curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome

Benjamin Frederick Ganzfried¹,², Markus Riester¹,², Benjamin Haibe-Kains³, Thomas Risch¹, Svitlana Tyekucheva¹,², Ina Jazic⁴, Xin Victoria Wang¹,², Mahnaz Ahmadifar¹, Michael J. Birrer⁴, Giovanni Parmigiani¹,², Curtis Huttenhower² and Levi Waldron¹,²*

¹Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, 02115, ²Department of Biostatistics, Harvard School of Public Health, Boston, MA, 02115, USA, ³Bioinformatics and Computational Genomics Laboratory, Institut de recherches cliniques de Montréal, Montréal, QC, H2W 1R7, Canada and ⁴Center for Cancer Research, Massachusetts General Hospital, Boston, MA, 02114, USA

These authors contributed equally to this work.

*Corresponding author: Email: levi@jimmy.harvard.edu

Submitted 10 January 2013; Revised 16 February 2013; Accepted 19 February 2013


This article introduces a manually curated data collection for gene expression meta-analysis of patients with ovarian cancer and software for reproducible preparation of similar databases. This resource provides uniformly prepared microarray data for 2970 patients from 23 studies with curated and documented clinical metadata. It allows users to efficiently identify studies and patient subgroups of interest for analysis and to perform meta-analysis immediately without the challenges posed by harmonizing heterogeneous microarray technologies, study designs, expression data processing methods and clinical data formats. We confirm that the recently proposed biomarker CXCL12 is associated with patient survival, independently of stage and optimal surgical debulking, which was possible only through meta-analysis owing to insufficient sample sizes of the individual studies. The database is implemented as the curatedOvarianData Bioconductor package for the R statistical computing language, providing a comprehensive and flexible resource for clinically oriented investigation of the ovarian cancer transcriptome. The package and pipeline for producing it are available from http://bcb.dfci.harvard.edu/ovariancancer.

Database URL: http://bcb.dfci.harvard.edu/ovariancancer

Introduction

A wealth of genomic data, in particular microarray data, is publicly available through diverse online resources. Major databases of gene expression data, e.g. the Gene Expression Omnibus (GEO) (1) and ArrayExpress (2), offer the potential to identify sets of genes predictive of cancer survival and of patient resistance to chemotherapy using thousands of samples from multiple laboratories. Such high numbers of samples are needed to robustly identify and validate gene signatures for incorporation into routine clinical practice (3). However, inconsistent formatting among database interfaces, expression data storage and clinical metadata annotations present formidable obstacles to making efficient use of these resources.

Existing resources aiming to make large-scale high-dimensional analysis across multiple studies tend to serve only a few specifically targeted needs. To develop reproducible biomarker discovery methods appropriate for clinical translation, a data resource must be accurate and retain clinical variables of known importance as much as possible. The insilicoDB (4) project provides many curated gene expression data sets; however, it is not a focused resource in terms of retention or quality assurance of clinical
annotations, or retention of all relevant data sets and clinical variables for any one cancer type. The other major database of curated gene expression studies, the Gene Expression Atlas (2), provides machine- rather than manually annotated data, resulting in reduced consistency of annotation across studies. These are among the only databases that offer basics such as uniform gene identifiers to enable cross-study analysis, and then for only the most common microarray technologies. Carey et al. (5) describe a framework for the curation, annotation and storage of microarray and high-throughput data in general. This framework allows, for example, institutions to provide researchers access to in-house and public data in a standardized and convenient fashion. However, there is no existing database that provides these resources for ovarian cancer.

Ovarian cancer is the fifth-leading cause of cancer deaths among women (6) and has been the focus of numerous clinical transcriptome investigations. The curatedOvarianData database is the result of a focused effort to enable meta-analysis of these studies and to provide the highest quality and most comprehensive gene expression data resource for any cancer. It provides standardized gene expression and clinical data for 2970 ovarian cancer patients from 23 studies spanning 11 gene expression measurement platforms, in the form of documented ExpressionSet objects for R/Bioconductor (7).

Gene expression data were collected from public databases and author websites, processed in a consistent manner and mapped uniformly to official Human Gene Nomenclature Committee (HGNC) (8) gene symbols. Curation of clinical annotations was machine-checked for correctness of syntax and human-checked by two individuals to ensure accuracy. This data package is geared primarily towards bioinformatic and statistical researchers, providing an ideal resource for development and assessment of algorithms for high-dimensional classification, clustering and survival analysis. It will also be valuable to ovarian cancer researchers for biomarker identification and validation. In addition to providing all publicly available gene expression studies with patient survival in common forms of ovarian cancer, it includes tumours of rare histologies, normal tissues and uncommon early-stage tumours. Special effort is made to retain the most important clinical variables from author-provided metadata and from the original publications: overall survival, optimal debulking surgery and tumour stage, grade and histology.

We also developed a software pipeline for automated and reproducible production of this and comparable data libraries. The pipeline includes a controlled language for curation of clinical annotations, defined by a template, which is intuitive for non-programmers to create and edit, but which is also used directly for machine syntax checking of curated annotations. The pipeline handles all steps of the process including data download, microarray preprocessing, merging of duplicate probe sets and sample technical replicates, up-to-date probe-set to gene mapping and building of the R/Bioconductor objects and package.

One important application of the database is testing of hypothesized prognostic markers of ovarian cancer using multiple independent studies. We validated a recently proposed independent prognostic indicator of ovarian cancer, CXCL12 (9), using 13 published studies, demonstrating for this biomarker that numerous studies are needed to overcome the lack of power in individual studies of smaller sample size. We provide code in the documentation of the curatedOvarianData package demonstrating how this comprehensive analysis, which was previously impractical to achieve, is a straightforward application of the database.

Methods and implementation

The pipeline for creating the data package from public databases (Table 1) is fully automated, with the exceptions of manual curation of clinical annotations (Figure 1). This manual curation was integrated in the pipeline with short R scripts that reformat user-provided annotations into a standardized template, which largely follows the format of The Cancer Genome Atlas (29). This template is provided in Table 2 and used as a unit test in the curatedOvarianData package, i.e. the curation is automatically checked for valid values in the package building process. Downloading phenotype data and expression data from GEO (1), syntax validation of curated clinical metadata, microarray data preprocessing, normalization, gene mapping and the creation of Bioconductor ‘ExpressionSet’ objects, which link gene expression data and phenotype annotations, were fully automated. The generation of the package is reproducible using the pipeline provided at https://bitbucket.org/lwaldron/curatedovariandata.

Data acquisition and curation

Our search for clinically annotated ovarian cancer microarray studies identified 21 published studies, which provided 23 publicly available data sets from various sources (Table 1). The search not only targeted studies of primary tumours annotated with patient survival but also included studies providing other potentially valuable clinical annotation. Other main factors of interest included drug resistance, outcome of the primary tumour debulking surgery, histology, stage and grade. We excluded studies not measuring gene expression (i.e. studies of genomic copy number), studies of cell lines, animal models, or non-primary tumours, and data sets not providing clinical information. Expression and clinical data were obtained from the two major public repositories GEO (i) and ArrayExpress (ii), otherwise from supplementary data of the original publications. Data from GEO were obtained using the GEOquery package (31). Clinical annotations...
were manually curated using one R script per data set, and original uncurated annotations were retained as a single field. Curated annotations were checked by syntax against a template, which standardized all the known clinically relevant indicators and allowable data values. Clinical data were twice independently curated (authors B.G. and T.R.), and all discrepancies were resolved for the final version. The availability of clinical data varied substantially across datasets (Figure 2).

**Gene expression processing and gene mapping**

Where raw data from Affymetrix U133a or U133 Plus 2.0 platforms were available, these were pre-processed by frozen Robust Multi-array Analysis (fRMA) (32), for other Affymetrix platforms by Robust Multi-array Average (RMA) (33), and otherwise we used pre-processed data as provided by the authors. Up-to-date maps from probe set IDs to gene symbols were obtained from BioMart (34). Where BioMart maps were not available but target sequences were provided for the microarray platforms, we used the BLAST algorithm (35) to map these sequences against the human genome (build GRCh37) and to identify the gene transcript targeted by each probe. Otherwise, the annotations provided with the platform on GEO were used.

In the curatedOvarianData version of the package, genes with multiple probe sets were represented by the probe set with the highest mean across all data sets of the sample platform (36), and this original probe set identifier was also stored in the ExpressionSet object (7). We selected the same representative probe set for all studies of a common microarray platform. Finally, we provide two alternative versions of the package: NormalizerVcuratedOvarianData, where redundant probe sets are averaged after filtering probe sets with low correlation to their redundant probe sets,
Figure 1. Flowchart of the data collection and curation pipeline. The software implementing this pipeline reproduces all steps from downloading of data to final packaging, requiring manual intervention only for identifying studies, curation of clinical metadata and documentation of the package.
<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Allowed values</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample_type</td>
<td>tumour, metastatic, cellline, healthy, adjacentnormal</td>
<td>Healthy, only from individuals without cancer; adjacentnormal, from individuals with cancer; ser, serous; endo, endometrioid; clearcell, mixture of ser + endo. Other includes sarcomatoid, endometrioid, papillary serous, adenocarcinoma, dysgerminoma.</td>
</tr>
<tr>
<td>histological_type</td>
<td>ser, endo, clearcell, mucinous, other, mix, undifferentiated</td>
<td>ser, serous; endo, endometrioid; clearcell, mixture of ser + endo. Other includes sarcomatoid, endometrioid, papillary serous, adenocarcinoma, dysgerminoma.</td>
</tr>
<tr>
<td>primary_site</td>
<td>ov, ft, other</td>
<td>Ov, ovary; ft, fallopian tube</td>
</tr>
<tr>
<td>arrayed_site</td>
<td>ov, ft, other</td>
<td>Ov, ovary; ft, fallopian tube</td>
</tr>
<tr>
<td>summary_grade*</td>
<td>low, high</td>
<td>low, 1, 2, LMP (low malignant potential); high, 3, 2/3</td>
</tr>
<tr>
<td>summarystage</td>
<td>early, late</td>
<td>early, FIGO I, II, I/II; late, FIGO III, IV, IIII, III/IV</td>
</tr>
<tr>
<td>tumourstage</td>
<td>1, 2, 3, 4</td>
<td>FIGO Stage (I–IV, translated to 1–4 for R usage) Substage (abcd) Grade (1–3)</td>
</tr>
<tr>
<td>substage</td>
<td>a, b, c, d</td>
<td></td>
</tr>
<tr>
<td>grade*</td>
<td>1, 2, 3</td>
<td>Age at initial pathologic diagnosis in years Patient treated with Platin Patient treated with Taxol Neoadjuvant treatment Time to recurrence or last follow-up in days Recurrence censoring variable Time to death or last follow-up in days Overall survival censoring variable Dichotomized overall survival time; as defined by study Dichotomized relapse variable; as defined by the study Site of the first recurrence Response to any kind of therapy Amount of residual disease (optimal ≤ 1 cm) Estimated percentage of normal cells; 20–≤ 20% Estimated percentage of stromal cells Estimated percentage of tumour cells; 80+≥ 80% Hybridization date or other available batch variable All original, uncurated metadata.</td>
</tr>
<tr>
<td>age_at_initial_pathologic_diagnosis</td>
<td>1-99</td>
<td></td>
</tr>
<tr>
<td>pltx</td>
<td>y/n</td>
<td>Patient treated with Platin Patient treated with Taxol Neoadjuvant treatment</td>
</tr>
<tr>
<td>tax</td>
<td>y/n</td>
<td>Patient treated with Taxol Neoadjuvant treatment</td>
</tr>
<tr>
<td>neo</td>
<td>y/n</td>
<td>Neoadjuvant treatment</td>
</tr>
<tr>
<td>days_to_tumour_recurrence</td>
<td>decimal</td>
<td>Time to recurrence or last follow-up in days</td>
</tr>
<tr>
<td>recurrence_status</td>
<td>recurrence, no recurrence</td>
<td>Recurrence censoring variable</td>
</tr>
<tr>
<td>days_to_death</td>
<td>decimal</td>
<td>Time to death or last follow-up in days</td>
</tr>
<tr>
<td>vital_status</td>
<td>living, deceased</td>
<td>Overall survival censoring variable</td>
</tr>
<tr>
<td>os_binary</td>
<td>short, long</td>
<td>Dichotomized overall survival time; as defined by study</td>
</tr>
<tr>
<td>relapse_binary</td>
<td>short, long</td>
<td>Dichotomized relapse variable; as defined by the study</td>
</tr>
<tr>
<td>site_of_tumour_first_recurrence</td>
<td>metastasis, locoregional, etc.</td>
<td>Site of the first recurrence</td>
</tr>
<tr>
<td>primary_therapy_outcome_success</td>
<td>completeresponse, etc.</td>
<td>Response to any kind of therapy</td>
</tr>
<tr>
<td>bebulking</td>
<td>optimal, suboptimal</td>
<td>Amount of residual disease (optimal ≤ 1 cm)</td>
</tr>
<tr>
<td>percent_normal_cells</td>
<td>0–100+/−</td>
<td>Estimated percentage of normal cells; 20–≤ 20%</td>
</tr>
<tr>
<td>percent_stromal_cells</td>
<td>0–100+/−</td>
<td>Estimated percentage of stromal cells</td>
</tr>
<tr>
<td>percent_tumour_cells</td>
<td>0–100+/−</td>
<td>Estimated percentage of tumour cells; 80+≥ 80%</td>
</tr>
<tr>
<td>batch</td>
<td>character</td>
<td>Hybridization date or other available batch variable</td>
</tr>
<tr>
<td>uncurated_author_metadata</td>
<td>character</td>
<td>All original, uncurated metadata</td>
</tr>
</tbody>
</table>

Additional study-specific details are provided in the package manual.

*Most ovarian cancer pathologists follow the FIGO grading system, although some exceptions (15, 22, 25) are noted in the package manual.
using the Normalizer function of the Sleipnir library for computational functional genomics (37), and FULLVcuratedOvarianData, which does not collapse redundant probe sets targeting the same gene transcript but instead provides a probe set to gene symbol map in the featureData slot of each ExpressionSet.

Final packaging

Technical replicate samples were merged by averaging. Microarray expression data and clinical metadata were then represented as ExpressionSet objects (7) for each study. The ExpressionSet objects were also populated with citations, platform identifiers and details, data preprocessing methods and warnings of retracted papers (27) and specimens also used in other studies (26, 28, 29, 38). ExpressionSets were packaged as the curatedOvarianData R library, which provides a reference manual including descriptions of the syntax template and summaries of the annotations, citation, microarray platform and other information for each study.

Discussion

We introduce a data package for the R/Bioconductor statistical programming environment that includes all current major ovarian cancer gene expression data sets (Table 1). The process of downloading clinically annotated public genomic data and proceeding to a final computational analysis is, despite recent efforts (4, 5), still long and prone to errors. This is particularly true when the various data sets need to be comparable for meta-analyses, which requires a fully standardized annotation. Our data resource provides a comprehensive and highly curated resource for efficient meta-analysis of the ovarian cancer transcriptome, for
biological analysis and bioinformatic methods development. It additionally provides a complete computational pipeline to reproduce this process for other cancers or data sources.

Two common problems of publicly available genomic data are the scarcity of clinical annotation and inconsistent definitions of clinical characteristics across independent data sets (5). In our review of original papers and curation of clinical annotations, we were however able to retain, in most studies, the clinical variables of proven importance: overall survival, age, optimal debulking surgery, tumour histology, grade and stage (Figure 2). Other characteristics such as detailed treatment information or recurrence free survival times were rarely available; however, ovarian cancer has a relatively standard treatment regimen of platinum chemotherapy and no radiotherapy. The most important clinical variables were in general consistently defined between studies, with these definitions provided in Table 2. Notably, all studies used the Federation of Gynecology and Obstetrics (FIGO) staging system, and all but one study (11) defined suboptimal debulking surgery as residual tumour mass > 1 cm (Table 2). The relatively large number of well-annotated data sets in this database may allow interesting future work, addressing the problem of recovering missing annotations from genomic data only (40).

One important use of this database is the assessment of prognostic biomarkers. As a demonstration, we examined a recent study by Popple et al. (9), which analysed the expression of the chemokine protein CXCL12 using a tissue microarray of 289 primary ovarian cancers. CXCL12/CXCR4 is a chemokine/chemokine receptor axis that has previously been shown to be directly involved in cancer pathogenesis (41, 42). Ovarian cancer constitutively expresses CXCL12 and CXCR4, and both tumour CXCL12/CXCR4 expression and stroma-derived CXCL12 expression have been reported to be prognostic factors in human ovarian cancer (41). Popple et al. found that high levels of CXCL12 protein were associated with significantly poorer survival compared with patients whose tumours produce low amounts of this chemokine, independently of stage, residual disease (optimal debulking) and adjuvant chemotherapy. The patient cohort was heterogeneous, with various histologic types, grades and stages, leaving open the question of whether this biomarker would be generalizable to other patient populations. Furthermore, differences in protein abundance may not be associated with RNA abundance.

To investigate these questions, we analysed CXCL12 expression in all primary tumour samples included in curatedOvarianData for which overall survival information was available. To ensure that the expression values were on the same scale across studies, all data sets were centred by their means and scaled by their standard deviations. A population hazard ratio (HR) was then pooled with a fixed-effects model, in which the HR for each cohort was weighted with the inverse of the standard error. This is visualized as a forest plot in Figure 3. Although the effect is only significant (P < 0.05) in three cohorts individually, the pooled HR is significantly larger than 1 (HR = 1.15, 95% CI 1.09–1.23). HR refers to the HR between patients differing by one standard deviation in CXCL12 expression. This confirms the hypothesis that upregulation of CXCL12 is associated with poor outcome in 2108 patients from 13 independent studies with mixed stage, grade and histologies. The effect is thus small but consistently detected, emphasizing the importance of biomarker validation in sufficiently large data collections. To assess the independence of CXCL12 with stage and residual disease, we also analysed the 1776 patients from 10 studies where both FIGO tumour stage and success of debulking surgery were known. Adjustment for these two established predictors in multivariate analysis had little effect on the observed association between CXCL12 and overall survival (HR = 1.13, 95% CI 1.05–1.21). These HRs are comparable in magnitude to that reported by Popple et al. for ‘moderate’ CXCL12 staining (HR = 1.215, 95% CI 0.892–1.655), but lower than reported for ‘high’ staining (HR = 1.684, 95% CI 1.180–2.404). This potentially reflects that the function of this gene is at the protein level. Consistent with previous reports (9, 38), we found no significant association of the receptor CXCR4 with overall survival (HR = 0.95, 95% CI 0.9–1.01, P = 0.09). These analyses are straightforward and fully reproduced as examples in the package documentation. Additional analyses limited to more homogeneous patient subsets, e.g., limited to tumours of the same histology, are needed, but they are another straightforward application of the package.

In constructing curatedOvarianData, we took several steps to minimize across-study batch effects. Where raw Affymetrix microarray data were available, we used a standardized pre-processing protocol. All data sets from the same platform were normalized with the same algorithms and parameters. For the Affymetrix U133A and U133 Plus 2.0, we chose the FrMA (32) normalizing algorithm, a variant of the standard RMA (33) algorithm that uses publicly available microarray databases to estimate probe-specific effects and variances, instead of using only the samples from the data set to be normalized. We provide example code in the database documentation for removing between-platform batch effects with the ComBat method (43). Such a batch effect removal is typically necessary when data sets are merged.

If different platforms are compared, then the mapping of probe sets to common identifiers such as gene symbols is a critical and error prone step. In particular when older platforms are considered, care must be taken to ensure that the probe sets target identical transcripts; gene identification is a persistent problem in genome-scale data integration. We used the BioMart database (34) to map
stable manufacturer probe set identifiers or Genbank IDs to current standard gene symbols. For cases in which no stable identifiers were available, we used the BLAST algorithm (35) to identify gene symbols from the probe oligonucleotide sequences. When many genes are targeted by more than one probe set, several approaches of collapsing probe sets to single genes have been proposed (36, 44, 45). In the main version of the package, we selected the probe set with highest mean across all data sets from the same platform to represent each gene transcript, a method shown to perform well (36) and with the advantage of being traceable back to a single oligonucleotide probe sequence for each platform. We also provide two alternative packages with averaged and un-collapsed probe sets. The version with un-collapsed probe sets provides current HGNC symbols in the featureData slot of the ExpressionSet objects, which makes the application of alternative methods for collapsing probe sets to unique gene symbols straightforward, e.g. with the WGCNA R package (46).

We demonstrated meta-analytical use of the package by showing a survival association of the recently proposed prognostic biomarker CXCL12 (9). Other possible uses include the validation of multi-gene signatures, and identification of novel gene signatures and biomarkers for patient survival and response to chemotherapy. Finally, this package enables rigorous assessment of high-dimensional machine-learning algorithms in terms of their performance and computational requirements. We plan to continually include newly published ovarian cancer data sets in future versions of this package.

### Conclusions

The curatedOvarianData package provides a comprehensive resource of curated gene expression and clinical data for the development and validation of ovarian cancer prognostic models, the investigation of ovarian cancer subtypes (10, 25, 29), and the comparative assessment of machine learning algorithms for gene expression data. This database greatly reduces the burden of time, expertise and error involved in assembling a compendium of curated gene expression data from tumours of known histopathology and from patients with known clinical progression. These advantages will be appealing to biostatisticians and bioinformaticians for development of analytical methods from high-dimensional genomic data, but the database will additionally provide a common, version-controlled and transparent platform for reproducible investigation of the ovarian cancer transcriptome. The pipeline for creating this database is published under an open license and will
facilitate creating similar resources for other cancers. As such, we hope this database and pipeline will provide one part of the solution to reproducibility in high-dimensional genomic research.

**Acknowledgements**

The authors thank Shaina Andelman for her contributions to graphic design, and also Steve Skates, Jie Ding and Dave Zhao.

**Funding**

National Cancer Institute at the National Institutes of Health [1RC4CA156551-01 to G.P. and M.B.]; the National Science Foundation [CAREER DBI-1053486 to C.H.]. M.R. acknowledges support from the National Cancer Institute initiative to found Physical Science Oncology Centers [US4CA143798]. Funding for open access charge: National Science Foundation [CAREER DBI-1053486 to C.H.].

**Conflict of interest**

None declared.

**References**


