Genomics xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Integrated analysis of multidimensional omics data on cutaneous melanoma prognosis

Yu Jiang [a,b,1], Xingjie Shi [c,1], Qing Zhao [d], Michael Krauthammer [e], Bonnie E. Gould Rothberg [f], Shuangge Ma [b,g,*]

[a] Division of Epidemiology, Biostatistics, and Environmental Health, School of Public Health, University of Memphis, Memphis, TN 38152, USA
[b] VA Cooperative Studies Program Coordinating Center, West Haven, CT 06516, USA
[c] Department of Statistics, Nanjing University of Finance and Economics, Nanjing, China
[d] Merck Research Laboratories, 126 East Lincoln Avenue, RY34, Rahway, NJ 07065, USA
[e] Department of Pathology, Yale University, New Haven, CT 06520, USA
[f] Cancer Center, Department of Internal Medicine, Pathology, Chronic Disease Epidemiology, Yale University, New Haven, CT 06520, USA
[g] Department of Biostatistics, Yale University, New Haven, CT 06520, USA

ARTICLE INFO

Article history:
Received 18 December 2015
Received in revised form 5 April 2016
Accepted 23 April 2016
Available online xxxx

Keywords:
Multidimensional omics data
Melanoma prognosis
Integration
The Cancer Genome Atlas (TCGA)

ABSTRACT

Multiple types of genetic, epigenetic, and genomic changes have been implicated in cutaneous melanoma prognosis. Many of the existing studies are limited in analyzing a single type of omics measurement and cannot comprehensively describe the biological processes underlying prognosis. As a result, the obtained prognostic models may be less satisfactory, and the identified prognostic markers may be less informative. The recently collected TCGA (The Cancer Genome Atlas) data have a high quality and comprehensive omics measurements, making it possible to more comprehensively and more accurately model prognosis. In this study, we first describe the statistical approaches that can integrate multiple types of omics measurements with the assistance of variable selection and dimension reduction techniques. Data analysis suggests that, for cutaneous melanoma, integrating multiple types of measurements leads to prognostic models with an improved prediction performance. Informative individual markers and pathways are identified, which can provide valuable insights into melanoma prognosis.

1. Introduction

Cutaneous melanoma poses a major public health concern. In 2015, an estimated 73,870 new cases of invasive melanoma are expected in the U.S., with an estimated 9940 deaths [33]. Cutaneous melanoma is the largest subtype, and Caucasians have a much higher risk and poorer prognosis. Despite extensive research, the understanding of melanoma prognosis is still very limited. Clinicopathologic features that have been suggested as prognostic include age at diagnosis, gender, Breslow tumor thickness, ulceration status, mitotic index, and presence of lymph node micrometastases [1,14]. Significant effort has been devoted to searching for omics markers that may contribute to melanoma prognosis independent of the aforementioned factors. Several multi-marker prognostic models have been published. Omics markers identified in the literature belong to the immunomodulation, DNA repair, signal transduction, melanoma endophenotypes, and other pathways.

Identifying prognostic omics markers has important implications. For basic scientists, it leads to a better understanding of the biological mechanisms underlying prognosis. For translational researchers and physicians, it assists patient stratification, treatment selection, and prediction of prognosis paths.

In the literature, multiple types of omics changes have been suggested as potentially associated with melanoma prognosis. For mRNA expression, Winnepenninckx et al. [41] identified 254 genes associated with distant metastasis-free survival. Gene expression studies also include Timar et al. [37], Gerami et al. [17], and others. Studies of tumor cells in melanoma patients have characterized prognostic alterations with a panel of five genes in copy number alteration (CNA; [10]). MicroRNA has also been implicated in melanoma prognosis. For example, the study by Streicher et al. [35] identified a fourteen-microRNA cluster on the X chromosome, the miRNA-506–514 cluster, and found that this cluster is critical in cancer cell growth and melanocyte transformation. DNA methylation profile has been investigated. Notable studies include Conway et al. [12] and a review study by Schinke et al. [31]. Sigalotti and others analyzed methylation data and constructed a seventeen-gene signature. For genetic mutations, the associations of several somatic variants – such as BRAF V600E and NRAS Q61R/L/H – with prognosis have been reported [3,43]. A whole-genome sequencing study found the RAC1 mutation as the third most frequent in sun-exposed melanomas and suggested its potential role in prognosis [23].

A common limitation shared by many of the existing studies, especially the early ones, is that they are "one-dimensional" in the sense

[*] Corresponding author at: 60 college ST, LEPH 206, New Haven, CT 06520, USA.
E-mail address: shuangge.ma@yale.edu (S. Ma).
[1] The two authors contributed equally to this work.

http://dx.doi.org/10.1016/j.ygeno.2016.04.005
0888-7543/© 2016 Elsevier Inc. All rights reserved.

that they profiled and analyzed only a single type of omics measurement. Multiple types of omics measurements are interconnected and have possibly overlapping but also independent information. For example, CNAs, microRNAs, methylation, and other changes affect gene expressions, which affect cancer outcomes/phenotypes through proteins. On the other hand, they can also directly affect protein expressions and functionalities through channels other than gene expressions. That is, they contain independent information on cancer outcomes not reflected in gene expressions. Analyzing a single type of omics measurement cannot comprehensively and accurately describe the biological processes underlying prognosis and may lead to suboptimal prognostic models and uninformative marker identification [44].

More recently, much effort has been devoted to multidimensional studies which profile multiple types of omics changes on the same subjects. A representative example is TCGA (The Cancer Genome Atlas) which is organized by NIH. For multiple cancer types such as breast cancer, ovarian cancer, and glioblastoma, the integrated analysis of TCGA data has been conducted. More accurate prognostic models have been constructed, and important markers missed by the existing studies have been identified [5–7]. For cutaneous melanoma, the TCGA data were very recently published, making it possible to conduct integrated analysis and more accurately describe its prognosis.

For several cancer types, multiple approaches have been applied to conduct the integrated analysis of multidimensional data. Some of the existing studies focus on the regulations among multiple types of omics measurements. Of special interest is the regulation of mRNA gene expression by miRNA, CNA, methylation, and other mechanisms [15,40], as gene expression is the downstream product and can be more directly related to clinical outcomes and phenotypes. Different from these studies, the present one is more concerned with linking omics measurements with prognosis, which is of more practical interest. Some other studies have analyzed each type of omics measurement separately and then compare results across multiple types of measurements. This is basically a meta-analysis strategy and suitable for identifying “hot zones” that host multiple omics changes. However as prognosis is affected by the joint effects of multiple types of omics changes, such an approach may not be effective in building prognostic models.

Overall, this study may complement the existing literature and be warranted in the following aspects. First, it provides a timely integrated analysis of the TCGA cutaneous melanoma data, and the results may provide insights into this clinically important disease. Second, it describes in detail how to conduct effective integrated analysis of multiple types of omics data using advanced statistical techniques and proper statistical packages, which are potentially applicable to many other datasets and diseases.

2. Methods

2.1. TCGA cutaneous melanoma data

TCGA is one of the largest and most comprehensive multidimensional cancer studies. For cutaneous melanoma, the goal was to collect data on about 500 samples. The protocols of TCGA sample and data collection have been described in detail elsewhere [8]. Data analyzed in this study were downloaded either directly from the TCGA website or from cbioportal using the CGDS-R package. Brief data information is provided in Table 1, and the flowchart of data processing is provided in the top part of Fig. 1.

For clinical and pathological variables, the preprocessed level 3 data were downloaded. The number of samples with available data is 422. In the analysis, only white metastatic samples are included. Data on the normal samples are excluded, and multiple data records on the same samples are merged. Only variables with missing rates below 40% are considered. Among them, those that have been suggested as potentially associated with melanoma prognosis include: gender, age at diagnosis, tumor status, Breslow thickness at diagnosis, Clark level at diagnosis, primary melanoma tumor ulceration, AJCC tumor pathologic stage, AJCC nodes pathologic stage, new tumor event, percent of lymphocyte infiltration, percent of monocyte infiltration, percent of necrosis, percent of stromal cells, percent of tumor cells, and percent tumor nuclei. The following variable recoding is conducted to facilitate analysis (by reducing cells with very small counts). The AJCC tumor pathologic stage is coded as 0 for T0 and Ts, 1 for T1–T3, and 2 for T4. The AJCC nodes pathologic stage is coded as 0 for N0 and Nx, 1 for N1, 2 for N2, and 3 for N3. After processing, data are available for 16 variables and 317 samples. To accommodate the remaining missing measurements, multiple imputation is conducted using the package *Amelia* [19].

Omics data were downloaded from cbioportal using the CGDS-R package. Mutation data are available on 278 samples. Following a recent study [21], mutation data on NRAS and BRAF are included in analysis. For a sample, the mutation status is coded as 1 if there is at least one mutation in the specific gene, and as 0 otherwise. In addition, attempt has been made to incorporate all mutation data in analysis. It is found that, with the extremely high dimensionality and noisy nature of mutation data, including all mutations leads to inferior prediction performance (details omitted). Thus, only the two most important mutations are analyzed. CNA measurements were obtained using the Affymetrix Genome-wide Human SNP array 6.0 platform. The loss and gain levels of copy number changes of tumors compared to normal tissues were identified using segmentation analysis and expressed in the log2 transformed form. A total of 21,699 measurements are available on 366 samples. DNA methylation at CpG sites was measured using the Illumina Human Methylation 450 platform. The available data contain the beta values, which represent the percentages of methylation, for 15,589 genes and 373 samples. The range of the beta values is from 0 (fully unmethylated) to 1 (fully methylated). mRNA gene expressions were measured using the Illumina Hiseq RNAseq V2 platform. The downloaded data are the robust Z-scores which have been lowess-normalized, log-transformed, and median-centered and represent the gene expression status (up or down regulated) in tumor samples relative to normal tissues. A total of 19,626 measurements are available on 371 samples.

Besides the aforementioned omics measurements, TCGA also has miRNA data, which, however, are not available from cbioportal. The miRNA data are not analyzed in this study with the concern on data source consistency. In addition, both the TCGA website and cbioportal have protein data. However measurements are only available on 129 protein expressions and 204 samples. A closer examination suggests

Table 1
Brief data information, before and after processing.

	Platform/method	Number of samples	Number of features before processing	Number of features after processing
Clinical-pathological	N.A.	317	83	16
Mutation	Mutation calling	278	15,861	2
CNA	Affymetrix Genome-wide Human SNP array 6.0	336	21,699	2500
Methylation	Illumina Human Methylation 450	373	15,589	2500
Gene expression	Illumina Hiseq RNAseq V2	371	19,626	2500