PLOS ONE

RESEARCH ARTICLE

Assessment of intratumoral heterogeneity with mutations and gene expression profiles

Ji-Yong Sung[1,2], Hyun-Tae Shin[1,2], Kyung-Ah Sohn[3], Soo-Yong Shin[4,5], Woong-Yang Park[1,2,6], Je-Gun Joung[1]*

1 Samsung Genome Institute, Samsung Medical Center, Seoul, Korea, **2** Department of Health Science and Technology, Samsung Advanced Institute of Health Science and Technology, Sungkyunkwan University, Seoul, Korea, **3** Department of Software and Computer Engineering, Ajou University, Suwon, Korea, **4** Department of Digital Health, Samsung Advanced Institute of Health Science and Technology, Sungkyunkwan University, Seoul, Korea, **5** Big Data Research Center, Samsung Medical Center, Seoul, Korea, **6** Department of Molecular Cell Biology, Sungkyunkwan University School of Medicine, Seoul, Korea

* jegun.joung@samsung.com

Citation: Sung J-Y, Shin H-T, Sohn K-A, Shin S-Y, Park W-Y, Joung J-G (2019) Assessment of intratumoral heterogeneity with mutations and gene expression profiles. PLoS ONE 14(7): e0219682. https://doi.org/10.1371/journal.pone.0219682

Editor: Ilya Ulasov, Sechenov First Medical University, RUSSIAN FEDERATION

Received: January 18, 2019

Accepted: June 30, 2019

Published: July 16, 2019

Data Availability Statement: Datasets and source codes in this paper are available at https://doi.org/10.5281/zenodo.1244203.

Funding: This research was supported by the Samsung Medical Center, and the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Science, ICT & Future Planning (JGJ: 2017R1A2B1007347) and (JGJ: 2018R1D1A1B07048531). The funders had no role in study design, data collection and analysis,

Abstract

Intratumoral heterogeneity (ITH) refers to the presence of distinct tumor cell populations. It provides vital information for the clinical prognosis, drug responsiveness, and personalized treatment of cancer patients. As genomic ITH in various cancers affects the expression patterns of genes, the expression profile could be utilized for determining ITH level. Herein, we present a novel approach to directly detect high ITH defined as a larger number of subclones from the gene expression pattern through machine learning approaches. We examined associations between gene expression profile and ITH of 12 cancer types from The Cancer Genome Atlas (TCGA) database. Using stomach adenocarcinoma (STAD) showing high association, we evaluated the performance of our method in predicting ITH by employing three machine learning algorithms using gene expression profile data. We classified tumors into high and low heterogeneity groups using the learning model through the selection of LASSO feature. The result showed that support vector machines (SVMs) outperformed other algorithms (AUC = 0.84 in SVMs and 0.82 in Naïve Bayes) and we were able to improve predictive power by using both combined data from mutation and expression. Furthermore, we evaluated the prediction ability of each model using simulation data generated by mixing cell lines of the Cancer Cell Line Encyclopedia (CCLE), and obtained consistent results with using real dataset. Our approach could be utilized for discriminating tumors with heterogeneous cell populations to characterize ITH.

Introduction

Intratumoral heterogeneity is defined as different tumor cells as being capable of exhibiting distinct morphological and phenotypic profiles, including cellular morphology, gene expression, metabolism, proliferation, and metastatic potential [1]. The research on tumor heterogeneity is essential to figure out the composition of the tumor for personalized treatment of cancer patients. The ITH studies are attracting attention as an important concept in the

decision to publish, or preparation of the manuscript.

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

understanding of drug resistance and cancer recurrence, which is a critical issue in cancer treatment, and in deciphering the mechanism of acquiring resistance to target chemotherapy [2]. Therefore, understanding the heterogeneity of tumors is very important, not only in designing the therapeutic approach to find the potential target for chemotherapy, but also in the field of clinical diagnosis for prediction of prognosis and therapeutic response of cancer patients [3].

Stomach adenocarcinoma (STAD) is the most common cancer in the world and has a very high mortality rate [4]. It is particularly interesting to understand the molecular characteristics of STAD from a genomic point of view through the application of next generation sequencing (NGS) technology and to study the causes and consequences of stomach adenocarcinoma through the assessment of ITH. In STAD, as well as in several other cancer types, ITH study is invaluable for identifying suitable biomarkers and correct therapeutic strategy because each subpopulation of cells (i.e., a subclone of cancer cells) contains distinct genetic information [5]. Currently, researchers have calculated the purity/ploidy of tumors [6] to know how many tumor cells are inherently contained on the basis of the copy number alteration data and somatic mutation data obtained from the carcinoma. In addition, the degree of tumor heterogeneity has been determined by estimating the number of sub-clones existing within a tumor [7].

Nonetheless, the measurement of ITH by identifying sub-clones using genome profiles requires complex and intensive analytical procedures, such as the calculation of tumor purity, ploidy, variant allele frequency (VAF), cancer cell fraction (CCF), and clustering. Recent study has suggested that high level of ITH affects the recurrence risk based on the gene expression profile [8]. Another approach has demonstrated that the degree of ITH could be measured by the biological network entropy using gene expression data [9]. Thus, their previous studies suggest that not only genomic data but also gene expression patterns can be used as invaluable information to identify ITH.

A method is needed for easily determining the tumor heterogeneity by utilizing various types of omics data. We propose a method to classify tumors with high heterogeneity exhibited as a larger number of subclones according to the mutation and expression profiles of genes that may be associated with stomach adenocarcinoma. We tried to distinguish samples with high tumor heterogeneity by combining the mutation data set and the gene expression data set to obtain meaningful insights for determining tumor cell diversity. During this process, our method could easily find potential biomarkers through a feature selection and conveniently determine candidate tumor samples with high heterogeneity.

Results

Relationship between ITH and RNA expression

Fundamentally, gene expression profiles have a functional importance for clonal evolution so that we have analyzed RNA expression profiles in 12 cancer types from the Broad GDAC Firehose (https://gdac.broadinstitute.org/). In our study, ITH was defined as the number of distinct tumor cell populations (subclones), measured by using EXPANDS and PyClone [10]. We performed the correlation analysis between clone number and gene expression profile in order to examine which cancer types are highly correlated to ITH. Overall expression profiles of many genes were significantly correlated ($p < 0.01$, Spearman's rank correlation) in Lung squamous cell carcinmoma (LUSC) and Stomach adenocarcinoma (STAD) than other cancer types (Fig 1A). 303 (1.47% of total genes) and 299 genes (1.45%) were significant, respectively. Among them, several oncogenes and tumor suppressors were included (5 and 10 in LUSC, 10 and 13 in STAD). Next, we identified enriched functions associated with those genes. STAD

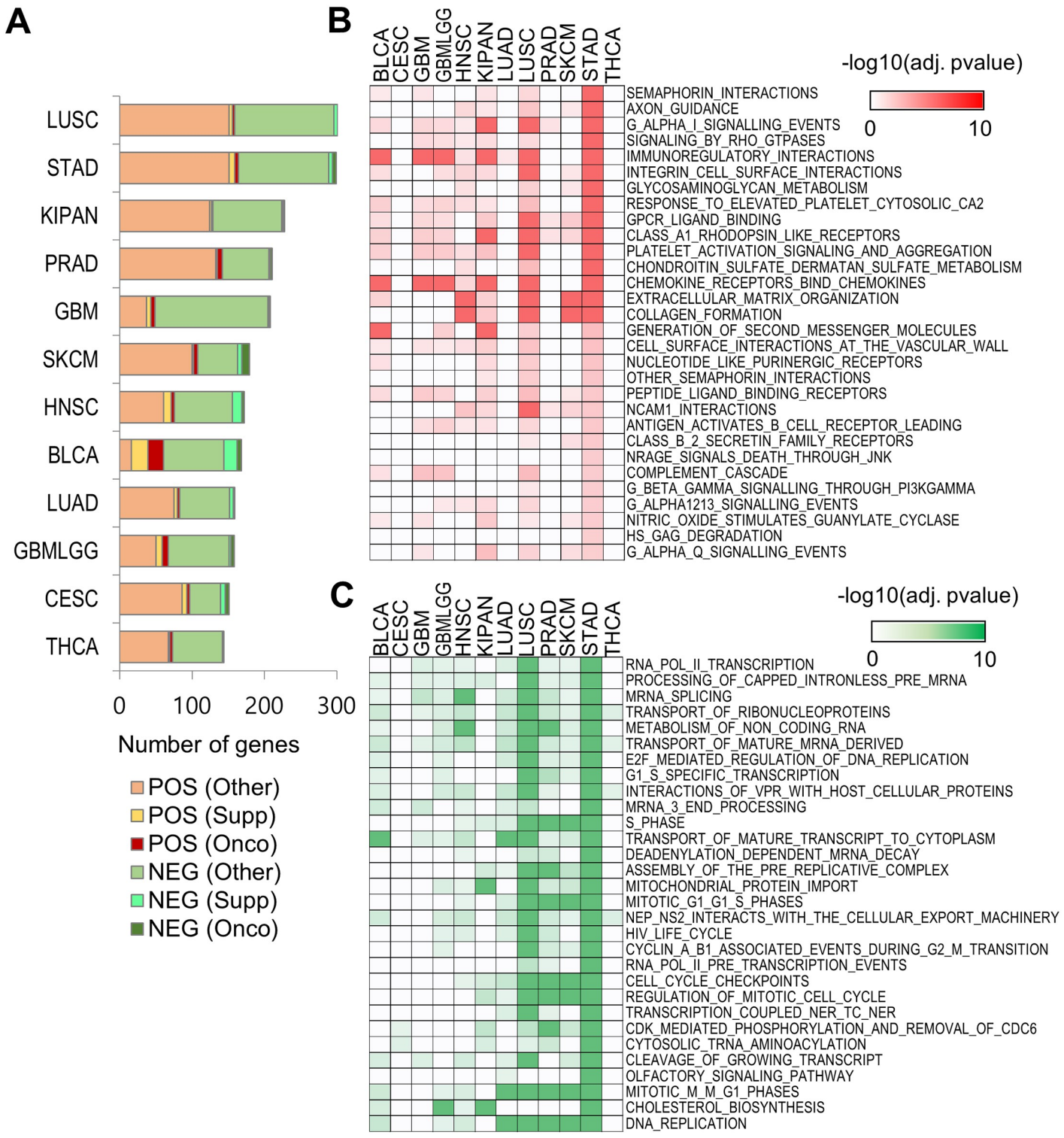

Fig 1. Association of RNA expression with ITH in 12 tumor types from TCGA. (A) Bar plot demonstrating comparison between the differential RNA gene expression in 12 tumor types from TCGA. BLCA: Bladder urothelial carcinoma; CESC: Cervical and endocervical cancers; GBM: Glioblastoma multiforme; GBMLGG: Glioma; HNSC: Head and neck squamous cell carcinoma; KIPAN: Pan-kidney cohort; LUAD: Lung adenocarcinoma; LUSC: Lung squamous cell carcinoma; PRAD: Prostate adenocarcinoma; SKCM: Skin cutaneous melanoma; STAD: Stomach adenocarcinoma; THCA: Thyroid carcinoma. POS: Positive; NEG: Negative; Onco: Oncogene; Supp: Suppress gene; Other: Other gene. (B) and (C) Heat map of functional categories of genes positively and negatively correlated with ITH, respectively.

https://doi.org/10.1371/journal.pone.0219682.g001

showed the most prominent association with both positively and negatively correlated genes with ITH (Fig 1B and 1C). And there are more samples in STAD with clone than other cancer types. CESC and THCA did not have enriched function. Based on high correlation and available enough data for subclone numbers with ITH in STAD, we performed following analysis for genomic and transcriptomic features of stomach adenocarcinoma. Here, this cohort is Stomach adenocarcinoma (STAD) rather than Stomach and Esophageal carcinoma (STES) in the Broad GDAC.

Characteristics of genomic and transcriptomic markers in ITH of stomach adenocarcinoma

A variety of approaches could be considered to determine the factors affecting the tumor heterogeneity of stomach adenocarcinoma. We examined the genomic and transcriptomic characteristics of 128 tumor samples with high heterogeneity and low heterogeneity. We assume that tumor suppressor genes and oncogenes are found frequently in tumors with high heterogeneity rather than in those with low heterogeneity. As expected, the distribution of both tumor-driver genes indicates that they were enriched in tumors with high heterogeneity (Fig 2). Among the 63 tumor suppressor genes, 12 genes had mutations and a total of 147 mutations were present in tumors with high heterogeneity. On the other hand, those genes harbored only 60 mutations in tumors with low heterogeneity. Most of the genes, except for *TP53*, had a higher mutation frequency in the high heterogeneity group. The frequency in oncogenes was also different between both the groups, showing more abundance (87 mutations) in the high heterogeneity group than in the low heterogeneity group (30 mutations). Among oncogenes, *MYCBP2* was the most prominent in the high heterogeneity group whereas *PIK3CA* was the most frequent in the low heterogeneity group (S1 Fig). In all the tumor stages, oncogene and tumor suppressor genes were prevalent in tumors with high heterogeneity (S2 Fig). We could confirm that there was genomic difference between both the groups in the tumor associated genes.

We also assume that the differences in the expression levels of many genes might cause functional differences associated with high intratumoral heterogeneity. Thus, we identified differentially expressed genes (DEGs) between the low and high heterogeneity groups using the R package, DEseq [11]. The *COL11A2* genes were highly expressed in the high heterogeneity group than in the low heterogeneity group and the *CLDN22* genes were expressed at low levels ($>$ 2-fold change and adj $p < 0.05$). We were able to confirm the difference in the gene expression in both the high and low heterogeneity groups. Next, we identified enriched functions through GSEA (Gene Set Enrichment Analysis) (Fig 3A). We could find several pathways showing functional differences, including "G-protein alpha signaling", "Extracellular matrix organization", "Degradation of extracellular matrix", "Integrin cell surface interactions", "Signaling by GPCR" and "Adherence junctions interactions". Among them, the G-protein alpha signaling pathway is mainly involved in signal transmission of cells [12] and low expression levels of genes in that pathway were associated with high ITH in our analysis (Fig 3B). Recent studies have shown that G protein alpha signaling affects the stomach adenocarcinoma growth through p53/p21 and MEK/ERK pathways [13].

Performance comparison of classifiers

We classified tumor samples into low and high heterogeneity groups by applying three prediction methods. Three types of datasets (mutation, RNA expression, and both the datasets combined) were tested (Fig 4). First, we compared the performance of the prediction methods according to the data types. The performance of each run was evaluated by 10-fold cross

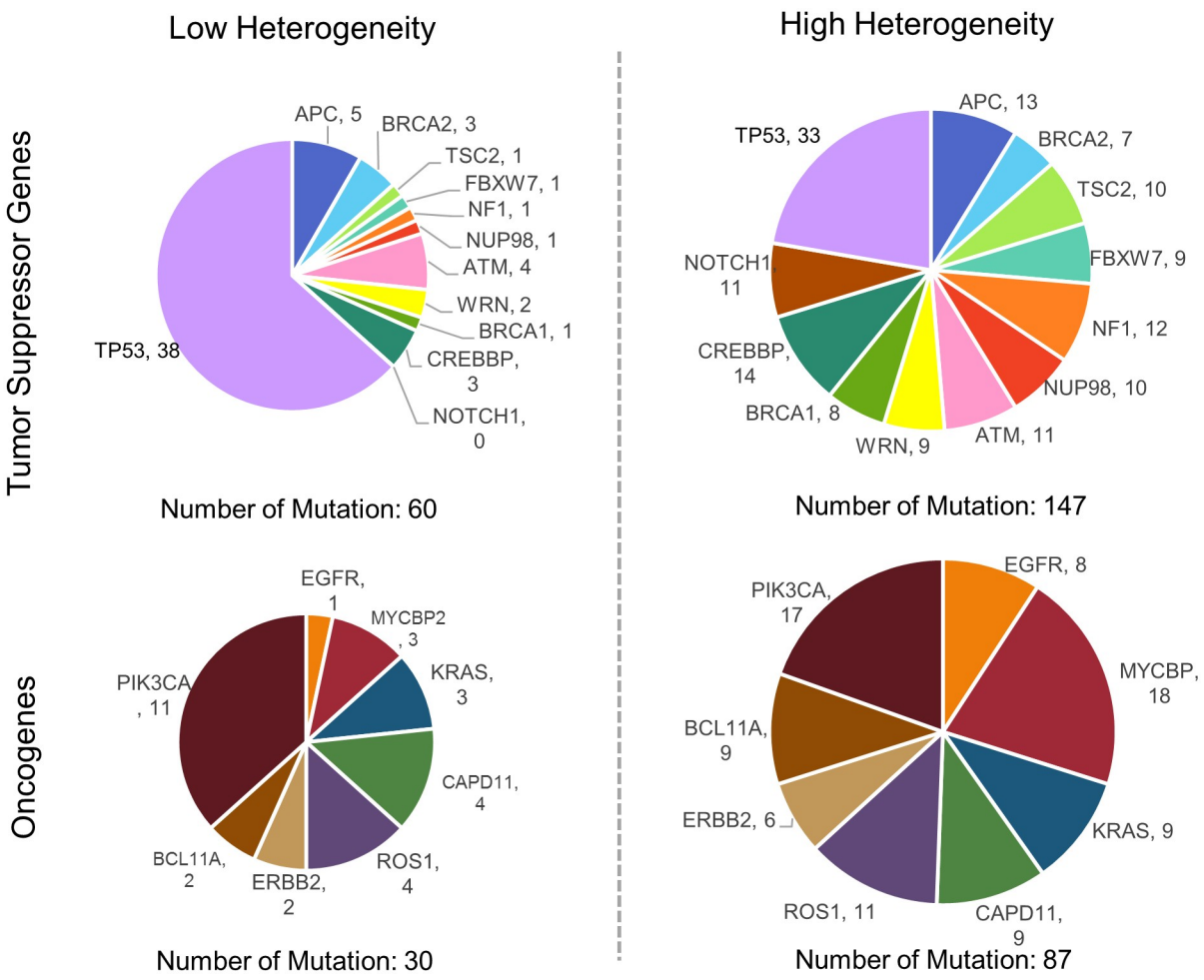

Fig 2. Pie charts representing the frequency of tumor suppressor genes and oncogenes between the low and high heterogeneity groups. Each gene name and its mutation count in a group are displayed.

https://doi.org/10.1371/journal.pone.0219682.g002

validation. The prediction accuracy is summarized in Table 1. Each value is the average of 10 runs. Overall, the performance was better when the two datasets were combined than when only one dataset was used. In terms of methods, the SVM outperformed the other methods for the two data types, exhibiting 71.02% in the mutation and 75.54% in the combined dataset. The SVM method achieved the highest area under the curve (AUC) (0.84) using the combined dataset (shown in Table 1 and Fig 5A). In the case of NB, the sensitivity in the combined dataset was the highest (0.90), but the specificity was the lowest (0.49). In conclusion, the best performance was when we applied SVM with the combined dataset.

Next, we tested whether the prediction was stable when a limited subset of genes, out of the total set, was applied, implying loss of information. A subset of genes was randomly selected from the total number of genes. As expected, the performance with regard to the classification of heterogeneity was decreased with high variance with the decrease in the size of gene set when we tested with SVM (Fig 5B). This suggests that it is important to generate a prediction model with the most available features before the feature selection.

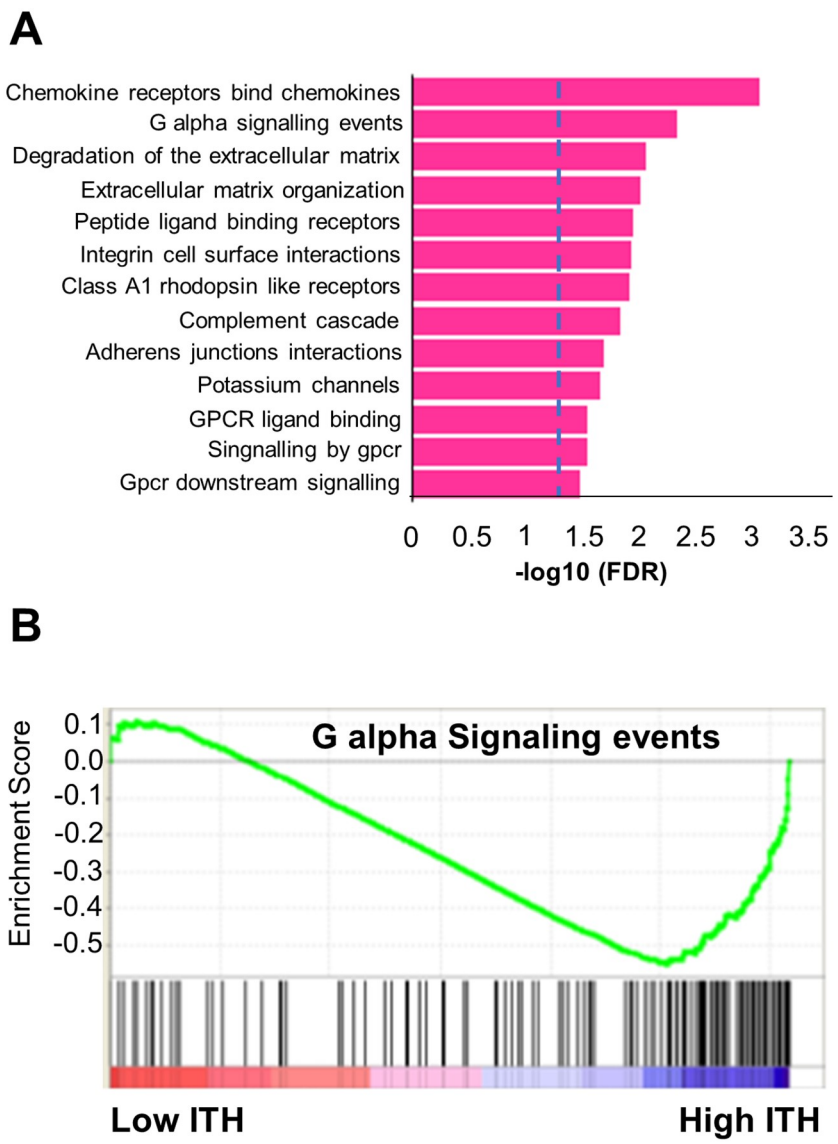

Fig 3. Functional differences between the low and high heterogeneity groups. (A) Enriched functions in genes showing expression difference between both the groups. (B) Enrichment of genes belonging to G alpha signaling events.

https://doi.org/10.1371/journal.pone.0219682.g003

Candidate genomic features for stomach adenocarcinoma with high heterogeneity

We were able to obtain a list of genes that contributed to the classification using the Lasso [14] feature selection method (shown in S1 Table). When we selected the top ranked genes, the *MKRN3* gene had the highest coefficient value in the mutation dataset and the *PDGFRA* gene had the highest value in the RNA expression dataset. Especially, the *PDGFRA* gene has been associated with tumor progression in stomach adenocarcinomas and has been associated with a poor prognosis [15]. The *PLEC* gene showed high difference in the mutation frequency between the low and high heterogeneity groups. The *PLEC* gene has been associated with binding of protein coding and actin binding. Also, *PLEC* is an important gene for cell pleomorphism and cytoskeleton disorganization [16].

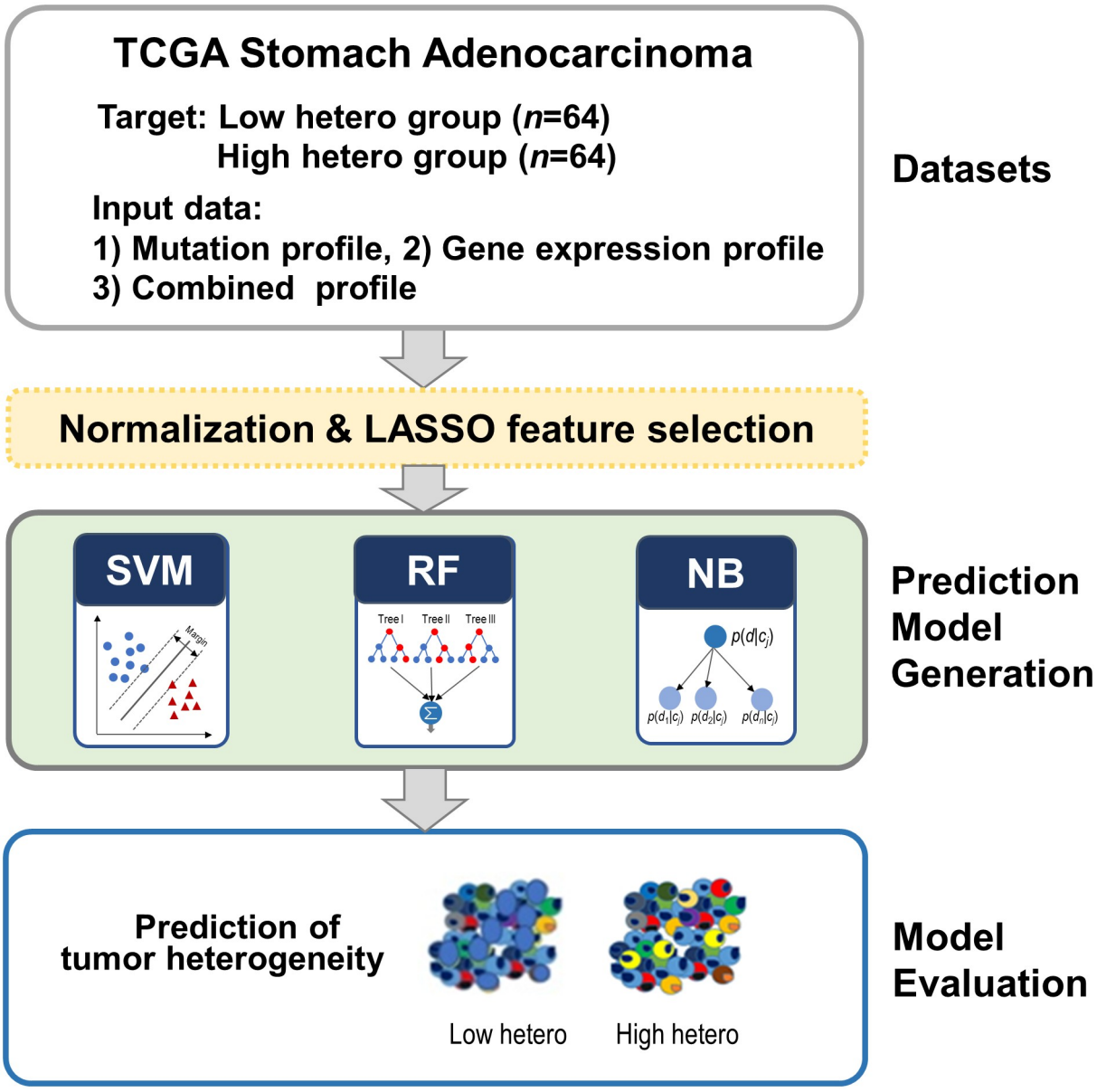

Fig 4. Schematic flow of tumor heterogeneity prediction. We performed machine learning in three steps: 1) preparation of mutation and RNA expression data set, 2) prediction model construction, and 3) evaluation of constructed model.

https://doi.org/10.1371/journal.pone.0219682.g004

Table 1. Comparison of the predictive performance according to the real datasets.

Data Type	Methods		
	Naïve Bayes	Support Vector	Random Forests
Mutation (Accuracy)	65.23%	71.02%	68.83%
RNA Expression (Accuracy)	71.33%	70.08%	69.84%
Combined Data (Accuracy)	69.38%	75.54%	72.87%
Combined Data (Sensitivity/Specificity)	0.90/0.49	0.73/0.78	0.75/0.69
Combined Data (AUC)	0.82	0.84	0.81

https://doi.org/10.1371/journal.pone.0219682.t001

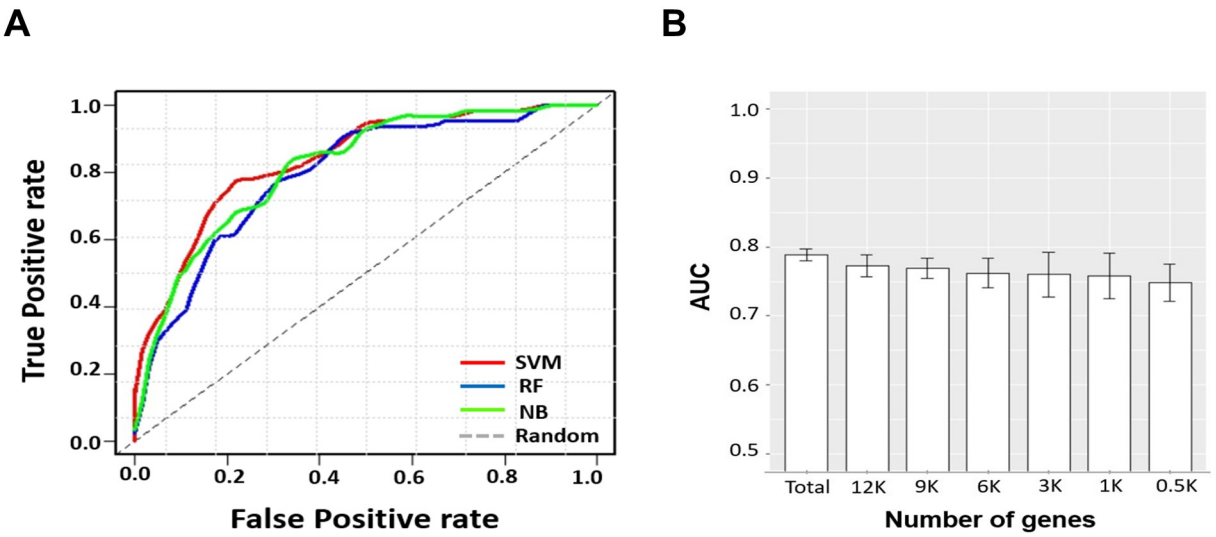

Fig 5. Performance comparison of classifiers. (A) Receiver operating characteristic curve of predictive performance on three classifiers. (B) Measurement of the performance with limited gene set in the RNA expression dataset with SVM.

https://doi.org/10.1371/journal.pone.0219682.g005

Performance test of predictive models with simulation data

We confirmed that the prediction model could distinguish the tumor heterogeneity when multiple tumor samples were randomly mixed. The simulation datasets were generated by using mutation profiles from the Cancer Cell Line Encyclopedia (CCLE) dataset and RNA-seq expression profiles from the CCLE and GTEx dataset, as described in the Methods section. A total of 150 mixed samples having diverse heterogeneity were used to evaluate the prediction models. As shown in Table 2, the performance was better in the RNA expression dataset than in the mutation dataset. Furthermore, the prediction models achieved the best performance in the combined dataset (showing an accuracy of 80.93% and AUC of 0.88 for SVM). This showed consistent results with the results obtained by using real datasets.

Among the parameters related to the composition of tumor samples, the distribution of tumor purity might affect the prediction performance. We examined as to how the prediction performance changed with the average tumor purity (Fig 6). Overall, the prediction ability was affected by tumor purity when using the mutation profile, and tended to decline at the setting of low purity. Among the three algorithms, the performance of NB was affected for all the dataset types. However, the performance of SVM and RF showed a small change according to the setting of various tumor purities when using RNA expression and combined dataset. This indicates that those algorithms are more tolerant to diverse tumor purities when using the RNA expression dataset.

Table 2. Comparison of the predictive performance according to the data type on simulation datasets.

Data Type	Methods		
	Naïve Bayes	Support Vector	Random Forests
Mutation (Accuracy)	58.27%	61.40%	59.87%
RNA Expression (Accuracy)	69.80%	73.87%	68.87%
Combined Data (Accuracy)	73.47%	80.93%	76.80%
Combined Data (Sensitivity/Specificity)	0.65/0.81	0.81/0.81	0.77/0.76
Combined Data (AUC)	0.83	0.88	0.85

https://doi.org/10.1371/journal.pone.0219682.t002

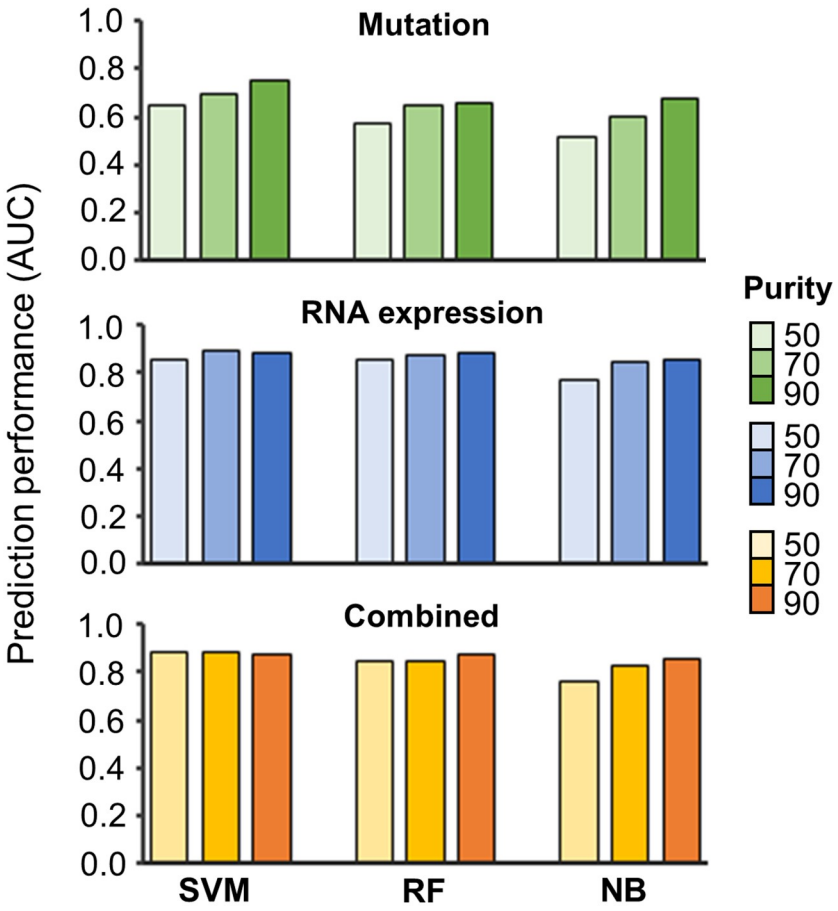

Fig 6. Measurement of the performance with limited gene set in the RNA expression dataset. The simulation shows the performance of each prediction method according to different tumor purity (50, 70 and 90).

https://doi.org/10.1371/journal.pone.0219682.g006

Discussion

In this study, we suggest a novel framework to classify tumor samples according to their heterogeneity, representing the degree of different cell populations. We demonstrated the first prediction approach that could distinguish samples with heterogeneity using the mutation and RNA expression data, as well as a combination of both the data. Our prediction models showed reliable performance when using a real dataset and the results were also consistent with the simulation data. Our approach could be utilized to conveniently find tumor samples showing high heterogeneity as well as for easily finding the genomic and transcriptomic markers through feature selection.

Our approach has several advantages. First, for systematic analysis, tumor samples with high heterogeneity could be found conveniently. Second, our method could improve the performance of distinguishing heterogeneity through the addition of different types of data sets including mutation and transcriptomics datasets. Finally, the features associated with tumor heterogeneity can be automatically extracted.

We explicitly examined the characteristics of each group in detail by dividing the samples into groups with low and high heterogeneity. In the group with high tumor heterogeneity, mutations in the tumor suppressor genes and oncogenes were relatively more than in the group with low heterogeneity. Most mutations were detected in both high heterogeneity group

and low heterogeneity group, with different percentage of mutations in *PIK3CA*, *KRAS* and *APC*. Those oncogenes were more found in high heterogeneity group than low group (*PIK3CA*: 26.6% in high-heterogeneity group and 17.2% in low-heterogeneity group; *KRAS*: 14.1% and 4.7%; *APC*: 20.3% and 7.8%). Previous study shows that mutational heterogeneity in *APC* and *KRAS* could cause polyclonality in early colorectal tumorigenesis [17]. Another study suggested that *PIK3CA*, *SMAD4* and *TP53* are most often associated with clonal divergence [18]. These driver genes may cause an increase of tumor heterogeneity via clonal expansion.

Also, the amount of RNA expression was significantly different between the high and low heterogeneity groups. Because there were apparently different genomic and transcriptomic alterations between both the groups, these important changes might contribute toward the high heterogeneity in tumor progression. Although the top-ranked mutated genes selected by feature selection are not well known in tumorigenesis, their occurrences were distinctly different between both the groups. The cause and effect of this association would prove invaluable for further studies.

There are different levels of ITH, e.g., phenotypic ITH such as heterogeneity in cell shape and morphology as well as non-genetic heterogeneity ITH such as heterogeneity in signaling [19] and epigenetic [20]. In this study, we focused on genetic aspect of ITH. However, considering different levels of ITH can help to understand the overall mechanism of tumor progress accurately and globally. In the future, it may be possible to conduct classification studies of ITH using these data.

In addition, there are the several issues on identifying ITH with single cell sequencing data. Recently they have been utilized for studying tumor heterogeneity [21–23]. The evaluation with single-cell RNA sequencing (scRNA-seq) data could be more reliable for studying ITH compared to that using bulk data. scRNA-seq analysis has an advantage that it can directly measure different tumor cell populations while analysis using the bulk sequencing analysis infers indirectly them. The profile of single cells can be deconvoluted with tools such as MuSiC [24] and CIBERSORT [25] to find out the cell type and the heterogeneity can be measured by grasping different molecular cell contents. This analysis will allow us to accurately evaluate tumor heterogeneity, as it provides more comprehensively the information of actual tumor microenvironment compared to analysis of bulk data or the mixture of cell lines.

As a future work, deep learning techniques will be applied [26]. The current machine learning approaches have a limitation with feature selection. To overcome this limitation and to improve the performance, we will try deep learning to classify ITH.

In summary, through the performance evaluation of the prediction models for ITH, we confirmed the plausibility of the proposed approach that could predict high tumor heterogeneity. We expect that this study will provide a novel insight for interpreting intratumoral heterogeneity so that it can be utilized for clinical implementation in diverse cancers.

Methods

Correlation analysis between ITH and gene expression

The Cancer Genome Atlas (TCGA) dataset (version 2016.01.28) for 12 cancer types was downloaded from Broad GDAC Firehose (http://firebrowse.org/). The full name and the abbreviation of these cancer types are as follows: bladder urothelial carcinoma (BLCA), cervical and endocervical cancers (CESC), glioblastoma multiforme (GBM), low grade glioma (GBMLGG), head and neck squamous cell carcinoma(HNSC), pan-kidney cohort (KIPAN), lung adenocarcinoma(LUAD), lung squamous cell carcinoma (LUSC), prostate adenocarcinoma (PRAD), skin cutaneous melanoma (SKCM), stomach adenocarcinoma (STAD), and thyroid

carcinoma (THCA). To obtain information on ITH, we used the number of clones calculated by PyClone and EXPANDS from previous studies [10]. PyClone performs Bayesian clustering by grouping somatic mutations into putative subclones. It infers the cellular prevalence of somatic mutations by taking both variant allele frequencies and copy number alterations [7]. EXPANDS estimates the number of subclones based on hierarchical clustering analysis with cell-frequency probability distributions of somatic mutations [27]. The correlations between ITH (i.e. the number of clones) and gene expression were measured by Spearman's correlation coefficient (PCC) and the significant genes ($p < 0.01$) were counted. In order to identify biological functions associated with ITH, gene set enrichment analysis (GSEA) (software.broadinstitute.org/gsea/) was performed on REACTOME pathway gene sets using pre-ranked gene list of PCC.

ITH prediction using genomic and transcriptomic datasets

The overview of the present study is presented in Fig 4. We first collected genomic and transcriptomic datasets from stomach adenocarcinoma patients and combined them into a single dataset. The target values for prediction were determined by the number of clones. We employed a machine learning method to predict the heterogeneity in real data. Furthermore, to evaluate the accuracy of prediction, we applied our predictive model to hypothetically heterogeneous tumor datasets generated from mixed cell lines. Our prediction model achieved the best performance in the combined dataset and demonstrated the plausibility of predicting high heterogeneity in stomach adenocarcinoma.

Evaluation data for ITH prediction

From TCGA stomach adenocarcinoma (STAD) cohort, we obtained the mutation profiles of 289 patients and RNA expression profiles of 450 patients. We selected 128 patients that had the same sample ID as that of data for mutation, gene expression, and clone number. To classify the tumors according to ITH in stomach adenocarcinoma, the label of tumors with high heterogeneity (number of clones ≥ 6) was set to 1 and the label of those with low heterogeneity (number of clones < 6) was set to 0, based on the presence of well-balanced labels for positive and negative sets (S3 Fig). We defined target labels, $y = \{y_1, y_2, \ldots, y_N\}$, $i = 1, \ldots, N$ and $y_i \in \{0, 1\}$, where N is the total number of patients. As input data for a model, three types of datasets were considered. The set of mutation was denoted by $M = \{m_1, m_2, \ldots, m_{N_m}\}$, the set of expression was denoted by $E = \{e_1, e_2, \ldots, e_{N_e}\}$, and the combined set of both was denoted by $EM = M \cup E$, where N_m, and N_e correspond to the total number of features for mutation and expression, respectively. The features with mutation frequency ≤ 5 were excluded and those with low expression (normalized read count < 20) over total samples were also excluded. When $X(M)$ was the input dataset of mutation profile, it could be represented as a $N \times N_m$ matrix and other datasets could also be represented in the same way. Both the combined datasets have the features of $N_{me} = 16{,}383$, merged from the mutation dataset ($N_m = 957$) and the RNA expression dataset ($N_e = 15{,}426$).

Machine learning methods

The feature selection was performed using the least absolute shrinkage and selection operator (LASSO) [14]. The LASSO regression has a function of variable selection that increases the prediction accuracy by reducing the regression coefficient [28], which is an advantage of ridge regression, and makes the regression coefficient values of irrelevant features easily zero at the same time. Therefore, Lasso regression is known as an analytical method that can provide high

prediction accuracy as well as the interpretive power of variable selection. We used the R package 'glmnet' [29] to select the important features. We did feature selection using LASSO from the training data, generated a classification model, and evaluated the performance using a test set. When applying it to the combined data, the features were selected separately from the mutation and expression datasets, and the selected features were merged so as to use more than one type of dataset.

To evaluate the performance of tumor heterogeneity prediction from the selected features, we compared three classification algorithms of Support Vector Machine (SVM), Random Forest (RF), and Naïve Bayes [30–32]. The performance of classifiers was evaluated based on 10-fold cross validation and was compared in terms of Area Under Curve (AUC). It was performed in R (version 3.4.3) and the parameter setting is as follows:

1. We applied radial basis function (RBF) kernel to a SVM classifier using R package 'e1071' (https://cran.r-project.org/web/packages/e1071/). Two parameters, C and sigma, for the RBF kernel SVM were determined by cross-validation with grid search of C = {0.1, 10, 100} and sigma = {0.01, 0.25, 0.5, 1}.
2. We used R package 'klaR' (https://cran.r-project.org/web/packages/klaR) to apply the Naïve Bayes classifier.
3. We used the R package 'randomForest' [32] to collect decision trees from a random subset of the data with standard settings. We determined the best hyper-parameter setting for the number of trees and the minimum size of terminal node through cross-validation for the number of trees = {200, 500, 1000, 2000} and the size of nodes = {2, 3, 4, 5}.

Generation of simulation data

To validate our predictive model using the simulation method, we downloaded mutation data of 38 STAD cell lines from CCLE (https://portals.broadinstitute.org/ccle) and mixed them to generate simulation datasets exhibiting tumor heterogeneity. With high or low heterogeneous tumors, a mutation profile of i-th sample, $X_i(M)$ was generated according to following procedure: 1) The number of clone, N_c, was selected according to random number generation from Poisson distribution with the parameter $\lambda = 6$; 2) the tumor purity value, P_t, was determined from normal distribution with mean value, $-P_t = 70$ and variance, $\sigma^2 = 100$; 3) the size of j-th clone, S_j, was determined by random number generation, where $S_j \in [20, 100]$; 4) the cell lines of N_c were randomly selected, the variant allele frequencies (VAFs) of each variant were mixed based on the proportion of clone size, and each mixed VAF was adjusted by purity; 5) finally, VAFs were converted to binary values by using $VAF_{cutoff} = 0.03$.

To generate mixed RNA expression datasets, we downloaded RNA expression data of tumor cell lines from CCLE (https://www.ebi.ac.uk/gxa/experiments/E-MTAB-2770/Results) as well as the RNA expression data of normal cell lines were downloaded from the GTEx Portal (https://www.gtexportal.org/home/datasets). We used 38 tumor cell lines and 2 normal cell lines. Each mixed expression profile was generated in the same way as described above except for procedure 4 and 5, which were modified as follows: 4) cell lines of N_c were randomly selected, expression value of each gene was mixed based on the proportion of clone size, and the mixed expression was adjusted by purity; 5) the expression values of normal samples were added to the mixed sample by considering the proportion of normal sample as $1 - (P_t/100)$.

A combined dataset of mutation and RNA gene expression was generated by performing two types of procedures simultaneously with the same cells. The two procedures are shown in S4 and S5 Figs, respectively.

Supporting information

S1 Fig. Occurrence of mutation in tumor suppressor genes and oncogenes per tumor sample.
(TIF)

S2 Fig. Comparison between the occurrence of mutations in tumors with high and low heterogeneity.
(TIF)

S3 Fig. Number of positive and negative samples according to the cutoff value for the number of subclones.
(TIF)

S4 Fig. Procedure for generating simulation data of mutations.
(TIF)

S5 Fig. Procedure for generating simulation data of RNA gene expression.
(TIF)

S1 Table. Top ranked genes chosen by the LASSO feature selection from mutation and RNA gene expression data.
(DOCX)

Acknowledgments

The results shown here are in part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov.

Author Contributions

Conceptualization: Hyun-Tae Shin, Je-Gun Joung.

Formal analysis: Ji-Yong Sung.

Investigation: Woong-Yang Park, Je-Gun Joung.

Writing – original draft: Ji-Yong Sung, Je-Gun Joung.

Writing – review & editing: Hyun-Tae Shin, Kyung-Ah Sohn, Soo-Yong Shin, Woong-Yang Park, Je-Gun Joung.

References

1. Marusyk A, Polyak K. Tumor heterogeneity: Causes and consequences. Biochimica et Biophysica Acta (BBA)—Reviews on Cancer. 2010; 1805(1):105–17.
2. Turner NC, Reis-Filho JS. Genetic heterogeneity and cancer drug resistance. The Lancet Oncology. 2012; 13(4):e178–e85. https://doi.org/10.1016/S1470-2045(11)70335-7 PMID: 22469128
3. Greaves M, Maley CC. Clonal evolution in cancer. Nature. 2012; 481(7381):306–13. https://doi.org/10.1038/nature10762 PMID: 22258609
4. Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2017. CA Cancer J Clin. 2017; 67(1):7–30. https://doi.org/10.3322/caac.21387 PMID: 28055103
5. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011; 144(5):646–74. https://doi.org/10.1016/j.cell.2011.02.013 PMID: 21376230
6. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nat Commun. 2015; 6:8971. https://doi.org/10.1038/ncomms9971 PMID: 26634437

7. Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, et al. PyClone: statistical inference of clonal population structure in cancer. Nat Methods. 2014; 11(4):396–8. https://doi.org/10.1038/nmeth.2883 PMID: 24633410
8. Gyanchandani R, Lin Y, Lin HM, Cooper K, Normolle DP, Brufsky A, et al. Intratumor Heterogeneity Affects Gene Expression Profile Test Prognostic Risk Stratification in Early Breast Cancer. Clin Cancer Res. 2016; 22(21):5362–9. https://doi.org/10.1158/1078-0432.CCR-15-2889 PMID: 27185370
9. Park Y, Lim S, Nam JW, Kim S. Measuring intratumor heterogeneity by network entropy using RNA-seq data. Sci Rep. 2016; 6:37767. https://doi.org/10.1038/srep37767 PMID: 27883053
10. Andor N, Graham TA, Jansen M, Xia LC, Aktipis CA, Petritsch C, et al. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. Nat Med. 2016; 22(1):105–13. https://doi.org/10.1038/nm.3984 PMID: 26618723
11. Anders. Differential expression analysis for sequence count data. Genome Biology. 2010; 11:R106. https://doi.org/10.1186/gb-2010-11-10-r106 PMID: 20979621
12. O'Hayre M, Degese MS, Gutkind JS. Novel insights into G protein and G protein-coupled receptor signaling in cancer. Curr Opin Cell Biol. 2014; 27:126–35. https://doi.org/10.1016/j.ceb.2014.01.005 PMID: 24508914
13. Wang Y, Xiao H, Wu H, Yao C, He H, Wang C, et al. G protein subunit alpha q regulates gastric cancer growth via the p53/p21 and MEK/ERK pathways. Oncol Rep. 2017; 37(4):1998–2006. https://doi.org/10.3892/or.2017.5500 PMID: 28350126
14. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc. 1996; vol. 58:(pg. 267–88).
15. Gialeli C, Nikitovic D, Kletsas D, Theocharis AD, Tzanakakis GN, Karamanos NK. PDGF/PDGFR signaling and targeting in cancer growth and progression: Focus on tumor microenvironment and cancer-associated fibroblasts. Curr Pharm Des. 2014; 20(17):2843–8. PMID: 23944365
16. Cheng CC, Lai YC, Lai YS, Chao WT, Tseng YH, Hsu YH, et al. Cell Pleomorphism and Cytoskeleton Disorganization in Human Liver Cancer. In Vivo. 2016; 30(5):549–55. PMID: 27566071
17. Gausachs M, Borras E, Chang K, Gonzalez S, Azuara D, Delgado Amador A, et al. Mutational Heterogeneity in APC and KRAS Arises at the Crypt Level and Leads to Polyclonality in Early Colorectal Tumorigenesis. Clin Cancer Res. 2017; 23(19):5936–47. https://doi.org/10.1158/1078-0432.CCR-17-0821 PMID: 28645942
18. Goswami RS, Patel KP, Singh RR, Meric-Bernstam F, Kopetz ES, Subbiah V, et al. Hotspot mutation panel testing reveals clonal evolution in a study of 265 paired primary and metastatic tumors. Clin Cancer Res. 2015; 21(11):2644–51. https://doi.org/10.1158/1078-0432.CCR-14-2391 PMID: 25695693
19. Kim E, Kim JY, Smith MA, Haura EB, Anderson ARA. Cell signaling heterogeneity is modulated by both cell-intrinsic and -extrinsic mechanisms: An integrated approach to understanding targeted therapy. PLoS Biol. 2018; 16(3):e2002930. https://doi.org/10.1371/journal.pbio.2002930 PMID: 29522507
20. Mazor T, Pankov A, Song JS, Costello JF. Intratumoral Heterogeneity of the Epigenome. Cancer Cell. 2016; 29(4):440–51. https://doi.org/10.1016/j.ccell.2016.03.009 PMID: 27070699
21. Karaayvaz M, Cristea S, Gillespie SM, Patel AP, Mylvaganam R, Luo CC, et al. Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. Nat Commun. 2018; 9(1):3588. https://doi.org/10.1038/s41467-018-06052-0
22. Kim KT, Lee HW, Lee HO, Kim SC, Seo YJ, Chung W, et al. Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. Genome Biol. 2015; 16:127. https://doi.org/10.1186/s13059-015-0692-3
23. Levitin HM, Yuan J, Sims PA. Single-Cell Transcriptomic Analysis of Tumor Heterogeneity. Trends Cancer. 2018; 4(4):264–8. https://doi.org/10.1016/j.trecan.2018.02.003 PMID: 29606308
24. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat Commun. 2019; 10(1):380. https://doi.org/10.1038/s41467-018-08023-x PMID: 30670690
25. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015; 12(5):453–7. https://doi.org/10.1038/nmeth.3337 PMID: 25822800
26. Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform. 2017; 18(5):851–69. https://doi.org/10.1093/bib/bbw068 PMID: 27473064
27. Andor N, Harness JV, Muller S, Mewes HW, Petritsch C. EXPANDS: expanding ploidy and allele frequency on nested subpopulations. Bioinformatics. 2014; 30(1):50–60. https://doi.org/10.1093/bioinformatics/btt622 PMID: 24177718
28. Knight Keith W F. asymptotics for lasso-type estimators. the annals of statistics. 2000; 28(5):1356–78.

29. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010; 33(1):1–22.

30. Scholkopf B. comparing support vector machines with gaussian kernels to radial basis function classifiers. 1996.

31. Domingos Pedro P M. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. machine learning. 1997; 29(2–3):103–30.

32. Breiman L. Random Forests. machine learning. 2001; 45(45):5–32.