Human Genomics

SOFTWARE REVIEW Open Access

Survival analysis tools in genomics research

Xintong Chen, Xiaochen Sun and Yujin Hoshida*

Abstract

There is an increasing demand to determine the clinical implication of experimental findings in molecular biomedical research. Survival (or failure time) analysis methodologies have been adapted to the analysis of genomics data to link molecular information with clinical outcomes of interest. Genome-wide molecular profiles have served as sources for discovery of predictive/prognostic biomarkers as well as therapeutic targets in the past decade. In this review, we overview currently available software, web applications, and databases specifically developed for survival analysis in genomics research and discuss issues in assessing clinical utility of molecular features derived from genomic profiling.

Keywords: Survival analysis, Software, Web application, Genomic database

Survival analysis in genomics research

With the increasing capability to perform genome-wide molecular characterization of clinical specimens, making clinical implication of genomic aberrations has become a more relevant topic. The decreasing cost of the assays has facilitated accumulation of genomic profiles of sizable clinical cohorts, with which more reliable molecular prognostic analysis has become possible. Also, expanding clinical contexts covered by the studies/datasets has enabled exploration of clinically more relevant predictive/prognostic biomarkers from genomic data [1]. Here, the major interest is the association of genomic features with clinical outcomes, including response to certain treatment and prognosis of the patients under specific clinical scenarios.

Clinical outcome especially prognosis is often presented as the time period between the start and end of the clinical observation in combination with a binary status information, indicating whether or not each patient had a clinical event of interest, e.g., death, cancer recurrence, and drug response. In contrast to laboratory experiment-derived data, clinical outcome data are generally incomplete because of the missing observation of the clinical event. For example, in the case of analyzing time to cancer recurrence after surgery, some patients who are still recurrence free during the study period may develop recurrence later, i.e., it is uncertain whether the patient should be classified into recurrence-positive or recurrence-negative group. Such situation, where a true outcome is still unknown, is treated as a censored observation, and the observation time is incorporated in the analysis. This type of analysis is called "survival" or "failure time" analysis, for which various biostatistical analysis methodologies are already available. These methodologies have been adapted for the analysis of genomic datasets with modifications to accommodate the high-dimensional data structure by utilizing correction methods for highly multiple hypothesis testing [2].

The accumulated genomic datasets with clinical outcome information have led to a new paradigm of biomarker research, i.e., *in silico* discovery and/or validation of predictive/prognostic molecular biomarkers. In this article, we overview currently available software, web applications, and databases specifically developed for integrative analysis of survival and genomic data. We also discuss current limitations mostly residing on the clinical study design side and how we could methodologically overcome these challenges to facilitate the development of molecular biomarkers with clinical utility.

Tools and resources for survival analysis in genomics research

The major tasks of survival analysis in genomics research include 1) survey/identify genomic feature(s) correlated with survival data and 2) evaluate/validate survival data correlation for predefined genomic feature(s). There are several freely available tools to complete the tasks for users

* Correspondence: yujin.hoshida@mssm.edu
Liver Cancer Program, Tisch Cancer Institute, Division of Liver Diseases, Department of Medicine, Icahn School of Medicine at Mount Sinai, 1470 Madison Avenue, Box 1123, New York, NY 10029, USA

with a wide range of informatics capability and fluency (Table 1). Significance Analysis of Microarrays (SAM) is one of the earliest software to identify genomic feature(s) correlated with biological and/or clinical phenotypes of interest, including time-to-event clinical outcome by using Cox score [3,4]. A similar algorithm is implemented as modules of the GenePattern software, a generic genomic data analysis environment and toolkit [5]. GenePattern LoocvSurvival module enables generation of a robust prognostic gene signature based on leave-one-out cross-validation scheme [6]. Cox regression-based method together with time-dependent receiver operating characteristic (ROC) curve analysis was also reported [7]. Net-Cox is a method based on Cox regression modeling using the information of co-regulated multiple genes, which was reported to improve replication of the prognostic model [8]. survcomp is an R-based Bioconductor [9] package for survival risk model comparison based on time-dependent ROC curve and c index [10].

The ever-expanding repositories of genomic datasets with clinical outcome information have been serving as resources to build web-based tools/resources for survival-related genomic analysis (Table 2). NCBI Gene Expression Omnibus (GEO) [16] and EBI ArrayExpress [17] are generic databases of a variety of genomic datasets with or without clinical outcome information. The Cancer Genome Atlas (TCGA) is a multi-institutional project generating a wide range of genomic data, which are made publicly available together with rich clinical annotations including outcome data [18]. Several survival analysis-focused web applications have also been built based on these resources. Oncomine is an intensively curated genomics database with a special focus on oncology research, providing functionalities of survival-related analysis for datasets with relevant sample annotations [19]. cBioPortal is a web-based resource that enables graphical user interface (GUI)-based intuitive interrogation of a wide range of omics datasets from TCGA and Cancer Cell Line Encyclopedia (CCLE) [20] datasets and, when available, survival data analysis including Kaplan-Meier curve and log-rank test [21]. Similar web-based resources combining genomic/clinical database and analysis tools that enable single/multiple gene-based prognostic assessment include Kaplan-Meier Plotter [22], PrognoScan [23], GOBO [24], Recurrence Online [25], PROGgene [26], bc-GenExMiner [27], ITTACA [28], SurvExpress [29], and G-DOC Plus [30]. These resources assembled publicly or privately available datasets from GEO, ArrayExpress, TCGA, and/or private solicitation/deposition and enable survival analysis based on prefixed or user-defined cutoff for prognostic subgrouping of the patients. Some of them support subgroup analysis and/or multivariable analysis with clinical prognostic variables when available. Some support survival classifier based on multiple genes (or gene signature) using preset algorithms such as averaging or multivariable Cox regression modeling. Breast Cancer Competition (BCC) is a collection of tools to facilitate collaborative genomic classifier building and testing, which was recently used to develop breast cancer prognostic models based on competition between multiple data analysis groups [31]. These tools are readily available to analyze user's own genes or survival models in a variety of diseases, tissue types, and clinical contexts when available.

Toward genome-based biomarkers with real clinical utility

In silico biomarker validation could be a substantially more cost-effective strategy for biomarker development, which typically requires costly and lengthy processes. Despite the exponentially expanding genomic databases and associated survival analysis tools and resources, clinically deployed genome-based biomarkers are still scarce, highlighting the unresolved challenges in biomarker development from genomic studies [43]. One major issue is the clinical study design, which derives the genomic dataset. Predictive/prognostic biomarkers must follow predefined specific study plan to demonstrate their validity and clinical utility. In general, such biomarkers and models should be clearly defined and independently

Table 1 Software for genomic feature-based survival analysis

Software	User interface (programming language)	Functionality	Reference	URL
Significance Analysis of Microarrays (SAM)	Graphical (Excel add-on), command-line (R)	Feature selection	[3,4]	[11]
GenePattern[a]	Graphical	Feature selection, assessment of survival association, model building	[5]	[12,13]
Partial Cox regression analysis	Command-line (R)	Feature selection, assessment of survival association, model building	[7]	[b]
Net-Cox	Command-line (Matlab)	Feature selection, assessment of survival association, model building	[8]	[14]
survcomp	Command-line (R)	Model comparison	[10]	[15]

[a]SurvivalGene, PrognosticGene, and LoocvSurvival modules deposited in [13].
[b]Source code available upon request to the authors.

Table 2 Web applications with database for genomic feature-based survival analysis

Web application/database	Analyzable genetic feature	Covered diseases	Reference	URL
Oncomine	Multiple	Cancer	[19]	[32]
cBioPortal	Multiple	Cancer (37 types)	[21]	[33]
Kaplan-Meier Plotter	Single	Cancer (breast, ovarian, lung)	[22]	[34]
PrognoScan	Single	Cancer (14 types)	[23]	[35]
GOBO	Multiple	Cancer (breast)	[24]	[36]
Recurrence online	Multiple	Cancer (breast)	[25]	[37]
PROGgene	Single/multiple	Cancer (21 types)	[26]	[38]
bc-GenExMiner	Single	Cancer (breast)	[27]	[39]
ITTACA	Single	Cancer (7 types)	[28]	[40]
SurvExpress	Multiple	Cancer (20 types)	[29]	[41]
G-DOC plus	Multiple	Cancer (9 types), non-cancer (3 types)	[30]	[42]

Accessed in October 2014.

evaluated in prospectively enrolled patients. The guidelines for assessment of prognostic marker (REMARK) [44], diagnostic marker (STROBE) [45], and cohort study (STARD) [46] are available to ensure the quality and validity of the biomarkers. However, a vast majority of available genomic datasets rarely meet these requirements because they were generated by using samples of convenience, i.e., biospecimens readily available to the researchers, which were retrospectively collected without predetermined intention of biomarker development or assessment. That is, prognostic genes identified through analysis of the databases may not or less likely to be clinically reliable or reproducible as biomarkers. Quality grading for the study design in the genomic databases such as the one proposed by Simon and colleagues, A (prospective study), B (retrospective analysis of previous prospective study samples), C (prospective/observational), and D (retrospective/observational) [47], will help speculate the reliability of the survival analysis result yielded from each specific dataset. Generation of future genomic data with special attention on these study design-related issues will enable highly reliable computational validation of new biomarkers.

Obviously, the primary goal of this type of exploratory analysis is to determine or speculate clinical outcome association of genomic features. However, if the features selected through the surveillance are further considered as candidates for clinical diagnostic development, there is another issue that needs to be considered. Clinical decision making is generally made according to well-defined, specific clinical contexts that are often summarized in a diagram or flow chart in the clinical practice guidelines. For a molecular biomarker to be considered as a clinical test to support the system of clinical decision making, the marker must demonstrate clinically meaningful utility in terms of magnitude of benefit, feasibility of clinical implementation, and cost in association with the system of existing clinical decision making system/algorithm. It will be technically feasible to incorporate such clinical framework in the aforementioned web-based tools of genomic survival analysis by engaging disease domain experts in their development.

Clinically applicable molecular biomarkers must yield reproducible and robust measurements in real-world clinical setting with clinically acceptable logistical complexity and cost to justify their use. The lack of reproducibility of the measurement especially for transcript-based biomarkers has been the major technical obstacle in clinical deployment of genome-based biomarkers [48]. Recent development of digital biomolecule counting technologies without target amplification has been overcoming this challenge by enabling a more sensitive and robust measurement of a variety of analytes, including DNA, RNA, and protein, as well as chemical modifications of these molecules [49]. Assay technologies that are specifically designed to generate genomic data from real-world clinical specimens, e.g., formalin-fixed paraffin-embedded tissues, will further expand the informatics resources with rich clinical contexts/scenarios and enhance our capability of *in silico* biomarker research. To accommodate requirements from the regulatory agencies for biomarkers such as FDA in the web-based resources may also help facilitate biomarker development. Two additional challenges in bringing genome-based prognostic biomarkers into clinics are reimbursement for the assays from health insurance companies and education of patients and physicians. To make the web-based genomic survival analysis resources accessible to broader communities outside of biomedical research by integrating them with clinical decision support system (CDSS) in electronic health record (EHR) may help resolve these issues and eventually facilitate clinical translation of genome-based prognostic biomarkers.

Abbreviations
FDR: False discovery rate; ROC: Receiver operating characteristic.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
XC, XS, and YH collected the materials, critically reviewed the relevant references, and drafted and proofread the manuscript. YH provided the overall supervision. All authors read and approved the final manuscript.

Acknowledgements
YH is supported by the National Institute of Health (R01 DK099558).

Received: 16 October 2014 Accepted: 11 November 2014
Published online: 25 November 2014

References

1. van't Veer LJ, Bernards R: **Enabling personalized cancer medicine through analysis of gene-expression patterns.** *Nature* 2008, **452**(7187):564–570.
2. Farcomeni A: **A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion.** *Stat Methods Med Res* 2008, **17**(4):347–388.
3. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci U S A* 2001, **98**(9):5116–5121.
4. Bair E, Tibshirani R: **Semi-supervised methods to predict patient survival from gene expression data.** *PLoS Biol* 2004, **2**(4):E108.
5. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP: **GenePattern 2.0.** *Nat Genet* 2006, **38**(5):500–501.
6. Hoshida Y, Villanueva A, Kobayashi M, Peix J, Chiang DY, Camargo A, Gupta S, Moore J, Wrobel MJ, Lerner J, Reich M, Chan JA, Glickman JN, Ikeda K, Hashimoto M, Watanabe G, Daidone MG, Roayaie S, Schwartz M, Thung S, Salvesen HB, Gabriel S, Mazzaferro V, Bruix J, Friedman SL, Kumada H, Llovet JM, Golub TR: **Gene expression in fixed tissues and outcome in hepatocellular carcinoma.** *N Engl J Med* 2008, **359**(19):1995–2004.
7. Li H, Gui J: **Partial Cox regression analysis for high-dimensional microarray gene expression data.** *Bioinformatics* 2004, **20**(Suppl 1):i208–i215.
8. Zhang W, Ota T, Shridhar V, Chien J, Wu B, Kuang R: **Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment.** *PLoS Comput Biol* 2013, **9**(3):e1002975.
9. **Bioconductor - open source software for bioinformatics.** In [http://www.bioconductor.org/]
10. Schroder MS, Culhane AC, Quackenbush J, Haibe-Kains B: **survcomp: an R/Bioconductor package for performance assessment and comparison of survival models.** *Bioinformatics* 2011, **27**(22):3206–3208.
11. **Significance Analysis of Microarrays - supervised learning software for genomic expression data mining.** In [http://statweb.stanford.edu/~tibs/SAM/]
12. **GenePattern - a powerful genomic analysis platform.** In [http://www.broadinstitute.org/cancer/software/genepattern/]
13. **GParc - GenePattern module repository.** In [http://gparc.org/]
14. **Net-Cox - network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment.** In [http://compbio.cs.umn.edu/Net-Cox/]
15. **survcomp - performance assessment and comparison for survival analysis.** In [http://www.bioconductor.org/packages/release/bioc/html/survcomp.html]
16. **Gene Expression Omnibus - a public functional genomics data repository.** In [http://www.ncbi.nlm.nih.gov/geo/].
17. **ArrayExpress - a database of functional genomics experiments.** In [http://www.ebi.ac.uk/arrayexpress/]
18. *The Cancer Genome Atlas (TCGA) Data Portal*; [https://tcga-data.nci.nih.gov/tcga/]
19. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, Barrette TR, Anstet MJ, Kincead-Beal C, Kulkarni P, Varambally S, Ghosh D, Chinnaiyan AM: **Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles.** *Neoplasia* 2007, **9**(2):166–180.
20. *Cancer Cell Line Encyclopedia (CCLE)*; [http://www.broadinstitute.org/ccle/home]
21. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N: **Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal.** *Sci Signal* 2013, **6**(269):l1.
22. Gyorffy B, Lanczky A, Szallasi Z: **Implementing an online tool for genome-wide validation of survival-associated biomarkers in ovarian-cancer using microarray data from 1287 patients.** *Endocr Relat Cancer* 2012, **19**(2):197–208.
23. Mizuno H, Kitada K, Nakai K, Sarai A: **PrognoScan: a new database for meta-analysis of the prognostic value of genes.** *BMC Med Genomics* 2009, **2:**18.
24. Ringner M, Fredlund E, Hakkinen J, Borg A, Staaf J: **GOBO: gene expression-based outcome for breast cancer online.** *PLoS One* 2011, **6**(3):e17911.
25. Gyorffy B, Benke Z, Lanczky A, Balazs B, Szallasi Z, Timar J, Schafer R: **RecurrenceOnline: an online analysis tool to determine breast cancer recurrence and hormone receptor status using microarray data.** *Breast Cancer Res Treat* 2012, **132**(3):1025–1034.
26. Goswami CP, Nakshatri H: **PROGgene: gene expression based survival analysis web application for multiple cancers.** *J Clin Bioinformatics* 2013, **3**(1):22.
27. Jezequel P, Campone M, Gouraud W, Guerin-Charbonnel C, Leux C, Ricolleau G, Campion L: **bc-GenExMiner: an easy-to-use online platform for gene prognostic analyses in breast cancer.** *Breast Cancer Res Treat* 2012, **131**(3):765–775.
28. Elfilali A, Lair S, Verbeke C, La Rosa P, Radvanyi F, Barillot E: **ITTACA: a new database for integrated tumor transcriptome array and clinical data analysis.** *Nucleic Acids Res* 2006, **34**(Database issue):D613–D616.
29. Aguirre-Gamboa R, Gomez-Rueda H, Martinez-Ledesma E, Martinez-Torteya A, Chacolla-Huaringa R, Rodriguez-Barrientos A, Tamez-Pena JG, Trevino V: **SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis.** *PLoS One* 2013, **8**(9):e74250.
30. Madhavan S, Gusev Y, Harris M, Tanenbaum DM, Gauba R, Bhuvaneshwar K, Shinohara A, Rosso K, Carabet LA, Song L, Riggins RB, Dakshanamurthy S, Wang Y, Byers SW, Clarke R, Weiner LM: **G-DOC: a systems medicine platform for personalized oncology.** *Neoplasia* 2011, **13**(9):771–783.
31. Bilal E, Dutkowski J, Guinney J, Jang IS, Logsdon BA, Pandey G, Sauerwine BA, Shimoni Y, Moen Vollan HK, Mecham BH, Rueda OM, Tost J, Curtis C, Alvarez MJ, Kristensen VN, Aparicio S, Borresen-Dale AL, Caldas C, Califano A, Friend SH, Ideker T, Schadt EE, Stolovitzky GA, Margolin AA: **Improving breast cancer survival analysis through competition-based multidimensional modeling.** *PLoS Comput Biol* 2013, **9**(5):e1003047.
32. *Oncomine*; [https://www.oncomine.org/resource/login.html]
33. *cBioPortal for Cancer Genomics*; [http://www.cbioportal.org/public-portal/]
34. **Kaplan-Meier Plotter - cancer survival analysis.** In [http://kmplot.com/analysis/]
35. **PrognoScan - a new database for meta-analysis of the prognostic value of genes.** In [http://www.abren.net/PrognoScan/]
36. *GOBO - Gene Expression-Based Outcome for Breast Cancer Online*; [http://co.bmc.lu.se/gobo/]
37. **Recurrence Online - transcriptome based breast cancer diagnostics.** In [http://www.recurrenceonline.com/]
38. *PROGgene - Pan Cancer Prognostics Database*; [http://watson.compbio.iupui.edu/chirayu/proggene/database/?url=proggene]
39. **bc-GenExMiner - platform for gene prognostic analyses in breast cancer.** In [http://bcgenex.centregauducheau.fr/BC-GEM/GEM_Accueil.php?js=1]
40. *ITTACA - Integrated Tumor Transcriptome Array and Clinical Data Analysis*; [http://bioinfo-out.curie.fr/ittaca/]
41. **SurvExpress - biomarker validation for cancer gene expression.** In [http://bioinformatica.mty.itesm.mx:8080/Biomatec/SurvivaX.jsp]
42. *G-DOC Plus - Georgetown Database of Cancer Plus other diseases*; [https://gdoc.georgetown.edu/gdoc/]
43. Hoshida Y, Moeini A, Alsinet C, Kojima K, Villanueva A: **Gene signatures in the management of hepatocellular carcinoma.** *Semin Oncol* 2012, **39:**473–485.
44. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM: **Reporting recommendations for tumor marker prognostic studies (REMARK).** *J Natl Cancer Inst* 2005, **97**(16):1180–1184.
45. Vandenbroucke JP, von Elm E, Altman DG, Gotzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M, Initiative S: **Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration.** *PLoS Med* 2007, **4**(10):e297.
46. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Moher D, Rennie D, de Vet HC, Lijmer JG, Standards for Reporting of Diagnostic Accuracy: **The STARD statement for reporting studies of**

diagnostic accuracy: explanation and elaboration. *Ann Intern Med* 2003, **138**(1):W1–W12.

47. Simon RM, Paik S, Hayes DF: **Use of archived specimens in evaluation of prognostic and predictive biomarkers.** *J Natl Cancer Inst* 2009, **101**(21):1446–1452.
48. Koscielny S: **Why most gene expression signatures of tumors have not been useful in the clinic.** *Sci Transl Med* 2010, **2**(14):14ps12.
49. Geiss GK, Bumgarner RE, Birditt B, Dahl T, Dowidar N, Dunaway DL, Fell HP, Ferree S, George RD, Grogan T, James JJ, Maysuria M, Mitton JD, Oliveri P, Osborn JL, Peng T, Ratcliffe AL, Webster PJ, Davidson EH, Hood L, Dimitrov K: **Direct multiplexed measurement of gene expression with color-coded probe pairs.** *Nat Biotechnol* 2008, **26**(3):317–325.

doi:10.1186/s40246-014-0021-z
Cite this article as: Chen *et al.*: Survival analysis tools in genomics research. *Human Genomics* 2014 **8**:21.