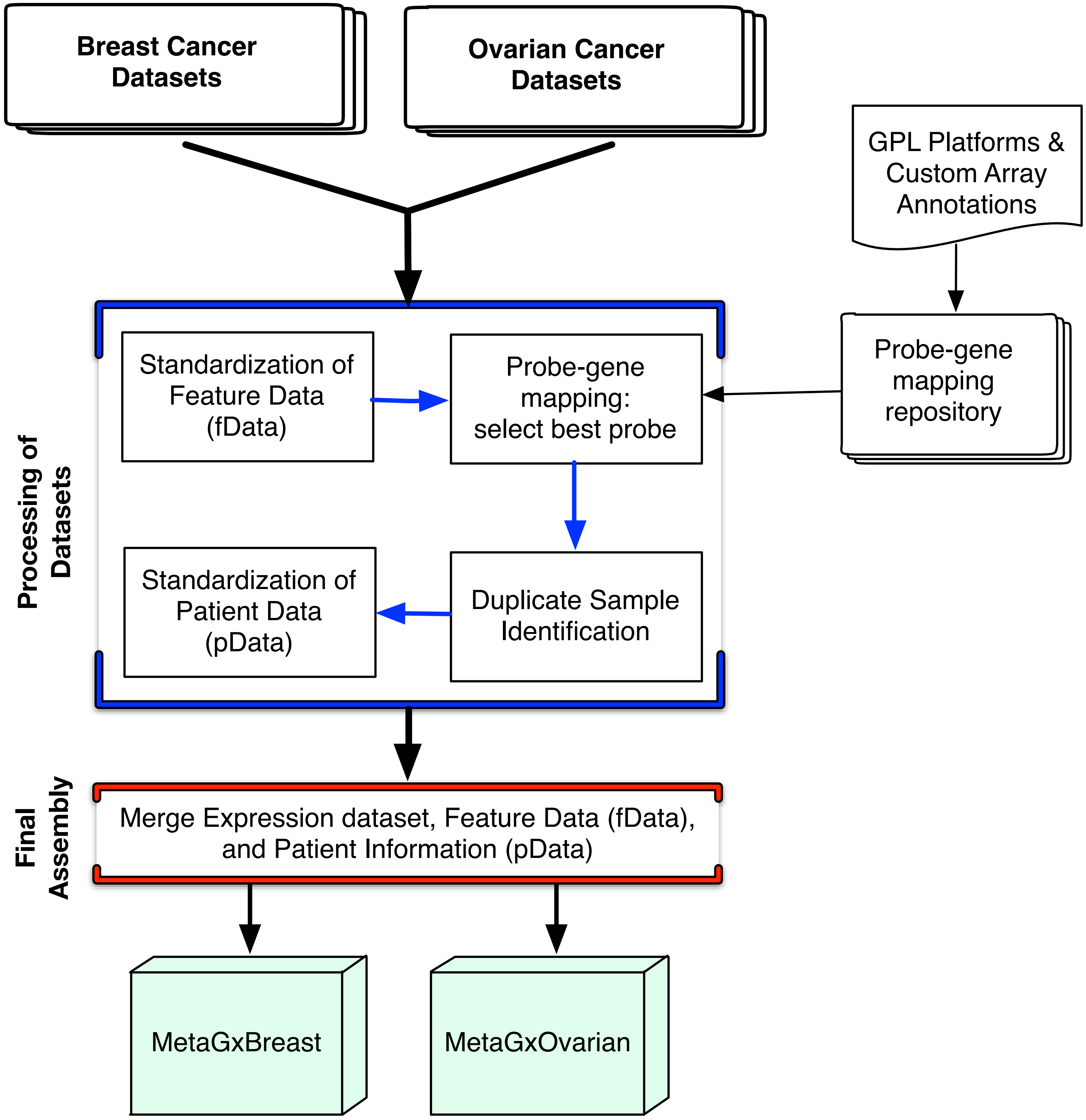

Breast Cancer Datasets
Ovarian Cancer Datasets
GPL Platforms & Custom Array Annotations
Probe-gene mapping repository
Processing of Datasets
Standardization of Feature Data (fData)
Probe-gene mapping: select best probe
Duplicate Sample Identification
Standardization of Patient Data (pData)
Final Assembly
Merge Expression dataset, Feature Data (fData), and Patient Information (pData)
MetaGxBreast
MetaGxOvarian

Figure 1: Diagrammatic representation of the enhancements in data assimilation and annotation within the MetaGxData framework. The process of downloading a dataset, and subsequent curation, annotation and integration into MetaGxData is depicted.

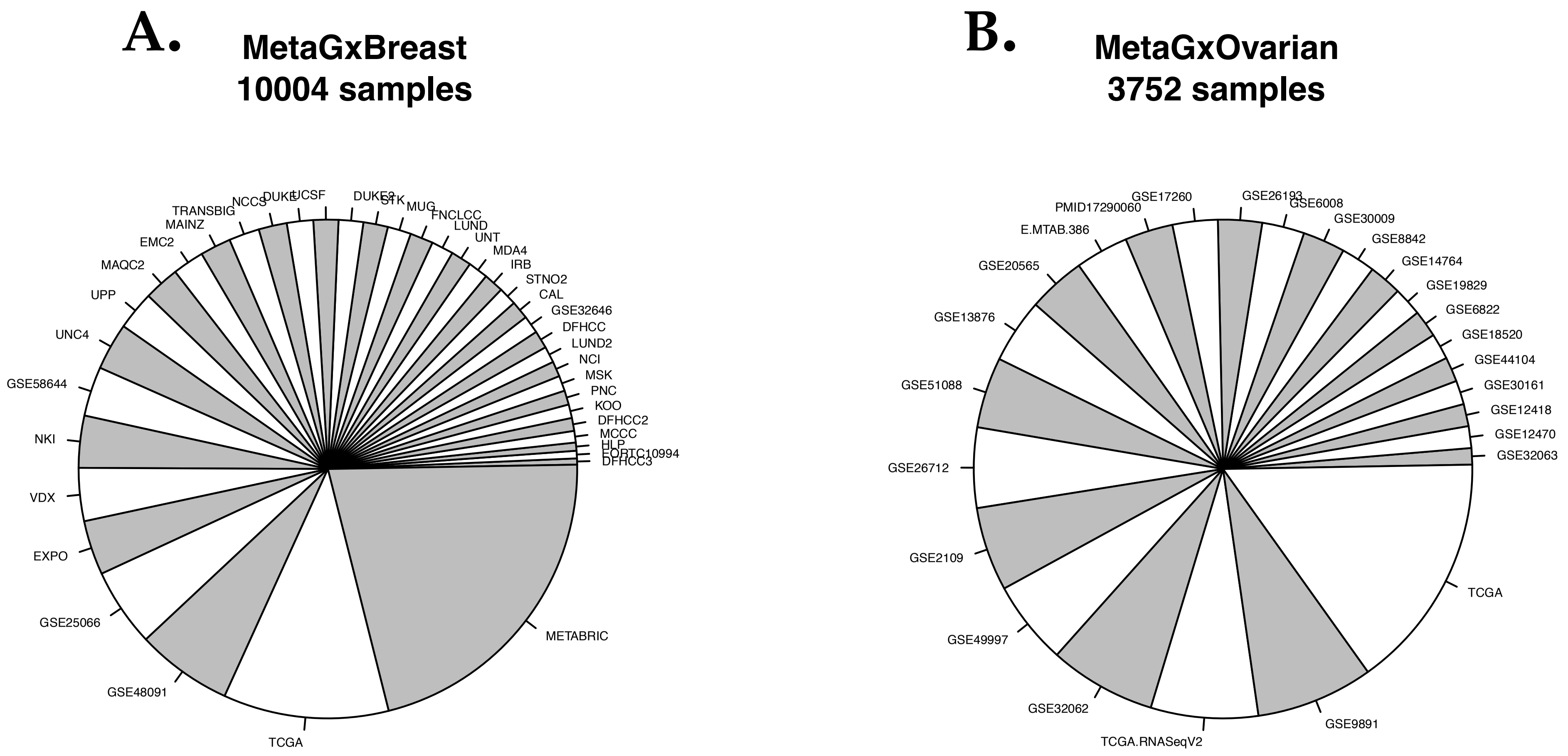


Figure 2: Distribution of the samples and gene expression datasets in the (A) MetaGxBreast and (B) MetaGxOvarian packages.

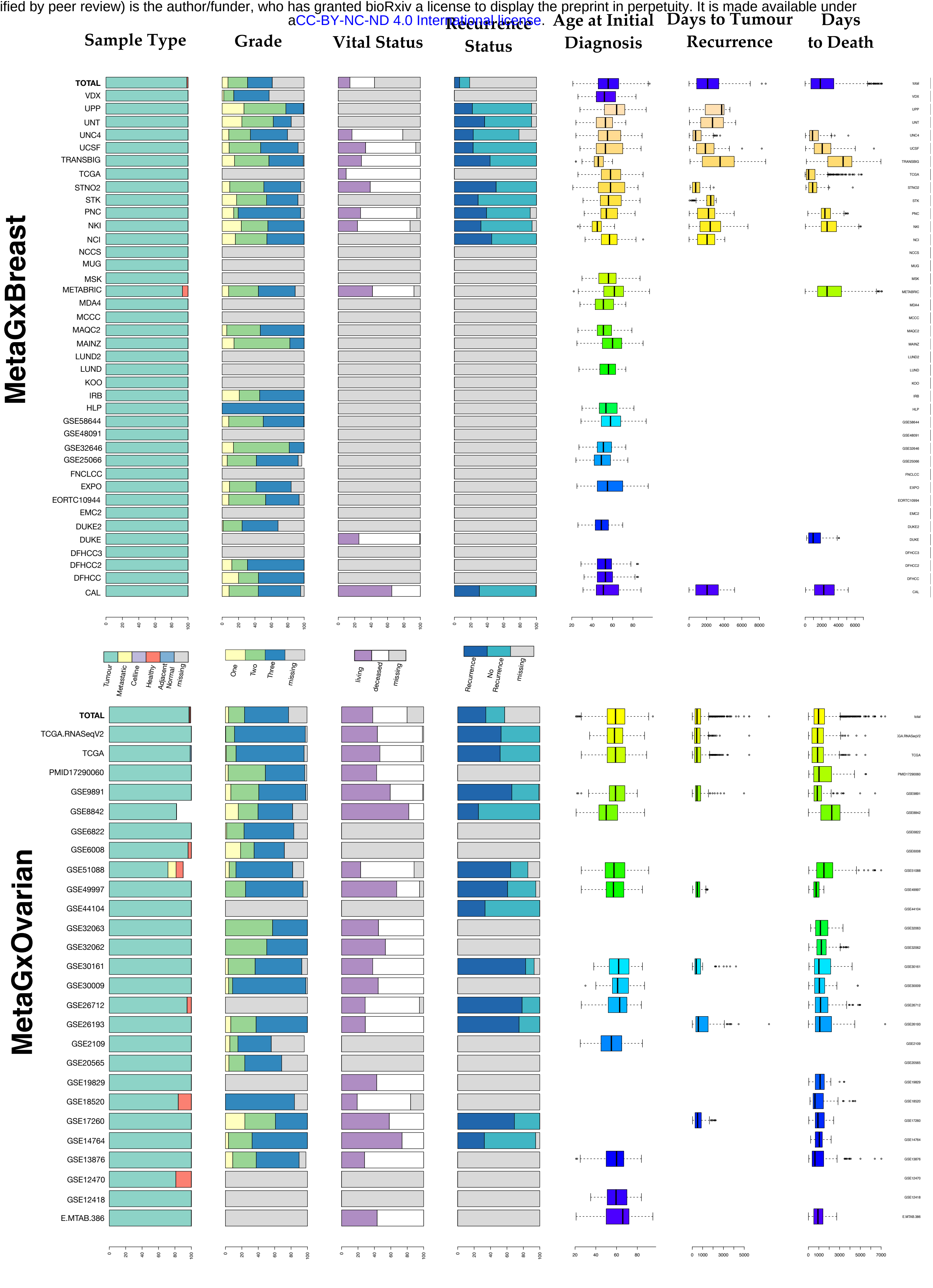


Figure 3: Schematic representation of the clinical variables (pData) that are available across gene expression datasets in both MetaGxBreast (A) and MetaGxOvarian (B) packages. Each row represents a dataset, and each column represents a clinical variable. Stacked bar plots indicate the percentage of samples in every dataset annotated with a particular variable designation. Continuous numeric values are represented as bar plots. Clinical variables common to both packages are first represented (left). Different variables relating to Treatment or Histology are highlighted in boxes.

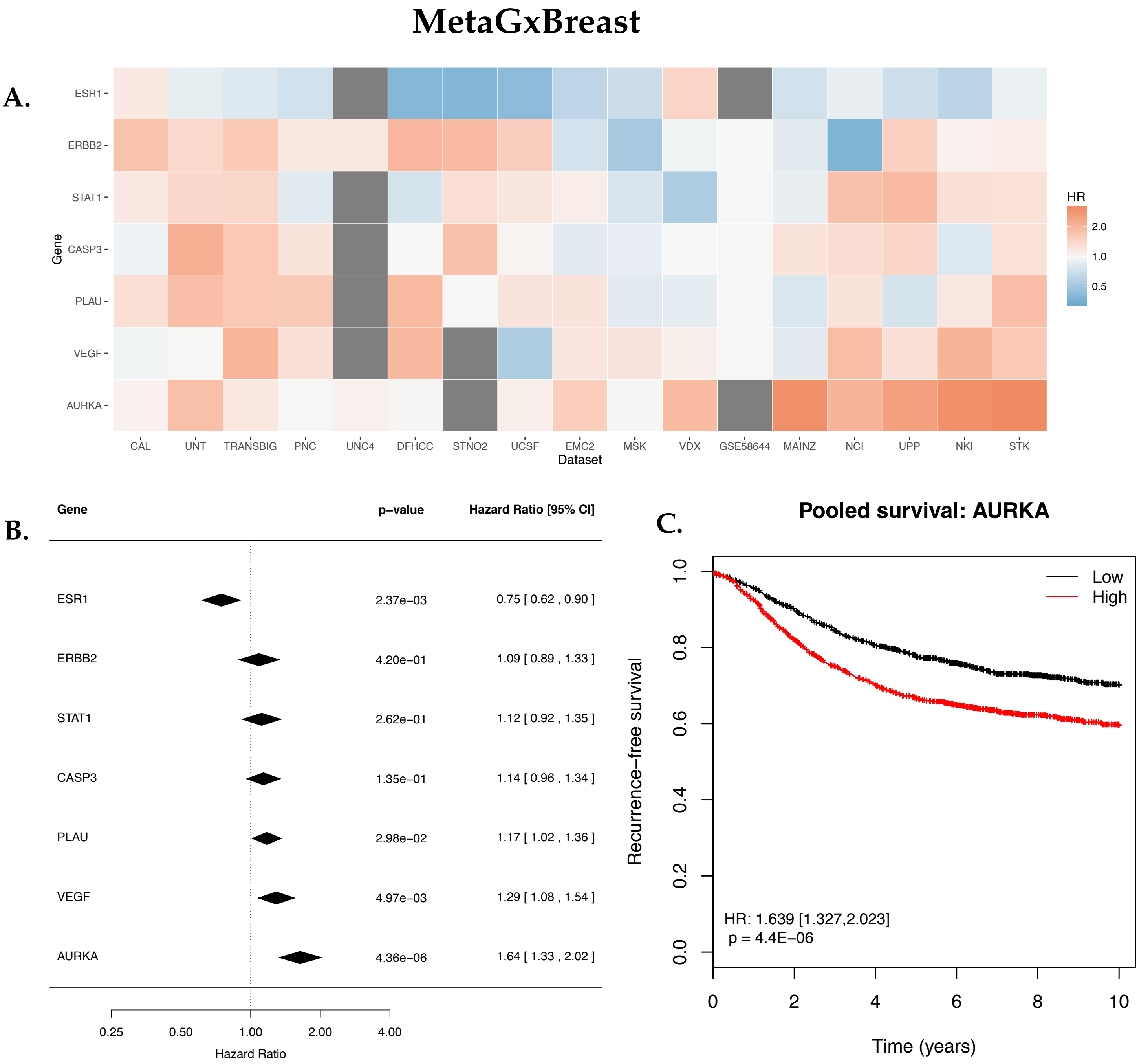


Figure 4: Assessment of the prognostic value of 7 key genes in breast cancer, using the MetaGxBreast package. (A) Heatmap representation hazard ratios for each gene, across 18 gene expression datasets. The estimate is presented as a log ratio, using a low baseline hazard. Ratios greater than 1 (blue) indicate worse prognosis for elevated expression levels of that gene in the respective datasets. (B) Random effects estimates of log hazard ratio, using the restricted maximum-likelihood estimator for each gene, pooled across all gene expression datasets. (C) Kaplan- Meier curve of the most prognostic gene, in this case AURKA, indicating survival across patients with high- or low-gene expression levels across all patients.

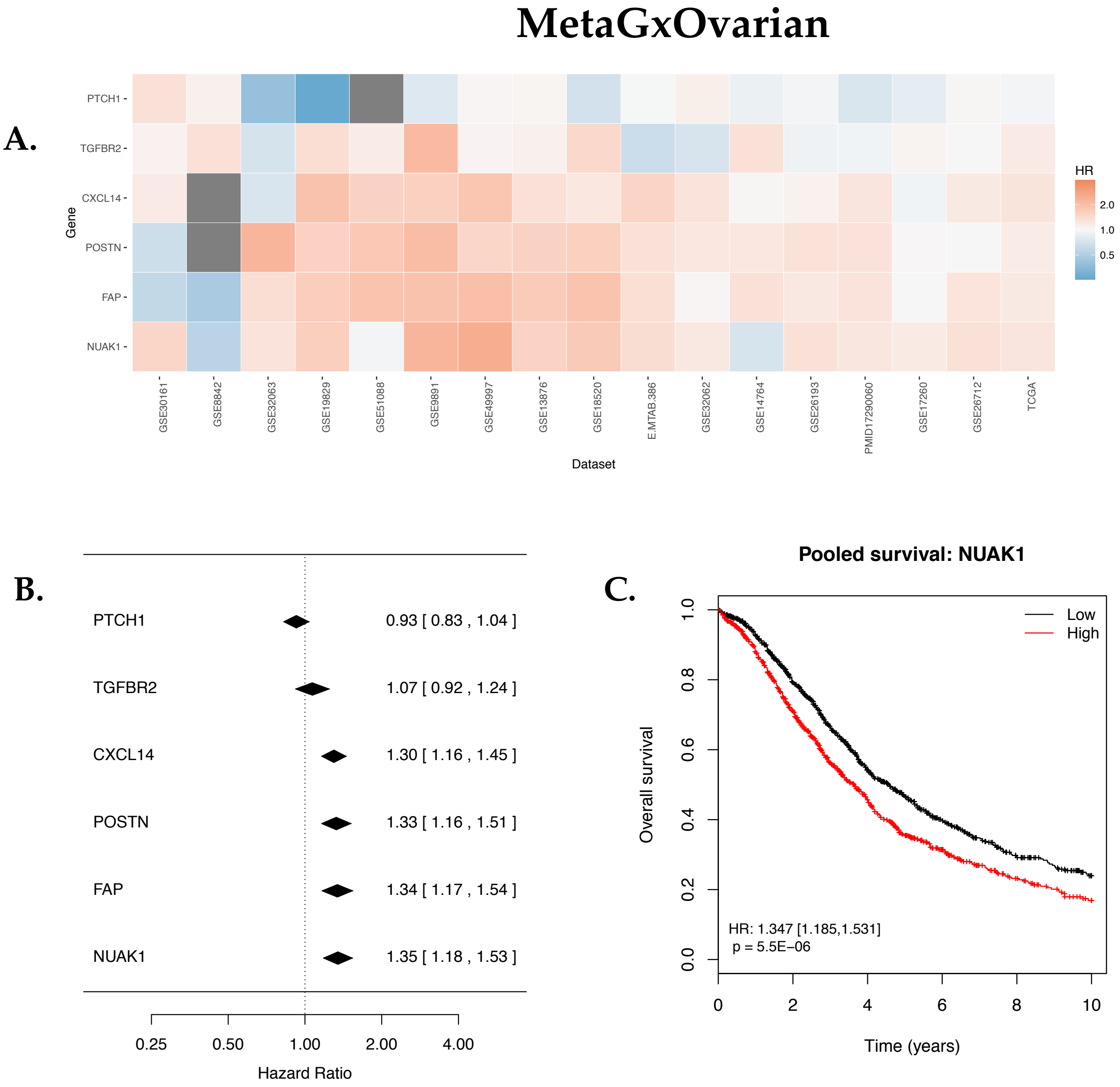


Figure 5: Assessment of the prognostic value of 6 key genes in ovarian cancer, using the MetaGxOvarian package. (A) Heatmap representation hazard ratios for each gene, across 17 gene expression datasets. The estimate is presented as a log ratio, using a low baseline hazard. Ratios greater than 1 (blue) indicate worse prognosis for elevated expression levels of that gene in the respective datasets. (B) Random effects estimates of log hazard ratio, using the restricted maximum-likelihood estimator for each gene, pooled across all gene expression datasets. (C) Kaplan- Meier curve of the most prognostic gene, in this case NUAK1, indicating survival across patients with high- or low-gene expression levels across all patients.

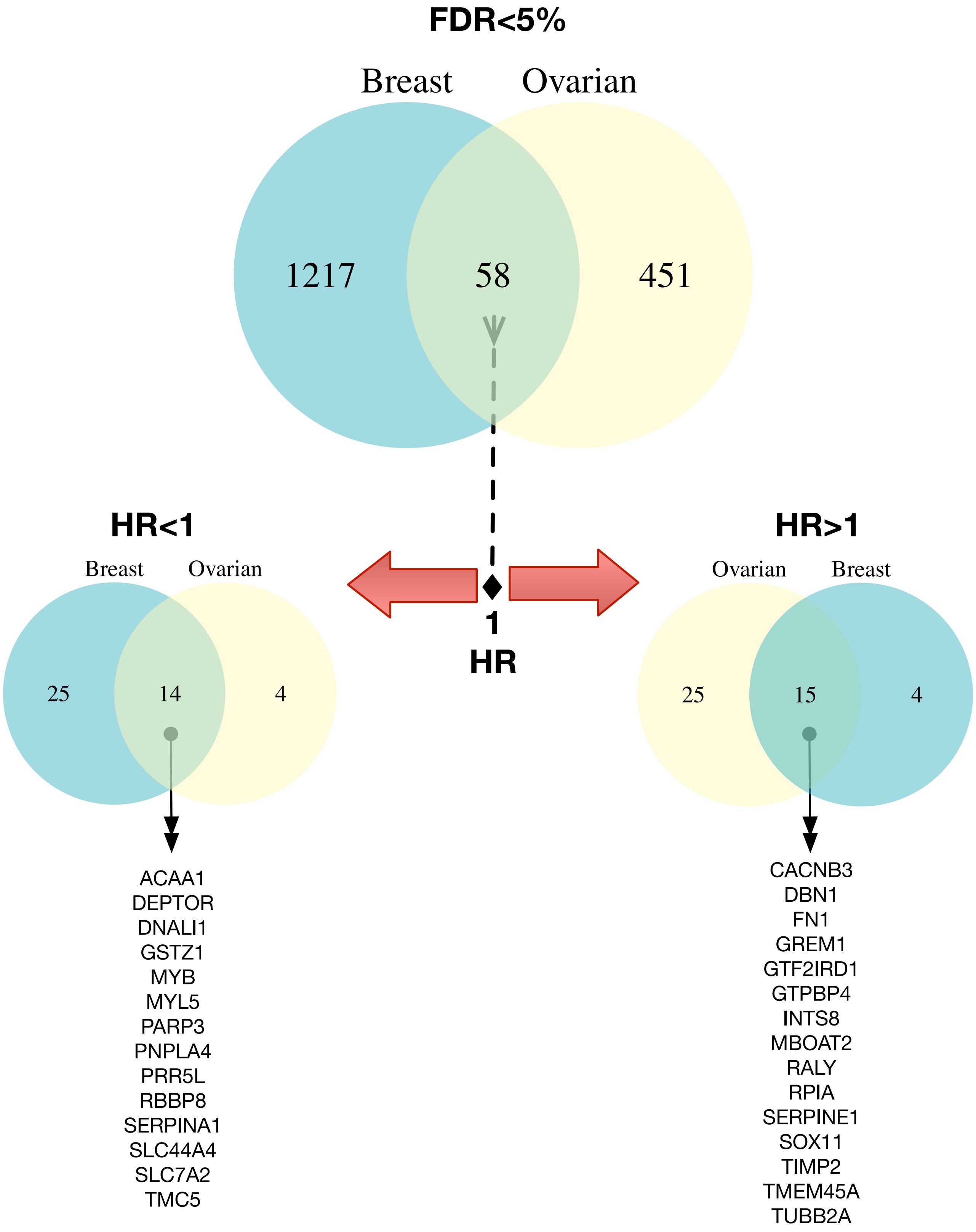


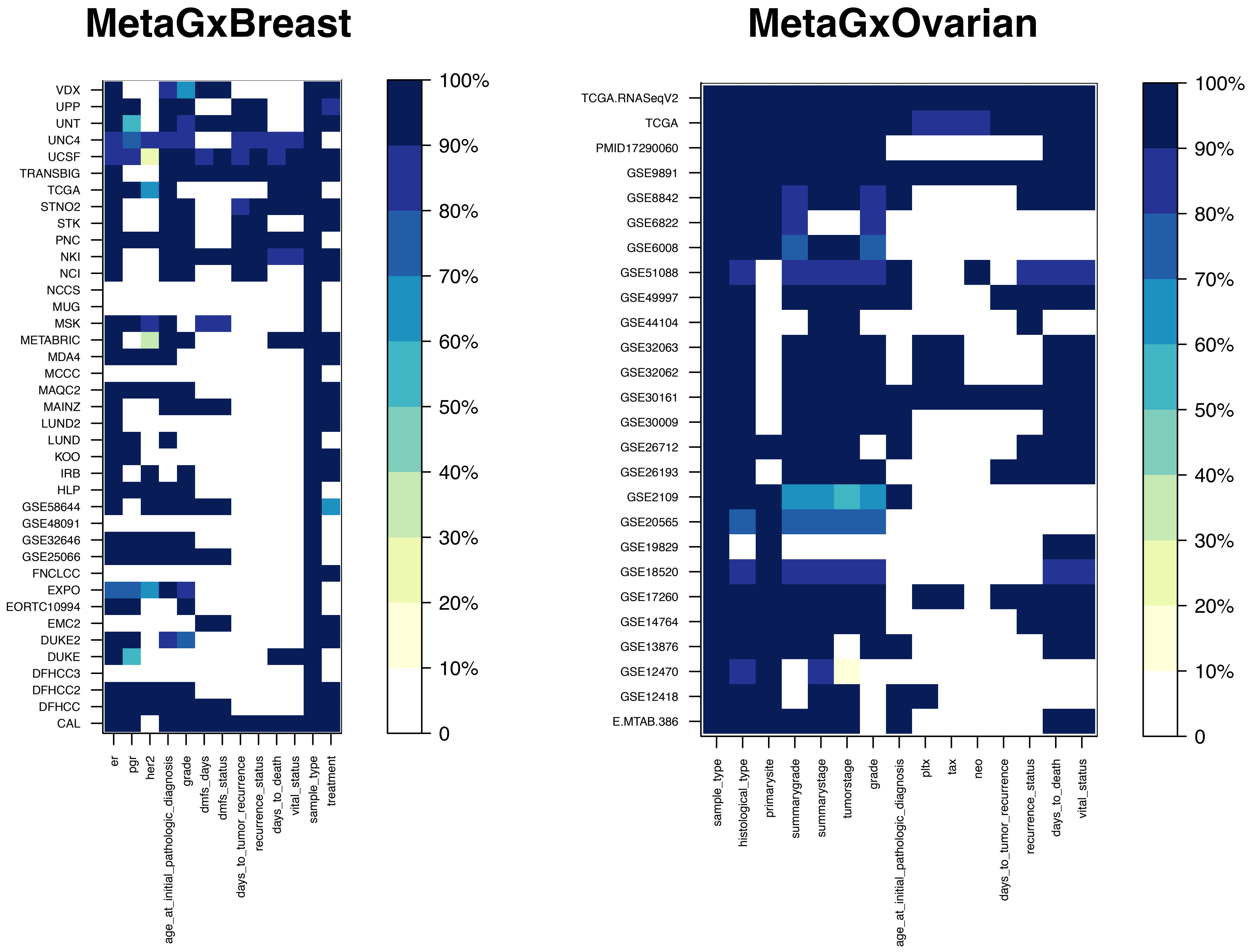
Figure 6: Genome-wide assessment of the prognostic value of 11,346 genes common to both the MetaGxBreast and MetaGxOvarian datasets. A Venn diagram of significant genes (FDR<0.05) in each tumour following calculation of the Hazards Ratio is indicated (top). A total of 509 and 1,275 significantly prognostic genes were identified for ovarian and breast cancer, respectively. Common significant genes between both tumour types (n=58) were further subdivided by their log hazard ratio, for each tumour type. Genes for which elevated expression levels are prognostic (HR>1) across both tumours, or genes for which down-regulated expression is prognostic (HR<1) are indicated.

	Dataset	PMID	Dataset_accessio	Platform_description	Platform	Notes	patients*	genes***	probes**	Publication_
1	CAL	17157792	E-TABM-158	Affymetrix HGU	Affymetrix HGU	Dataset of breast cancer patients from the Un	118	12688	21169	2006
2	DFHCC	20098429	GSE19615	Affymetrix HGU	GPL570	Dana-Farber Harvard Cancer Center (United St	115	20282	42447	2010
3	DFHCC2	20100965	GSE18864	Affymetrix HGU	GPL570	Dana-Farber Harvard Cancer Center (United St	84	20282	42447	2010
4	DFHCC3	16473279	GSE3744	Affymetrix HGU	GPL570	Dana-Farber Harvard Cancer Center (United St	40	20282	42447	2010
5	DUKE	16273092	GSE3143	Affymetrix HGU95	GPL8300	Duke university hospital (United States)	171	8836	12085	2006
6	DUKE2	18024211	GSE6861	Affymetrix X3P	GPL1352	Duke university hospital (United States)	160	19700	45490	2007
7	EMC2	19421193	GSE12276	Affymetrix HGU	GPL570	Erasmus Medical Center (The Netherlands)	204	20282	42447	2009
8	EORTC10994	15897907	GSE1561	Affymetrix HGU	GPL96	Trial number 10994 from the European Organ	49	12752	20967	2005
9	EXPO	Erin Curley (	GSE2109	Affymetrix HGU	GPL570	Expression project for oncology, large dataset	353	20282	42447	NA
10	FNCLCC	17659439	GSE7017	In-house cDNA	GPL4819	F_d_ration Nationale des Centres de Lutte cor	150	5107	6064	2008
11	HLP	19688261	E-TABM-543	Illumina	Illumina	University Hospital La Paz (Spain)	53	19451	26536	2010
12	IRB	18297396	GSE5460	Affymetrix HGU	GPL570	Dana Farber Cancer Institute	129	20282	54675	2006
13	KOO	12747878	Authors' website	Affymetrix HGU95	Affymetrix HGU95	Koo Foundation Sun Yat-Sen Cancer Centre (Ta	88	254	280	2003
14	LUND	18430221	GSE31863	Swegene	GPL14374	Lund University Hospital (Sweden)	143	10388	11154	2008
15	LUND2	17452630	GSE5325	Swegene	GPL3883	Lund University Hospital (Sweden)	105	7913	22008	2007
16	MAINZ	18593943	GSE11121	Affymetrix HGU	GPL96	Mainz hospital (Germany)	200	12752	20967	2008
17	MAQC2	20064235	GSE20194	Affymetrix HGU	GPL96///GPL570///GPL1	Microarray quality control consortium (United	230	12752	20967	2010
18	MCCC	19960244	GSE19177	Illumina	GPL6106	Peter MacCallum Cancer Centre (Australia)	75	14953	19048	2010
19	MDA4	16896004	MDACC DB	Affymetrix HGU	Affymetrix HGU	MD Anderson Cancer Center (United States)	129	12688	21169	2006
20	MSK	16049480	GSE2603	Affymetrix HGU	GPL96	Memorial Sloan-Kettering (United States)	99	12752	20967	2005
21	MUG	18592372	GSE10510	Operon	GPL6486	Medical University of Graz (Austria)	152	10715	14288	2009
22	NCCS	18636107	GSE5364	Affymetrix HGU	GPL96	National Cancer Centre of Singapore (Singapor	183	12752	20967	2008
23	NCI	12917485	Authors' website	In-house cDNA	In-house cDNA	National Cancer Institute (United States)	99	4112	5154	2003
24	NKI	12490681, 1	Rosetta Inpharma	Agilent	Agilent	National Kanker Instituut (The Netherlands)	337	13116	14960	2002
25	PNC	21910250	GSE20711	Affymetrix HGU	GPL570		92	20282	42447	2011
26	STK	16280042	GSE1456	Affymetrix HGU	GPL97///GPL96	Stockholm	159	18434	36178	2005
27	STNO2	12829800	GSE4382	In-house cDNA	GPL180///GPL2776///G	Stanford/Norway (United States and Norway)	118	3228	3663	2003
28	TRANSBIG	17545524	GSE7390	Affymetrix HGU	GPL96	Dataset collected by the TransBIG consortium	198	12752	20967	2007
29	UCSF	17428335, 1	Authors' website	In-house cDNA	In-house cDNA	University of California, San Francisco (United	162	6275	8015	2007
30	UNC4	20813035	GSE18229	Agilent	GPL885///GPL887///GP	University of Northern California (United State	305	5025	5420	2007
31	UNT	16478745, 1	GSE2990	Affymetrix HGU	Affymetrix HGU	Cohort of untreated breast cancer patients frc	133	18009	36084	2010
32	UPP	16141321	GSE3494	Affymetrix HGU	GPL570	Uppsala hospital (Sweden)	251	18434	36178	2005
33	VDX	17420468, 1	GSE2034/GSE532	Affymetrix HGU	Affymetrix HGU	Veridex (The Netherlands)	344	12688	21169	2007
34	METABRIC	22522925		METABRIC	METABRIC		2136	24924	36155	2012
35	TCGA	23000897		TCGA	TCGA	The Cancer Genome Atlas	1073	19405	19504	2012
36	GSE25066	21558518	GSE25066	Affymetrix HGU	GPL96	Nuvera Biosciences	508	12752	20967	2010
37	GSE32646	22320227	GSE32646	Affymetrix HGU	GPL570	Osaka University	115	20282	42437	2012
38	GSE58644	25284793	GSE58644	Affymetrix Gene1.0ST	GPL6244	McGill University	321	20202	21462	2014
39	GSE48091	26077471	GSE48091	Affymetrix RSTA	GPL10379	Karolinska Institutet	623	12917	23246	2015

Table 1A: List of Datasets that constitute the *MetaGxBreast*

	Dataset	PMID	Dataset_acc	Platform_description	Platform	patients*	genes***	probes**	Pulication_da
1	E.MTAB.386	22348002	E.MTAB.386	III. HumanRef-8 v2	GPL6104	129	10572	12449	2012
2	GSE12418	16996261	GSE12418	SWEGENE v2.1.1_27k	GPL5886	54	9544	11276	2006
3	GSE12470	19486012	GSE12470	Agilent G4110b	GPL887	53	13667	15952	2008
4	GSE13876	19192944	GSE13876	Operon Human v3	GPL7759	157	13846	20870	2009
5	GSE14764	19294737	GSE14764	Affymetrix U133A	GPL96	80	12752	20967	2009
6	GSE17260	20300634	GSE17260	Agilent G4112a	GPL6480	110	19596	30936	2010
7	GSE18520	19962670	GSE18520	Affymetrix U133 Plus 2.0	GPL570	63	20282	42447	2009
8	GSE19829	20547991	GSE1929	Affymetrix U133 Plus 2.0/Affymetr	GPL570///GP	70	20339	54253	2010
9	GSE20565	20492709	GSE20565	Affymetrix U133 Plus 2.0	GPL570///GP	140	20282	42447	2010
10	GSE2109	Erin Curley e	GSE2109	Affymetrix U133 Plus 2.0	GPL570	204	20282	42447	NA
11	GSE26193	22101765	GSE26193	Affymetrix U133 Plus 2.0	GPL570	107	20282	42447	2011
12	GSE26712	18593951	GSE26712	Affymetrix U133A	GPL96	195	12752	20967	2008
13	GSE30009	22492981	GSE30009	TaqMan qRT-PCR 380	GPL13728	103	359	363	2012
14	GSE30161	22348014	GSE30161	Affymetrix U133 Plus 2.0	GPL570	58	20282	42447	2012
15	GSE32062	22241791	GSE32062	Agilent G4112a	GPL570///GP	260	19596	30936	2012
16	GSE32063	22241791	GSE32063	Agilent G4112a	GPL6480	40	19596	30936	2012
17	GSE44104	23934190	GSE44104	Affymetrix U133 Plus 2.0	GPL570	60	20282	42447	2014
18	GSE49997	22497737	GSE49997	ABI Human Genome Survey Microa	GPL2986	204	16760	18439	2012
19	GSE51088	24368280	GSE51088	Agilent G4110B	GPL7264	172	16747	18703	2014
20	GSE6008	17418409	GSE6008	Affymetrix U133A	GPL96	103	12752	20967	2007
21	GSE6822	Jaroslav Petr	GSE6822	Affymetrix Hu6800	GPL80	66	5342	6407	NA
22	GSE8842	19047114	GSE8842	Agilent G4100A	GPL5689	83	6251	7809	2008
23	GSE9891	18698038	GSE9891	Affymetrix U133 Plus 2.0	GPL570	285	20282	42447	2008
24	PMID158975	15897565	PMID158975	Affymetrix U133A	GPL96	63	12752	20967	NA
25	PMID172900	17290060	PMID172900	Affymetrix U133A	GPL96	117	12752	20967	NA
26	PMID193184	19318476	PMID193184	Affymetrix U133A	GPL96	42	12752	20967	NA
27	TCGA		TCGA	Affymetrix HT U133A	GPL4685	578	12833	21260	NA
28	TCGA.RNASeqV2		TCGA.RNASe	Illumina	Illumina	261	20446	20446	NA

Table 1B: List of Datasets that constitute the *MetaGxOvarian*



Supplementary Figure 1: Heatmap representation of clinical variables availability across gene expression datasets of MetaGxBreast and MetaGxOvarian. Datasets are represented as rows and clinical variables as columns. The percentage of samples in each dataset that is annotated with a particular variable is represented.