Driver gene mutations based clustering of tumors: methods and applications

Wensheng Zhang¹,*, Erik K. Flemington² and Kun Zhang¹,*

¹Department of Computer Science, Bioinformatics Facility of Xavier NIH RCMI Cancer Research Center, Xavier University of Louisiana, New Orleans, LA 70125, USA and ²Department of Pathology, Tulane School of Medicine, Tulane Cancer Center, Tulane University, New Orleans, LA 70112, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Somatic mutations in proto-oncogenes and tumor suppressor genes constitute a major category of causal genetic abnormalities in tumor cells. The mutation spectra of thousands of tumors have been generated by The Cancer Genome Atlas (TCGA) and other whole genome (exome) sequencing projects. A promising approach to utilizing these resources for precision medicine is to identify genetic similarity-based sub-types within a cancer type and relate the pinpointed sub-types to the clinical outcomes and pathologic characteristics of patients.

Results: We propose two novel methods, ccpwModel and xGeneModel, for mutation-based clustering of tumors. In the former, binary variables indicating the status of cancer driver genes in tumors and the genes’ involvement in the core cancer pathways are treated as the features in the clustering process. In the latter, the functional similarities of putative cancer driver genes and their confidence scores as the ‘true’ driver genes are integrated with the mutation spectra to calculate the genetic distances between tumors. We apply both methods to the TCGA data of 16 cancer types. Promising results are obtained when these methods are compared to state-of-the-art approaches as to the associations between the determined tumor clusters and patient race (or survival time). We further extend the analysis to detect mutation-characterized transcriptomic prognostic signatures, which are directly relevant to the etiology of carcinogenesis.

Availability and implementation: R codes and example data for ccpwModel and xGeneModel can be obtained from http://webusers.xula.edu/kzhang/ISMB2018/ccpw_xGene_software.zip.

Contact: wzhang@xula.edu or kzhang@xula.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Precision cancer medicine relies on the precise understanding of the biological and genetic characteristics of individual patient’s tumor(s). Somatic mutations in proto-oncogenes and tumor suppressor genes constitute a major category of causal genetic abnormalities in tumor cells. The mutation spectra of thousands of tumor samples have been generated by The Cancer Genome Atlas (TCGA) and other whole genome (exome) sequencing projects. A promising approach for utilizing these sources of information is to establish genetic similarity-based clusters (or sub-types) within a cancer type and then relate the sub-types to clinical outcomes and pathologic characteristics of patients.

Mutation-based clustering analysis is still in its infancy, as the following issues have not been sufficiently addressed. First, the mutated genes in the tumor samples of a patient cohort are usually numerous but the mutation events present on a single gene are generally sparse, thus conventional clustering algorithms cannot be directly applied to such type of data. Second, unlike the case of gene expression profiling, the functional similarity between genes cannot be reflected in the mutation profile whose element usually takes a binary value (i.e. 0 or 1). However, such information is crucial in calculating the between-sample similarity and ignorance of this information could disassociate two tumors that have different mutation profiles (on driver genes) but share a common or similar mechanism of tumorigenesis. Third, the relevance of a non-synonymous mutation to the formation and progression of tumors intuitively depends on the confidence in its host gene as a true cancer gene. If the confidence variability across the genes is not considered, the contribution of a putative cancer gene to the similarity of two tumors could be under- or over- estimated in the computation. Lastly, in order to make the statistical results interpretable and meaningful for precision medicine, the clustering process should be

© The Author(s) 2018. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

i404
Driver gene mutations based clustering of tumors

2 Methods and data

2.1 Data

2.1.1 TCGA data

Among the 33 cancer types with clinically-annotated multi-omic data (http://cancergenome.nih.gov/), 16 are studied in this work by considering the genetic diversity of patients. Each of the selected cancer types has at least 14 patients from a minority population (i.e. black American or Asian) besides the dominant white Americans. The studied cancer types include bladder urothelial carcinoma (BLCA), glioblastoma multi-forme (GBM), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), breast invasive carcinoma (BRCA), ovarian serous cystadenocarcinoma (OV), uterine corpus endometrial carcinoma (UCEC), colon adenocarcinoma (COAD), thyroid cancer (THCA), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), esophageal carcinoma (ESCA), cervical kidney renal papillary cell carcinoma (KIRP), liver hepatocellular carcinoma (LIHC) and stomach adenocarcinoma (STAD). These cancer types have 168 to 967 patient samples whose somatic mutation profiles are available (Supplementary Table S1). Additional information of the used somatic mutation data is summarized in Supplementary Table S2 and Supplementary Text S1. Synonymous mutations and those under the categories of ‘intron’ and ‘rna’ are excluded from further analysis.

We also download the Level-3 RNA-Seq gene expression profiling (RNAsSeqV2 data) of BLCA and LIHC to show how the results from the clustering analysis of the mutation spectra could be used to identify a mutation-based transcriptomic prognostic signature. Logarithm transformation and between-tumors normalization of the expression data using cyclic loess are performed before the advanced analysis.

2.1.2 Gene expression data for validation

From the GEO database (accessed on June 28, 2016), we obtain four microarray gene expression datasets to validate the prognostic signatures pinpointed using the TCGA data. They include two datasets for bladder cancer, i.e. GSE31684 (N = 93; Riemser et al., 2012) and GSE13507 (N = 165; Kim et al., 2010), and two datasets for LIHCCs, i.e. GSE14520 (N = 225; Roessler et al., 2012) and GSE54236 (N = 88; Villa et al., 2016). These datasets have been quantile normalized and logarithm transformed, thus no further pre-processing is performed.

2.1.3 Catalogue of cancer driver genes and the confidence scores

The driver gene catalogue was generated by (Tamborero et al., 2013a) according to their comprehensive analysis of 12 TCGA cancer types. The original collection contains 435 potential cancer driver genes. Hereafter, we call these genes the putative cancer driver genes. Among them, 13 have not been annotated to any highly-specific GO terms (http://geneontology.org/) by June 10, 2016 (Section 2.3), and thus are excluded from further study. As to the confidence in a gene being a true cancer driver gene, we do not use the partition given by Tamborero et al., who grouped the genes into two categories (i.e. ‘high-confidence’ and ‘candidate’). Instead, a confidence score (μ) is calculated by the formula, i.e. \( \frac{c + \sum_{i=1}^{4} d_i}{5} \), where c and d are binary variables (0/1). c indicates if a gene is included as a census cancer gene in the COSMIC database (http://cancer.sanger.ac.uk/cosmic) and d indicates if the gene is predicted by the 4th computational tools as a cancer driver gene. We consider five tools in this study and they are MuSiC (Dees et al., 2012), OncodriveFM (Gonzalez-Perez and Lopez-Bigas, 2012), OncodriveCLUST (Tamborero et al., 2013b), ActiveDriver (Reimand and Bader, 2013) and MutSig (Lawrence et al., 2013). Because both MutSig and MuSiC measure mutation recurrences, we combine their results into a single 0/1 variable in estimating the confidence scores of a driver gene. This variable takes the value 1 if a gene is identified as a tumor driver by any one of these two tools. Otherwise, its value is zero.

2.2 Overview of the clustering methods

We propose two novel methods (i.e. ccpwModel and xGeneModel) to cluster tumors based on the somatic mutation spectra of the putative cancer driver genes. In the applications, we compare them with a naive method (2geneModel) and the network-based stratification model (nbsModel). In the following, we outline the major points of the 2GeneModel, nbsModel and ccpwModel and present xGeneModel in Section 2.3.

2.2.1 GeneModel

In this method, tumor samples within a cohort are partitioned into four groups according to the genotypes (i.e. wild and mutant) of the two most frequently mutated cancer driver genes for the specific cancer type. This method is a naive extension of the single crucial driver gene (such as TP53) based stratification of tumors, which has
been widely studied in previous work (Robles and Harris, 2010; Zhang et al., 2016a).

2.2.2 nbsModel
The inputs of this method are the mutation spectra of a cancer cohort and the protein interaction network of the putative cancer driver genes. The network is retrieved from the STRING database (Szklarczyk et al., 2011). The distance between tumor samples is calculated by the method in (Hofree et al., 2013) and the nbs v0.2 software. The final partition of tumors is obtained by the hclust() function in the R package 'stat' with ward.2D as the argument.

2.2.3 ccpwModel
In this method, binary variables indicating the status (i.e. presence or absence) of cancer driver genes in tumor samples and the involvement of those genes in a dozen core cancer pathways (CCPWs) are considered as features in the Ward’s hierarchical clustering. The CCPW catalogue and the relevance to ~200 driver genes are retrieved from (Vogelstein et al., 2013).

2.3 xGeneModel
In this method, the functional similarities of the putative cancer driver genes and their confidence scores as the ‘true’ driver genes are integrated with the mutation events to calculate the genetic distance between tumors. Like 2GeneModel, the clustering process is transparent since the distance of two tumors is calculated from the genotypes of a few pairs of cancer genes in an explicit way. The flowchart of xGeneModel is presented in Figure 1. The inputs of this method include three relationship matrices (M1, M2 and M3) and a weight vector (W).

- M1 is a zero-one filled matrix, indicating the involvement of a driver gene in the biological process (BP) or molecular function (MF) by GO terms. It is split into two blocks. The left block represents the pre-selected BP terms and the right block represents the pre-selected MF terms. A selected BP (or MF) term is at least five layers away from the root term GO: 0008150 (GO: 0003674) in the structure graph.
- M2 is a block diagonal matrix. The left-top (or right-bottom) block represents the similarity measures of the BP terms (or MF terms) with the values ranging from 0 to 1. Those values are calculated by the mgoSim() function in the R package ‘GOSemSim’ (Yu et al., 2010).
- M3 is a zero-one filled matrix, indicating the mutation status of the driver genes in the tumor samples of a cancer patient cohort.
- W is a vector containing the confidence scores of the putative cancer driver genes being the ‘true’ cancer driver genes (Section 2.1).

Given the pre-prepared input matrices and vector, xGeneModel contains the following four steps.

i. **Step-1** Based on M1 and M2, the similarity score (correlation) between two genes (e.g. g<sub>i</sub> and g<sub>j</sub>) is calculated using the formula \( s = \frac{b + m}{2} \), where \( b \) represents the average correlation between the BP terms of \( g_i \) and BP terms of \( g_j \). \( m \) is similarly defined for MF terms. In this way, we generate the functional similarity matrix (M4) for all the putative cancer driver genes.

ii. **Step-2** M4 is adjusted to obtain a weighted between-gene similarity matrix (M5) by the Equation \( M5 = \text{diag}(W) \times M4 \times \text{diag}(W) \). The \( ith \) row \( jth \) column element of M5, \( m_{ij}^{(5)} \), represents the ‘strength’ of the genes \( g_i \) and \( g_j \) for connecting two tumor samples that have a mutation on either of them, respectively.

iii. **Step-3** Based on M3 and M5, the mutation similarity matrix (M6) is calculated element by element for the tumors in a cohort. Given a pair of tumors \( t_p \) and \( t_q \) and their mutation profile vectors \( m_p^{(3)} \) and \( m_q^{(3)} \) (i.e. the \( pth \) and \( qth \) rows of M3), M5 is firstly diluted by assigning zeros to the entries at the non-informative rows (R) and columns (C) to generate a ‘kernel’ matrix \( K \). \( R \) includes the rows corresponding to the zero elements of \( m_p^{(3)} \), indicating the unmutated genes in tumor \( p \). \( C \) includes the columns corresponding to the zero elements of \( m_q^{(3)} \), indicating the unmutated genes in tumor \( q \). Then, \( n \) determining elements \( (a_1, a_2, ..., a_n) \) are selected sequentially from \( K \), where \( n \) is the previously specified number of gene pairs to be considered. Specifically, \( a_1 \) is the maximum entry of \( K, a_2 \) is the maximum entry of the matrix \( K^{(-1)} \) generated by removing the column and row of \( a_1 \) from \( K \), \( a_3 \) is the maximum entry of the matrix \( K^{(-2)} \) generated by removing the column and row of \( a_2 \) from \( K^{(-1)} \) and so on. Finally, \( m_{pq}^{(6)} \), the mutation similarity score of the \( pth \) and \( qth \) tumors is quantified by the average of \( a_1, a_2, ..., a_n \).

iv. **Step-4** Ward’s hierarchical clustering analysis is performed with \( J-M6 \) as the input distance matrix of tumors, where \( J \) is an all-ones matrix of the same size as \( M6 \).

2.4 Association analysis
We perform the survival analysis using the R package ‘survival’. Except for where specifically stated in the Results Section, \( P \)-values for the association between tumor clusters and the overall survival months of patients is obtained by the logRank test and Cox proportional hazard (Cox-PH) regression model, in which patient age is included as a covariate to be corrected. The Kaplan–Meier survival curves is created by the ‘survfit()’ function, with the censored observations being marked by a vertical tick. The association between tumor clusters and patient racial groups is evaluated by the Fisher’s exact test.

2.5 Transcriptomic prognostic signatures
Within a cancer cohort, the mutation-based clusters are combined into two survival-characterized cluster aggregates according to the results of the association analysis. Prognostic marker genes (PM-genes) are individually selected by a t-test based on the difference of the expression levels between these two aggregates. The prediction strength of a transcriptomic signature (i.e. the expression profiles of the PM-genes) for patient survival is tested by a clustering-based method and a Singular Value Decomposition (SVD) based procedure similar to that used in (Zhang et al., 2016b). In the former, tumor samples are firstly stratified into two or three groups based on their expression matrix of the PM-genes and then the survival curves of these groups are compared. In the latter, a macro measure quantified with the leading SVD left vector of the expression matrix of the PM-genes is considered as the interested predictive variable in the Cox-PH regression model.

3 Results
3.1 Mutation-based clusters
In Sections 2.2 and 2.3, we present the major points of four mutation-based clustering methods. Several implementation details are further clarified here. The first is how many clusters the tumor samples within a cohort should be partitioned into. Theoretically, personalized medicine prefers a partition in which the number \( k \) of clusters is relatively large. However, a previous study showed that,
for many cancer types, setting $k$ to be 3 or 4 could lead to the more clinically meaningful results (Hofree et al., 2013). In this regard, we consistently set $k$ to be 4 to facilitate the comparison of these four methods. The second is the number ($N_{gp}$) of gene pairs that should be considered in computing the genetic similarity of two tumor samples, which is a specific parameter for xGeneModel. While some tumors in the TCGA data have over 10 mutations in the putative cancer genes, a recent study showed that most cancer types only require 3 driver mutations (Tomasetti et al., 2015). As such, we run the model three times for each cancer type, which corresponds to setting $N_{gp}$ at 2, 3, or 4. That is, in the three scenarios, the information of 2–4, 3–6 or 4–8 genes is used to evaluate the similarity of a specific tumor pair, respectively. The third is how to evaluate the association between the cluster and patient survival. Kim et al. computed a $P$-value by comparing the patient (tumor) cluster with the best survival to the cluster with the worst survival (Kim et al., 2015). Given the limited sample sizes and the death events within a cohort in the TCGA data, this approach may exaggerate the association. As a result, in this work, we determine the significance of an association by comparing an individual cluster with the others or comparing the aggregates of two clusters with the others.

A comprehensive evaluation of these clustering methods is given in Table 1. The details are graphically demonstrated by a series of figures in Supplementary Figure Sets S1, S2, S3 and S4, each of which corresponds to the results of one method. In particular, xGeneModel performs slightly better (in terms of the association between the obtained partitions of tumor samples and patient survival) when $N_{gp}$ is set at 2 rather than 3 or 4. Therefore, only the results for $N_{gp} = 2$ are reported in this study.

Regarding the association between the determined tumor clusters and patient survival, xGeneModel performs nearly as well as nbsModel and the results of these two methods are complementary to each other. Statistically significant ($P < 0.05$) results are obtained by both methods in four cancer types, i.e., HNSC, LUAD, BRCA and KIRP. Statistically significant ($P < 0.05$) results in LUSC, OV, COAD and CESC or BLCA, KIRC, UCEC, LIHC and STAD are obtained by xGeneModel or nbsModel, respectively. The performance of 2GeneModel and ccpwModel is relatively less desired. Each of them achieves statistically significant results ($P < 0.05$) in six cancer types. Nevertheless, they still show special strength in clustering tumor samples of some cancer types. For example, the survival stratification of the partition obtained by ccpwModel in LIHC is much clearer than that from nbsModel. Based on the result, we identify a robust prognostic signature for liver cancer (Sub-section 3.2). Another example is the result that 2GeneModel obtains in UCEC. It clearly shows that the UCEC patients with tumors in which both PTEN and PICK3A genes are mutated demonstrate a desired survival curve, a highly useful signature for predicting the prognosis of UCEC patients (Page-15 of Supplementary Material).

Compared to other methods, xGeneModel is also superior in that it is the only algorithm achieving survival-associated partitions in OV and CESC samples (Page-30 and 34 of Supplementary Material). Especially for CESC, the racial disparities of mutations are also found in KIRC by xGeneModel and ccpwModel. The racial disparity in COAD is uniquely identified by xGeneModel.

### 3.2 Extended applications

As shown in Table 1, clinically meaningful stratifications are generated by at least one clustering method in 13 (out of 16) cancer types. The identified tumor clusters constitute a catalogue of potential cancer sub-types. Compared to the sub-types identified by gene expression profiling, a mutation-based sub-type may be more relevant to the etiology of carcinogenesis and therefore, is more useful for personalized therapy. Inspired by this perception, we extend the study to pinpoint the prognostic signatures that are not only quantified by
gene expression levels but also characterized by somatic mutation spectra of driver genes. Since numerous gene expression data of cancer samples has been deposited in public repositories while large-scale mutation collections are rather limited, this type of investigation would facilitate the utilization of the available genome-wide information in precision medicine. Hereby, as a showcase, we report some encouraging results obtained for BLCA and LIHC.

3.2.1 A mutation-characterized transcriptomic prognostic signature in BLCA

Using xGeneModel, we partition the BLCA samples into four clusters (Fig. 2). Among them, Cluster-3 is of our special interest because ~95% of patients in this group have mutated TP53 gene and its survival profile is poorer than that of the other clusters with a modest significance ($P = 0.07$, logRank test). By comparing Cluster 3 with the aggregate of the other three clusters and setting the cutoff of the $t$-test $P$-value at 0.00001, we select a prognostic signature including 217 genes (Supplementary Text S3) whose expression profiling in a cohort constitutes a transcriptomic prognostic signature. Figure 3 demonstrates the results obtained on the TCGA data and two external microarray datasets of bladder cancer. Plot A presents the survival curves of two transcriptomic signature-based clusters of an enlarged TCGA BLCA set, which contains 233 samples with both mutation and expression information available and 144 samples with only expression profiling. The contrast is extreme-ly significant ($P < 0.01$) in both scenarios when patient age is corrected (Cox-PH analysis) or is not corrected (logRank test). By comparing Cluster-2 in the nbsModel partition. On the other hand, there is a 67% overlap between Cluster-3 (xGn-3) in the xGeneModel partition and the aggregate (nbs-12) of Cluster-1 and Cluster-2 in the nbsModel partition. On the other hand, when the same analysis as described in the previous paragraph is applied to nbs-12, the prediction strength (for patient survival) of the selected signature (i.e. 219 genes with $t$-test $P < 0.0002$) lacks robustness (Supplementary Fig. Set-S5).

3.2.2 A mutation-characterized transcriptomic prognostic signature in LIHC

Using ccpwModel, we identify a lethal LIHC sub-type (Cluster 4), in which tumor cells are characterized by the mutation-disturbed RAS/PI3K pathways (Fig. 4). Another sub-type (Cluster-3), in which tumor cells are characterized by the mutation-disturbed DNA damage control and cell cycle/apoptosis mechanisms, has a relatively low short-term (within 20 months) survival rate. By comparing the aggregate of these two hyper-mutated sub-types with the other tumors in the TCGA cohort and setting the cutoff of the $t$-test $P$-values at 0.0005, we select 78 marker genes (Supplementary Text S4) whose expression profiling in a cohort constitutes a transcriptomic prognostic signature. Similar to the case in BLCA, the signature’s prediction strength is confirmed and validated by the survival analysis of an enlarged TCGA LIHC dataset and two external microarray datasets of liver cancer (Fig. 5). Applying nbsModel to the same LIHC data also leads to a survival-related partition of tumors, as presented on Page-69 of Supplementary Material, where a good-survival cluster is characterized by the lack of commonly mutated driver genes. However, the transcriptomic prognostic signature obtained by comparing the gene expression levels of this cluster with those of the aggregate of the other three clusters lacks robustness (Supplementary Fig. Set-S6). This situation cannot be improved by simply changing the signature size (i.e. gene number) based on different cutoffs of the $t$-test $P$-values.

| Cancer | Cluster-survival association | Cluster-race association |
|--------|-----------------------------|-------------------------|
|        | 2GeneModel | xGeneModel | ccpwModel | nbsModel | 2GeneModel | xGeneModel | ccpwModel | nbsModel |
| BLCA   | *           | *           | *           | **        | *           | *           | *           | **        |
| GBM    | *           | *           | *           | *         | *           | *           | *           | **        |
| HNSC   | **          | **          | **          | **        | **          | **          | **          | **        |
| KIRC   | *           | *           | *           | *         | *           | *           | *           | *         |
| LUAD   | *           | *           | *           | *         | *           | *           | *           | *         |
| LUSC   | *           | *           | *           | *         | *           | *           | *           | *         |
| BRCA   | **          | **          | *           | *         | **          | **          | **          | **        |
| OV     | *           | *           | *           | *         | *           | *           | *           | *         |
| UCEC   | *           | *           | *           | *         | *           | *           | *           | *         |
| COAD   | *           | *           | *           | *         | *           | *           | *           | *         |
| THCA   | *           | *           | *           | *         | *           | *           | *           | *         |
| CESC   | *           | *           | *           | *         | *           | *           | *           | *         |
| ESCA   | *           | *           | *           | *         | *           | *           | *           | *         |
| KIRP   | *           | *           | *           | *         | *           | *           | *           | *         |
| LIHC   | *           | *           | *           | *         | *           | *           | *           | *         |
| STAD   | *           | *           | *           | *         | *           | *           | *           | *         |

Note: *0.01<=P<0.05; **0.001<=P<0.01 and ***P<0.001.
A major goal of somatic mutation-based clustering analysis of tumor samples is to identify cancer sub-types that can mediate the genetic etiology of disease cells as well as clinical outcomes and pathologic characteristics of patients. In this regard, biomedical evaluation of the clustering results is more relevant to precision medicine than statistical optimization. Due to the varied carcinogenic mechanisms, not a single clustering method could outperform others in all cancer types, as shown in this study. In fact, the results of various methods tend to be complementary to each other. Those results collectively constitute a catalogue of tumor stratification patterns, which may represent potential cancer sub-types that warrant further investigation.

The primary contributions of this study are the two newly proposed methods. In the ccpwModel, the binary variables indicating the status of cancer driver genes in tumor samples and those genes’ involvement in the CCPWs are considered as features in the unsupervised learning process. In xGeneModel, the functional similarities of the putative cancer driver genes and their confidence scores as the ‘true’ driver genes are integrated with the mutation events to calculate the genetic distance between tumor samples. From these unique angles, we address the four issues faced by the mutation-based clustering of tumors as discussed in the Introduction Section. A remaining challenge is how to further improve the sophistication of several operations, such as determining cancer-relevant functional correlations between genes and calculating the confidence score of a putative cancer driver gene.
Elimination of racial disparities in cancer screening, diagnosis, treatment, and mortality is an essential step toward the improvement of health outcomes for all cancer patients (Koh, 2009). The promise of this objective depends on addressing and identifying the underlying social-economic and biological causes. In this study, we find statistically significant associations between the mutation-based tumor clusters and the racial groups of patients in six cancer types. Due to the limited sizes of cancer cohorts in TCGA data, these associations can hardly be detected by simply comparing the mutation frequencies of a single gene between racial groups. To our knowledge, we are the first to investigate cancer racial disparities in such a manner. Taking the results of bladder cancer (BLCA) as an example, we can infer the biomedical relevance of this analysis as follows. Using xGeneModel, 233 BLCA tumor samples are partitioned into four clusters (sub-types), which not only stratify in survival time but are significantly associated with patient race (Fig. 2). In particular, Cluster-3 is characterized by 95% mutation rate in the TP53 gene. Suppose that, in the future, a therapy targeting this cancer sub-type is developed; it would be more effective for a white patient than for an Asian patient. This is because over 50% of white patients are assigned to Cluster-3 but the proportion of Asian patients in this cluster is about 25%.

Another unique aspect of this study is that it demonstrates a way to integrate the results of mutations based tumor clustering with the widely available gene expression data for prognostic signature.
Driver gene mutations based clustering of tumors

Lawrence, M.S. et al. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature, 499, 214–218.
Lee, D.D. and Seung, H.S. (1999) Learning the parts of objects by non-negative matrix factorization. Nature, 401, 788–791.
Monti, S. et al. (2013) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Mach. Learn, 52, 91–118.
Reimand, J. and Bader, G.D. (2013) Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Mol. Syst. Biol., 9, 637.
Rister, M. et al. (2012) Combination of a novel gene expression signature with a clinical nomogram improves the prediction of survival in high-risk bladder cancer. Clin. Cancer Res., 18, 1323–1333.
Robles, A.I. and Harris, C.C. (2010) Clinical outcomes and correlates of TP53 mutations and cancer. Cold Spring Harb. Perspect. Biol., 2, a001016.
Roessler, S. et al. (2012) Integrative genomic identification of genes on 8p associated with hepatocellular carcinoma progression and patient survival. Gastroenterology, 142, 957–966, e912.
Szklarczyk, D. et al. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res., 39, D561–D568.
Tamibororo, D. et al. (2013a) Comprehensive identification of mutational cancer driver genes across 12 tumor types. Sci. Rep., 3, 2650.
Tamibororo, D. et al. (2013b) OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics, 29, 2238–2244.
Tomasetti, C. et al. (2015) Only three driver gene mutations are required for the development of lung and colorectal cancers. Proc. Natl. Acad. Sci. USA, 112, 118–123.
Vauthey, J.N. et al. (2013) RAS mutation status predicts survival and patterns of recurrence in patients undergoing hepatectomy for colorectal liver metastases. Ann. Surg., 258, 619–626, discussion 626–617.
Villa, E. et al. (2016) Neangiogenesis-related genes are hallmarks of fast-growing hepatocellular carcinomas and worst survival. Results from a prospective study. Gut, 65, 861–869.
Vogelstein, B. et al. (2013) Cancer genome landscapes. Science, 339, 1546–1558.
Yu, G. et al. (2010) GOSeSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics, 26, 976–978.
Zhang, W. et al. (2016a) Integrative genomics and transcriptomics analysis reveals potential mechanisms for favorable prognosis of patients with HPV-positive head and neck carcinomas. Sci. Rep., 6, 24927.
Zhang, W. et al. (2016b) The modularity and dynamicity of miRNA-mRNA interactions in high-grade serous ovarian carcinomas and the prognostic implication. Comput. Biol. Chem., 63, 3–14.