frontiers

TOOme: a novel computational framework to infer cancer tissue-of-origin by integrating both gene mutation and expression

Binsheng He[1], Jidong Lang[2], Bo Wang[2], Xiaojun Liu[2], Qingqing Lu[2*], Jianjun He[1], Wei Gao[3], Pingping Bing[1], Geng Tian[2], Jialiang Yang[2]

[1]Changsha Medical University, China, [2]Geneis (Beijing) Co. Ltd, China, [3]Fujian Provincial Cancer Hospital, China

Submitted to Journal:
Frontiers in Bioengineering and Biotechnology

Specialty Section:
Bioinformatics and Computational Biology

Article type:
Original Research Article

Manuscript ID:
536281

Received on:
19 Feb 2020

Revised on:
29 Mar 2020

Frontiers website link:
www.frontiersin.org

Conflict of interest statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest

Author contribution statement

JY, GT and PB conceived the concept of the work. BH, XL, BW and JL performed the experiments. BH and XL wrote the paper. QL, WG and JH reviewed the paper. All authors approved the final version of this manuscript.

Keywords

tissue-of-origin, somatic mutation, Gene Expression, random forest, Cross-validation

Abstract

Word count: 343

Metastatic cancers require further diagnosis to determine their primary tumor sites. However, the tissue-of-origin for around 5% tumors could not be identified by routine medical diagnosis according to a statistics in the United States.

With the development of machine learning techniques and the accumulation of big cancer data from TCGA and GEO, it is now feasible to predict cancer tissue-of-origin by computational tools. Metastatic tumor inherits characteristics from its tissue-of-origin, and both gene expression profile and somatic mutation have tissue specificity. Thus, we developed a computational framework to infer tumor tissue-of-origin by integrating both gene mutation and expression (TOOme). Specifically, we first perform feature selection on both gene expressions and mutations by a random forest method. The selected features are then used to build up a multi-label classification model to infer cancer tissue-of-origin. We adopt a few popular multiple-label classification methods , which are compared by the 10-fold cross validation process.

We applied TOOme to the TCGA data containing 7,008 non-metastatic samples across 20 solid tumors. 74 genes by gene expression profile and 6 genes by gene mutation are selected by the random forest process, which can be divided into two categories: (1) cancer type specific genes and (2) those expressed or mutated in several cancers with different levels of expression or mutation rates. Function analysis indicates that the selected genes are significantly enriched in gland development, urogenital system development, hormone metabolic process, thyroid hormone generation prostate hormone generation and so on. According to the multiple-label classification method, random forest performs the best with a 10-fold cross-validation prediction accuracy of 96%. We also use the 19 metastatic samples from TCGA and 256 cancer samples downloaded from GEO as independent testing data, for which TOOme achieves a prediction accuracy of 89%. The cross-validation validation accuracy is better than those using gene expression (i.e., 95%) and gene mutation (53%) alone.

In conclusion, TOOme provides a quick yet accurate alternative to traditional medical methods in inferring cancer tissue-of-origin. In addition, the methods combining somatic mutation and gene expressions outperform those using gene expression or mutation alone.

Contribution to the field

Metastatic cancers require further diagnosis to determine their primary tumor sites. However, the tissue-of-origin for around 5% tumors could not be identified by routine medical diagnosis according to a statistics in the United States. With the development of machine learning techniques and the accumulation of big cancer data from TCGA and GEO, it is now feasible to predict cancer tissue-of-origin by computational tools. Metastatic tumor inherits characteristics from its tissue-of-origin, and both gene expression profile and somatic mutation have tissue specificity. Thus, we developed a computational framework to infer tumor tissue-of-origin by integrating both gene mutation and expression (TOOme). TOOme provides a quick yet accurate alternative to traditional medical methods in inferring cancer tissue-of-origin. In addition, the methods combining somatic mutation and gene expressions outperform those using gene expression or mutation alone.

Ethics statements

Studies involving animal subjects

Generated Statement: No animal studies are presented in this manuscript.

Studies involving human subjects

Generated Statement: No human studies are presented in this manuscript.

Inclusion of identifiable human data

Generated Statement: No potentially identifiable human images or data is presented in this study.

Data availability statement

Generated Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://dcc.icgc.org/releases/release_26/,https://dcc.icgc.org/releases/release_28/.

TOOme: a novel computational framework to infer cancer tissue-of-origin by integrating both gene mutation and expression

Binsheng He[1,#,*], Jidong Lang[2,#], Bo Wang[2], Xiaojun Liu[2], Qingqing Lu[2], Jianjun He[1], Wei Gao[3], Pingping Bing[1,*], Geng Tian[2,*], Jialiang Yang [2,*]

[1] Academician Workstation, Changsha Medical University, Changsha 410219, China
[2] Genies Beijing Co., Ltd., Beijing 100102, China.
[3] Fujian Provincial Cancer Hospital, Fuzhou 350011, China

[#] These authors contributed equaled to this study

***Correspondence:**
Corresponding Author
hbscsmu@163.com
bpping@163.com
tiang@geneis.cn
yangjl@geneis.cn

Keywords: tissue-of-origin; somatic mutation; gene expression; random forest; cross-validation

Abstract

Metastatic cancers require further diagnosis to determine their primary tumor sites. However, the tissue-of-origin for around 5% tumors could not be identified by routine medical diagnosis according to a statistics in the United States.

With the development of machine learning techniques and the accumulation of big cancer data from The Cancer Genome Atlas(TCGA) and Gene Expression Omnibus(GEO), it is now feasible to predict cancer tissue-of-origin by computational tools. Metastatic tumor inherits characteristics from its tissue-of-origin, and both gene expression profile and somatic mutation have tissue specificity. Thus, we developed a computational framework to infer tumor tissue-of-origin by integrating both gene mutation and expression (TOOme). Specifically, we first perform feature selection on both gene expressions and mutations by a random forest method. The selected features are then used to build up a multi-label classification model to infer cancer tissue-of-origin. We adopt a few popular multiple-label classification methods , which are compared by the 10-fold cross validation process.

We applied TOOme to the TCGA data containing 7,008 non-metastatic samples across 20 solid tumors. 74 genes by gene expression profile and 6 genes by gene mutation are selected by the random forest process, which can be divided into two categories: (1) cancer type specific genes and (2) those expressed or mutated in several cancers with different levels of expression or mutation rates. Function analysis indicates that the selected genes are significantly enriched in gland development, urogenital system development, hormone metabolic process, thyroid hormone generation prostate hormone generation and so on. According to the multiple-label classification method, random forest performs the best with a 10-fold cross-validation prediction accuracy of 96%. We also use the 19 metastatic samples from TCGA and 256 cancer samples downloaded from GEO as independent testing data, for which TOOme achieves a prediction accuracy of 89%. The cross-validation validation accuracy is better than those using gene expression (i.e., 95%) and gene mutation (53%) alone.

In conclusion, TOOme provides a quick yet accurate alternative to traditional medical methods in inferring cancer tissue-of-origin. In addition, the methods combining somatic mutation and gene expressions outperform those using gene expression or mutation alone.

Introduction

Metastatic cancer is a common clinical challenge for limited evidence to determine its primary origin. Patients with carcinoma of unknown primary (CUP) account for about 5% of total cancer patients[1]. CUP are usually heterogeneous, and can lead to dilemmas in diagnosing and treatment since the original tumor site is unknown [2]. Clinically, CUP patients are generally treated with non-selective empirical chemotherapy, which usually leads to low survival rates [3]. Thus, identifying cancer tissue-of-origin (TOO) is critical in improving the treatment of cancer patients and extending their surviving time [4-6].

There are several ancillary examinations in CUP identification, among which immunohistochemistry (IHC) is an important one. However, this method relies on the experiences of pathologists and is labor-intensive. As a result, it is inaccurate in most of the times[7-11]. Positron emission tomography (PET) and computed tomography (CT) are also commonly used in the identification of CUP[12-14]. The detection rate of conventional radiological imaging on primary carcinoma reach 20%–27%, and that of PET reach 24%–40% [15]. The detection accuracy of PET/CT is awfully low that it rarely brings help to identify the primary origin. Obstacles in image technology cause much difficulty of effective use of relative Carcinoma image to help tracing cancer tissue origin.

Molecular profiling of tissue-specific genes is also being used in CUP work-up. Quantities of large-scale profiles of different tumors have been used for diagnose. Molecular profiling is as well as or better than IHC, in terms of poorly differentiated or undifferentiated tumors [16]. Therefore, making use of molecular profiling has become a popular way for diagnosis of unknown origin. Comprehensive molecular profiles displayed in The Cancer Genome Atlas (TCGA) including copy number variation, somatic mutation, gene expression, microRNA expression, DNA methylation, and protein expression, are used to identifying human tumor types [17]. By analysis of tumor types from data of methylation and copy number variation, tissue of origin and molecular classification can be revealed [18]. The methylation profile of metastasis in a meningeal melanocytic tumor is similar to that of primary tumor, and it is suggest that particular copy number variations may be associated with metastatic behavior [19]. Methylation and copy number variation are DNA-level molecular profiling, which brought great help to identify tumor origins.

The copy number profile and gain or loss in specific chromosome regions have been researched by hybridization and cytogenetic-based methods [20, 21]. An *IDH1* somatic mutation in genomic profiling was revealed to bring great benefit to the diagnosis of cholangiocarcinoma and trace the primary origin in a malignancy[22]. Marquard *et al.* obtained classification accuracy of 69% and 85% on 6 and 10 primary sited with somatic mutation respectively, based on PM and CN classifier(classifiers with both point mutations and copy number aberrations) with cross-validation[23]. Mutation of tumor-specific enrichment in certain genes, has been utilized to infer tumor localization, and Dietlein & Eschner developed a tool with mutation spectra to infer cancer origins with a prediction specificity of 79% [24, 25]. As a DNA-level molecular profiling, SNP, that is somatic mutation, can be used as a very useful tool to infer the tissue of origins.

A lot of RNA-level gene expression profile have been explored to identify the cancer tissue of origin [26-30]. Erlander et al, have demonstrated that the gene expression value of samples detected in metastatic tumor is similar to that in the original tumor under condition of carcinoma of unknown primary [31]. Centeno et al, developed a hybrid model by integrating expression profiling and immunohistochemistry for

microRNA-based qRT-PCR test on identification of cancer tissue origin, with 85% of the cases correctly identified [32]. Bloom, G et al, utilized artificial neural networks (ANNs) to predict the unknown cancer tissue origin with mean accuracy of 83-88% in different platforms[33].

Numerous researches have utilized molecular profiles, such as copy number variation, somatic mutation, gene expression, and so on for predicting cancer tissue origin. However, the accuracy of prediction was not satisfying. Identifying cancer tissue origin by combining somatic mutation and gene expression profiling on DNA level and RNA level respectively is first proposed in this study. Firstly, we obtained the data of somatic mutation and gene expression profiling from International Cancer Genome Consortium(ICGC) Database. Machine learning methods can help to improve the performance on prediction of cancer tissue origin. We aim to obtain better performance in predicting cancer tissue origin, by the combination of somatic mutation and gene expression profiling, based on random forest. Machine learning algorithm, such as logistic regression can be used to select gene [34]. However, random forest algorithm [35] was chosen as the gene selection algorithm in this study due to its advantage, good robustness and easy to use. Finally, we used random forest algorithm for classification of cancers. Experiment results showed that higher accuracy can be obtained by using the method proposed in this study.

Materials and methods

Gene expression data

Gene expression profile was downloaded from ICGC Database version release-26 (https://dcc.icgc.org/releases/release_26/). Each gene is named by Gene Symbol ID. The value of gene expression in each labeled sample is normalized by TPM. After deduplication, samples were extracted for combination with SNP samples.

Somatic mutation data

The somatic mutation data was downloaded from ICGC Database version release-28 (https://dcc.icgc.org/releases/release_28/). Each gene is named by Ensembl Gene ID. For Gene Symbol ID is most widely used in paper, the Ensembl Gene ID of gene name in somatic mutation data was converted to Gene Symbol ID. The samples are deduplicated according to information of ICGC-donor-ID, chromosome, and locus in chromosome and gene-affected. Each sample was labeled by its type of cancer.

Data combination

The gene expression and somatic mutation data were merged into one feature matrix. For labeled samples with gene expression array data only involves in 21 cancer types, and samples with Skin Cutaneous Melanoma(SKCM) were removed for it contributes to the major metastasis cancers. The sample with somatic mutation data whose label was not included in these 20 cancer types was removed. Then, the shared sample data was chosen, therefore the samples data after filtering is obtained from 20 different cancer types. An M*N matrix was generated, where M and N represents the number of sample and gene respectively.

Gene selection

Because gene sequencing and mutation detection are costly and time consuming, a scale reduction of gene number is necessary. There are many feature selection algorithms, like Lasso, PCA [36, 37] and etc. The Random forest [35, 38] was a supervised learning algorithm, which is an ensemble learning algorithm based on decision tree and was used to select genes. Best performance was obtained by using 80 selected genes. $\sqrt{n}$ genes were used in a tree, where n represents the number of genes. At the process of splitting node, Gini index was used, which is calculated by formula:

$$\mathrm{Gini}(\mathrm{p}) = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} {p_k}^2 \qquad (1)$$

Where p represents the weight referring to frequencies of cancers in a node, k represents the number of cancers and p_k represents the weight of the kth cancer. The variable importance measures of ith gene in node m, that is the Gini index variation after splitting of node m, is calculated by formula:

$$VIM_{im}^{(Gini)} = GI_m - GI_l - GI_r \qquad (2)$$

Where m is a node in M, which is a set of nodes, $VIM_{im}^{(Gini)}$ represents variable importance measures of ith gene in node m, the GI_m represents the Gini index before splitting, GI_l and GI_r represents the Gini index of two new node after splitting respectively. The importance of the ith gene , in the tth tree is calculated by formula:

$$VIM_{ti}^{(Gini)} = \sum_{m \in M} VIM_{im}^{(Gini)} \qquad (3)$$

Where $VIM_{ti}^{(Gini)}$ represents the importance of the ith gene in the tth tree. If the set of trees is T, the importance of the ith gene in all the tree is calculated by formula:

$$VIM_i^{(Gini)} = \sum_{t=1}^{T} VIM_{ti}^{(Gini)} \quad (4)$$

Where $VIM_i^{(Gini)}$ is the importance of the ith gene in all trees. We sorted the importance scores of all genes, then the top H genes were selected, where H is the variable number of genes that can be set to find the best result.

Multi-classifier Random Forest

The random forest is actually a special method of bagging that using the decision tree as a model in bagging[38, 39]. First, the bootstrap method is used to generate m training sets, which is a set of samples. Then, each training set is used to construct a tree. $\sqrt{n}$ genes are used in a tree, where n represents the number of selected genes. When splitting a node, not all the genes are used to optimize the metric Gini index used in this study, a part of genes is randomly extracted instead. An optimal solution can be found among the extracted genes, and applied to node splitting. Leaf node in the tree records which gene is used to determine the cancer type, and each leaf node represents the last judged cancer type. The predicted cancer type is given by maximum votes from decision tree.

Statistical Analysis

The metric of precision, recall and F1 score were used to evaluate the performance of the model. True-positive, false-positive, true-negative and false-negative are abbreviated as TP, FP, TN and FN respectively. Precision is calculated by $(\mathrm{TP})/(\mathrm{TP} + \mathrm{FP})$, which indicates the ability of classifier to differentiate positive from negative cases. Recall is calculated by $(\mathrm{TP})/(\mathrm{TP} + \mathrm{FN})$, which indicates the ability of classifier to recognize all positive cases. The $F1$ score is calculated by $(2 * \mathrm{recall} * \mathrm{precision})/(\mathrm{recall} + \mathrm{precision})$. Each individual cancer type is calculated by these metrics, and the cohort metric adopt the mean report. The entire cohort is calculated by accuracy, reported as $(\mathrm{TP} + \mathrm{TN})/(\mathrm{total\ cases})$. 10 times 10-fold cross validation is used to obtain the metric report, whose average is treated as the result metric.

Gene annotation

The functions annotation of specific gene set was given. Geno ontology [40, 41] was used as enrichment analysis database. Gene clustering and visualization was realized by R package cluaterProfiler and gogadget[42, 43].

Results

The workflow of TOOme

The complete workflow of prediction on cancer tissue origin is shown in Fig 1. The process can be split into three steps. At the first step, we download the raw data from ICGC Database, and extracted the effective information to obtain preliminary data of somatic mutation and gene expression profiling. At the second step, we filtered the data of somatic mutation and gene expression profiling respectively. Then, samples with both somatic mutation data and gene expression proofing were used to form feature matrix. As a result, the generated feature matrix was used for gene selection. At the third step, most of the samples were utilized to train the model with 10-time 10 folds cross validation by using random forest classification algorithm. We carried out numerous experiments to evaluate the performance of the proposed method.

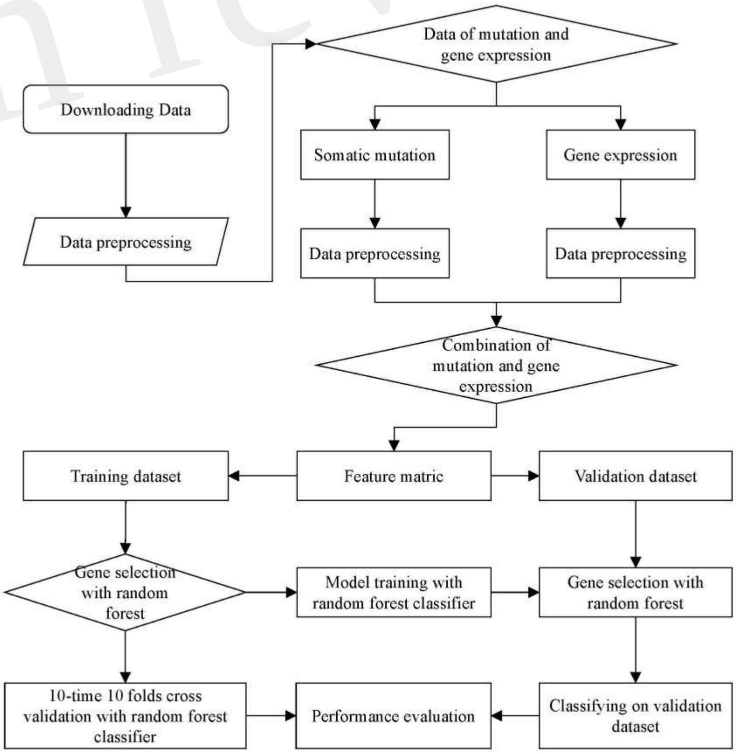
Downloading Data
Data preprocessing
Data of mutation and gene expression
Somatic mutation
Gene expression
Data preprocessing
Data preprocessing
Combination of mutation and gene expression
Training dataset
Feature matric
Validation dataset
Gene selection with random forest
Model training with random forest classifier
Gene selection with random forest
10-time 10 folds cross validation with random forest classifier
Performance evaluation
Classifying on validation dataset

Fig 1. The complete workflow of prediction on cancer tissue origin.

Data used in this study

We used ICGC version 26 and 28 databases, with Gene expression profile and somatic mutation information to classify tumor samples. The allele mutation in somatic mutation data can be A/G, C/T, C/A and etc. For it is hard to distinguish mutation types with limited relative information and tools, we consider all kinds of allele mutation as gene mutation and count the number of gene mutation of each sample. Different from somatic mutation data, Gene expression profile array data is directly used. The sample distribution of each cancer is showed in Table 1, where samples suffer from BRCA are much more than from other cancers. Considerable prediction results can be obtained by our model. The precision, recall and F_1 score, showed in Table 2, reach 99.86%, 99.47% and 99.67% respectively.

In this study, there are 371 samples with metastasis, where 352 samples are SKCM. To avoid unbalanced distribution of samples, we removed all the SKCM samples with metastasis. Only 19 samples with metastasis were used as test dataset.

Table 1. Sample distribution of each cancer from ICGC database.

Available Cancer Types	Abbreviation	samples	
		Amount	Percentage
Bladder urothelial carcinoma	BLCA	294	4.20%
Breast invasive carcinoma	BRCA	970	13.84%
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	241	3.44%
Colon adenocarcinoma	COAD	390	5.57%
Glioblastoma multiforme	GBM	148	2.11%
Head and Neck squamous cell carcinoma	HNSC	460	6.56%
Kidney renal clear cell carcinoma	KIRC	345	4.92%
Kidney renal papillary cell carcinoma	KIRP	216	3.08%
Acute Myeloid Leukemia	LAML	121	1.73%
Brain lower grade glioma	LGG	433	6.18%
Liver hepatocellular carcinoma	LIHC	282	4.02%
Lung adenocarcinoma	LUAD	475	6.78%
Lung squamous cell carcinoma	LUSC	411	5.87%
Ovarian serous cystadenocarcinoma	OV	185	2.64%
Pancreatic adenocarcinoma	PAAD	134	1.91%
Prostate adenocarcinoma	PRAD	374	5.34%
Rectum adenocarcinoma	READ	137	1.95%
Stomach adenocarcinoma	STAD	412	5.88%
Thyroid carcinoma	THCA	486	6.93%
Uterine corpus endometrial carcinoma	UCEC	494	7.05%
Total		7008	100%

Performance evaluation

The classification accuracies obtained by using data of somatic mutation, gene expression profiling and both of them, under condition of using different number of genes, have been compared in Fig 2. Motivated by Ma, Patel et al that 5 genes can be used to solve a 32-type classification problem[44], 5 was chosen as the minimum number of genes. For gene sequencing and mutation detection are costly and time consuming, 120 was chosen as the maximum number of genes. A lot of experiments have been done using the prepared data between the interval from 5 to 120. For using small number of genes did not obtain satisfying classification performance, the interval between number of genes was set to 10 or even larger until the number of genes equals to 50. Then the interval was set to 5 for fine tuning, based on small fluctuation by changed number of genes.

Results with 10-time 10 folds cross validation on training dataset are shown in Fig 2 that accuracy of using data of both somatic mutation and gene expression profiling is always higher than that of only using one of it. The best result of them are 95.77%, 53.51% and 89.28%, obtained by using 80, 120 and 105 genes respectively. Results shows that gene expression can make much contribution to obtain higher accuracy than data of somatic mutation. However, a combination of them achieved best classification performance.

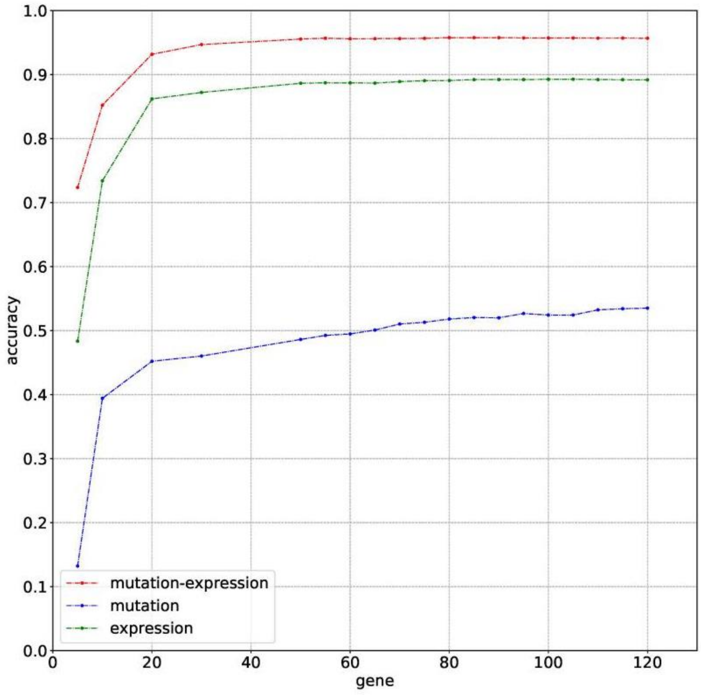

Fig 2. The classification accuracy of using somatic mutation, gene expression and combination of somatic mutation and gene expression respectively.

As for the test dataset, we conducted experiments by using the chosen 80 genes in training model. The overall classification accuracy is 89.47%. Table 3 shows the prediction probabilities of each sample on each cancer. The value on the table highlighted by color of green, yellow and pink presents high, middle and low probabilities respectively of predicting a sample to a cancer type. We obtained considerable prediction accuracy on sample with BRCA and THCA. Each sample was correctly predicted to the same as the true label. A sample whose true label is CESC was predicted to UCEC. A sample whose true label is BRCA was predicted to LGG with a terrible probability 1.65%. In this condition, we considered that little error on classification is tolerable.

Table 2. Performance of classification of combination of somatic mutation and gene expression by using 80 genes.

Cancer Type	Precision	Recall	F1-score	Support	Specificity
BLCA	0.8906	0.9354	0.9124	294.0000	0.9950
BRCA	0.9987	0.9947	0.9967	970.0000	0.9998
CESC	0.9148	0.8859	0.9001	241.0000	0.9971
COAD	0.7548	0.9644	0.8468	390.0000	0.9815
GBM	0.9940	1.0000	0.9970	148.0000	0.9999
HNSC	0.9916	1.0000	0.9958	460.0000	0.9994
KIRC	0.9850	0.9516	0.9680	345.0000	0.9992
KIRP	0.9344	0.9630	0.9485	216.0000	0.9979
LAML	1.0000	1.0000	1.0000	121.0000	1.0000
LGG	0.9926	0.9977	0.9952	433.0000	0.9995
LIHC	0.9925	0.9844	0.9884	282.0000	0.9997
LUAD	0.9358	0.9448	0.9403	475.0000	0.9953
LUSC	0.9408	0.9000	0.9199	411.0000	0.9965
OV	1.0000	0.9946	0.9973	185.0000	1.0000
PAAD	0.9378	0.9552	0.9464	134.0000	0.9988
PRAD	0.9973	1.0000	0.9987	374.0000	0.9998
READ	0.7569	0.1591	0.2627	137.0000	0.9990
STAD	0.9947	0.9976	0.9961	412.0000	0.9997
THCA	1.0000	0.9979	0.9990	486.0000	1.0000
UCEC	0.9673	0.9816	0.9744	494.0000	0.9975
Accuracy	0.9577	0.9577	0.9577	0.0000	

Table 3. Prediction probabilities of each samples on each cancer.

Cancer	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
BLCA	0.0005	0.0015	0.0005	0	0.1825	0.162	0.0665	0.0155	0.002	0.001	0.034	0	0	0	0	0.0015	0.0005	0	0
BRCA	0.993	0.9675	0.9995	0.999	0.6375	0.1195	0.045	0.066	0.0015	0.0005	0.0085	0.001	0.0005	0	0	0	0	0	0
CESC	0.0005	0.004	0	0	0.047	0.101	0.8	0.086	0.0275	0.002	0.1115	0	0	0	0	0.0015	0	0	0.001
COAD	0	0.001	0	0.0005	0.005	0.01	0.008	0.002	0.7015	0.001	0.009	0	0	0	0	0.001	0	0	0
GBM	0	0	0	0	0.001	0.0035	0	0	0	0	0.001	0	0	0	0	0	0.0005	0	0
HNSC	0.0005	0	0	0	0.0065	0.011	0.0055	0.0015	0	0.993	0.754	0	0	0	0	0	0	0	0.001
KIRC	0	0	0	0	0.0015	0.0535	0.001	0.003	0.0005	0	0.001	0	0.0005	0	0	0.0015	0.001	0	0
KIRP	0	0	0	0	0.004	0.038	0.001	0.0045	0.0005	0	0	0	0	0	0	0.0005	0.0015	0	0
LAML	0	0.006	0	0	0.0155	0.0055	0	0.005	0.001	0	0.0005	0	0	0	0	0	0	0	0
LGG	0	0	0	0	0.0125	0.165	0.0055	0.01	0.0005	0.0005	0.0035	0	0	0	0	0.001	0	0	0.0005
LIHC	0	0.0005	0	0	0.003	0.0365	0.0045	0.0045	0.0095	0	0.001	0	0	0	0	0	0	0	0
LUAD	0.0025	0.006	0	0	0.011	0.0225	0.009	0.012	0.001	0	0.0055	0.0065	0	0	0	0.0025	0.001	0.001	0.001
LUSC	0.001	0.008	0	0.0005	0.017	0.0735	0.0375	0.008	0	0	0.024	0.001	0.0005	0	0	0.0015	0.0005	0.0005	0.002
OV	0	0	0	0	0.002	0.0005	0	0.001	0	0	0	0	0	0	0	0	0.002	0	0
PAAD	0	0.0005	0	0	0.0095	0.0775	0.004	0.0045	0.0075	0	0.001	0	0	0	0	0.0005	0	0	0
PRAD	0	0.0005	0	0	0.003	0.004	0.002	0.001	0	0	0.0005	0	0	0	0	0	0	0	0.001
READ	0	0.002	0	0	0.0005	0.001	0.003	0.0005	0.242	0.0005	0.0065	0	0	0	0	0	0	0	0
STAD	0	0	0	0	0.0055	0.0025	0.0005	0.0005	0.0045	0	0.004	0	0	0	0	0	0	0	0
THCA	0	0	0	0	0.0015	0.0035	0	0.0065	0	0	0.0005	0.991	0.9985	1	1	0.9875	0.9925	0.9985	0.992
UCEC	0.002	0.0025	0	0	0.034	0.1095	0.007	0.768	0.0005	0.0015	0.034	0.0005	0	0	0	0.001	0.0005	0	0.0015
LOW_CONFIDENCE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
predicted_label	BRCA	BRCA	BRCA	BRCA	BRCA	LGG	CESC	UCEC	COAD	HNSC	HNSC	THCA	THCA	THCA	THCA	THCA	THCA	THCA	THCA
true_label	BRCA	BRCA	BRCA	BRCA	BRCA	BRCA	CESC	CESC	COAD	HNSC	HNSC	THCA	THCA	THCA	THCA	THCA	THCA	THCA	THCA
correct	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1

Mean value of gene expression and somatic mutations on each cancer

We plotted the heatmap of mean value of gene expression and somatic mutations on each cancer. In Fig 3, the rows represent 74 genes of gene expression and columns denote the cancers. In Fig 4, the rows represent 6 genes of somatic mutation and columns represent the cancers. The mean value of gene expression and somatic mutation on a logarithmic scale was plotted with relative color. A color bar was used to display the value difference. Cancers that fell into cluster at horizontal axis had a similar value between gene expression or mutation number. The genes were also clustered at vertical axis based on the similarity between cancers.

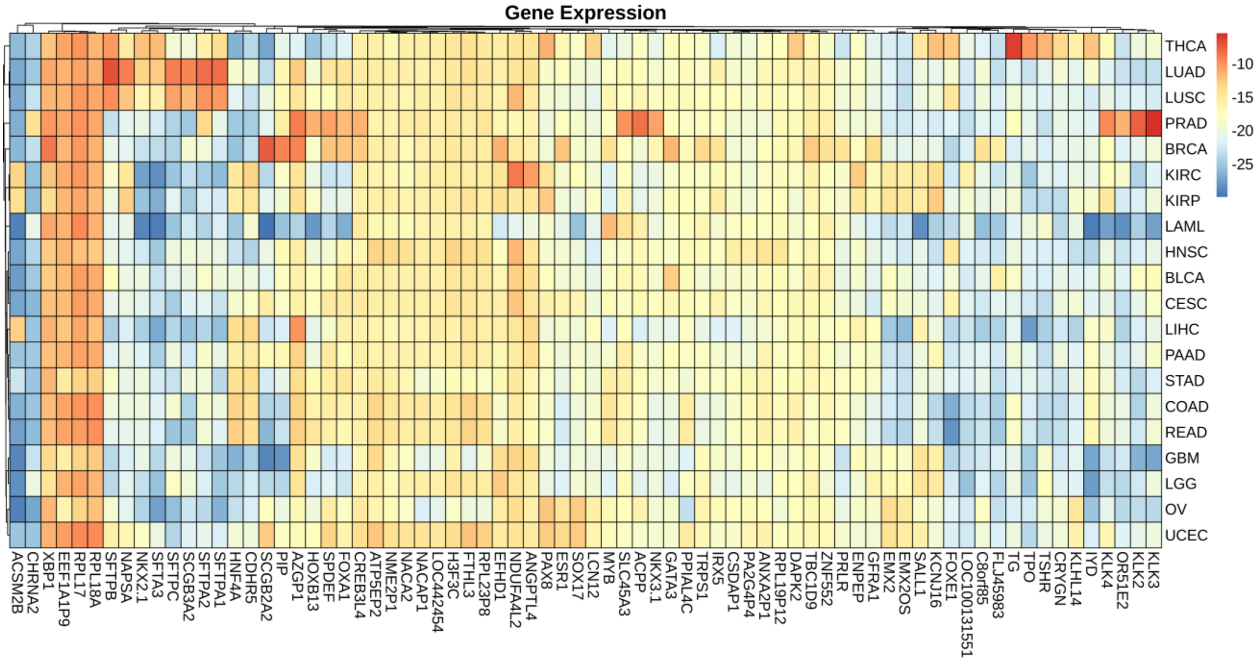

Fig 3. Heatmap of mean value of gene expression on each cancer.

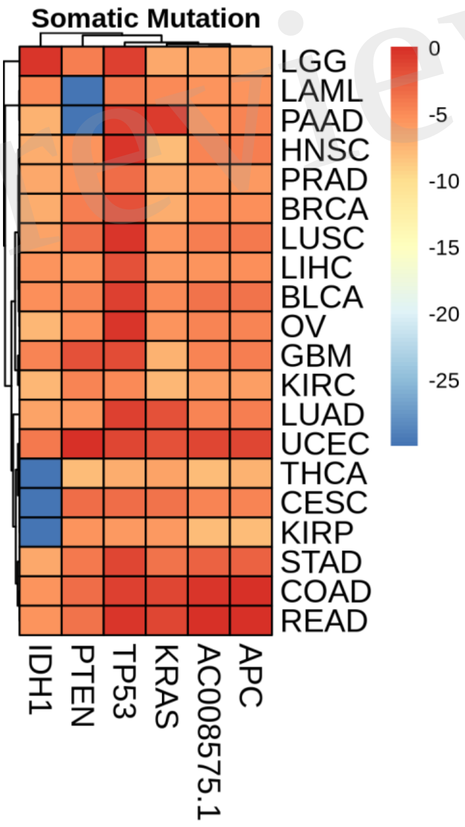

Fig 4. Heatmap of mean value of somatic mutations on each cancer.

Discussion

Data of somatic mutation and gene expression profiling can be used to identify the primary site of tumors. However, it was the first time to identify the cancer tissue origin by using both data of somatic mutation and gene expression profiling. We carried out experiments by using 7008 samples with combination of data of somatic and gene expression profiling among 20 cancers. By comparing the performance of them, we obtained highest accuracy by leveraging both of the data of somatic mutation and gene expression profiling.

The primary analysis tool we used was random forest [35, 38], a machine learning algorithm that can be used for gene selection and tumor classification. We chose top-rank 80 genes, where 6 genes and 74 genes are corresponding to mutation and expression respectively, for classification. Therefore, it showed that data of somatic mutation performs worse than gene expression profiling on prediction of cancer tissue origin. Our method obtained 96% overall accuracy on the training dataset. The performance is maintained considerably on the external cohorts, and the overall accuracy on sample with metastatic disease is 89%. Our model cannot provide good performance on physiologically proximal cancers, such as uterine corpus endometrial carcinoma and cervical squamous cell carcinoma and endocervical adenocarcinoma. The endometrial and ovarian endometrioid carcinomas evolve from similar precursor endometrial epithelial cells; many researches are involved in the molecular pathogenesis of the endometrial and ovarian endometrioid carcinomas[45].

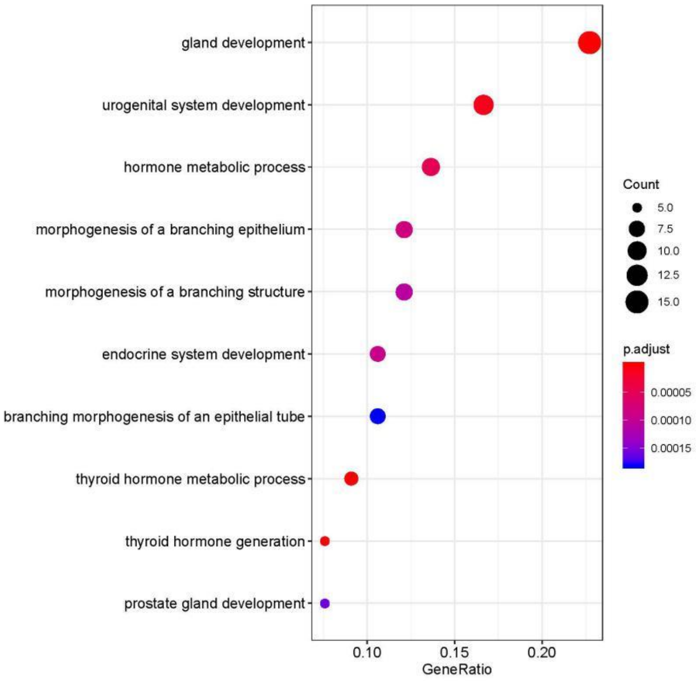

Fig 5. Selected top-rank 80 genes enriched in cellular component, biological process and molecular function.

We studied the role that gene plays in cellular component, biological process and molecular function. Fig 5 shows the top-rank 80 genes selected by random forest algorithm. The selected genes were enriched in hormone metabolic process, tissue and organ development and hormone-mediated signaling pathway, specifically in gland development, urogenital system development, hormone metabolic process, morphogenesis of a branching epithelium, morphogenesis of a branching structure, endocrine system development, branching morphogenesis of an epithelial tube， thyroid hormone metabolic process, thyroid hormone generation and prostate gland development. For example, *APC* plays a significant role in discovering pathogenesis of soft tissue tumors[46]. Birnbaum et al investigated what role the *APC* gene play in colorectal cancer, at the investigation of 183 colon adenocarcinomas, point mutations were found in 73% of cases[47]. We obtained the similar conclusion that mutation of *APC* gene may be the important impact of colorectal cancer, as heatmap shown in Fig 4 that the mean number of *APC* gene mutation in colorectal cancer is more than that in other cancers except rectum adenocarcinoma. It can be explained that they are two physiologically proximal cancers. Mutation in *IDH1* gene can reduce cell survival, proliferation and invasion of human glioma [48]. Mutation in *IDH1* gene is an oncogenic driver in a majority of lower-grade gliomas and have an impact on brain lower grade glioma with different genetic pathway [49-51]. The same conclusion was acquired in Fig 4 that the mean number of *IDH1* gene mutation in Brain lower grade glioma is more than that in other cancers.

ACPP gene plays a vital key in prostate adenocarcinoma [52-54]. From the heatmap, it is clear that the level of *ACPP* gene expression in prostate adenocarcinoma is higher than that in other cancers. The expression levels of TG were found to be altered in all kinds of thyroid carcinomas [55]. From Fig 3, we obtained similar results that the level of *TG* gene expression in thyroid carcinomas is higher than that in other cancers.

Molecular profiling of tissue-specific genes can be utilized to identify the primary site of tumor. Combination of data of somatic mutation and gene expression profiling were first proposed in this study to predict the primary origin. We obtained considerable prediction performance, and therefore this research can bring great help to the identification of cancer tissue origin. However, we did not carry out research to discover the relationship between data of gene expression and somatic mutation. Our method cannot classify physiologically proximal cancers yet. And it is also a future work to employing other machine learning algorithms that can improve the classification performance.

Conclusion

Identification of cancer tissue origin is a challenging work recently and in the future. With a lot of molecular profiling available, we can make use of them alone and combine some of them to improve performance of identification primary site of tumor. Machine learning algorithm is also an effective tool to help classifying the cancers. The prediction performance can be tremendously affected by the number of features used.

In this study, we used both molecular data of somatic mutation and gene expression profiling to generate a feature matrix. Then the optimal number of genes was obtained and the data was trained, based on random forest algorithm. The performance of using our method was also compared to only by using data of somatic mutation or gene expression profiling. Our method achieved highest accuracy. Experiment results shows that our method can be an effective tool for primary origin tracing.

Conflict of Interest

Author BW, JL, XL, QL, GT and JY were employed by the company Geneis Beijing Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest

Authors' contributions

JY, GT and PB conceived the concept of the work. BH, XL, BW and JL performed the experiments. BH and XL wrote the paper. QL, WG and JH reviewed the paper. All authors approved the final version of this manuscript.

Acknowledgement

This research was funded by Hunan Provincial Innovation Platform and Talents Program (No. 2018RS3105), the Natural Science Foundation of China (No. 61803151), the Natural Science Foundation of Hunan province (No. 2018JJ3570), and the Project of Scientific Research Fund of Hunan Provincial Education Department (No 19A060 and 19C0185).

Reference

1. Shaw, P.H.S., et al., *A Clinical Review of the Investigation and Management of Carcinoma of Unknown Primary in a Single Cancer Network.* Clinical Oncology, 2007. **19**(1): p. 87-95.
2. Rizwan, M. and M. Zulfiqar, *Carcinoma of unknown primary.* JPMA. The Journal of the Pakistan Medical Association, 2010. **60**(1): p. 598-9.
3. Kurahashi, I., et al., *A microarray-based gene expression analysis to identify diagnostic biomarkers for unknown primary cancer.* PLoS One, 2013. **8**(5): p. e63249.
4. Hyphantis, T., et al., *Psychiatric manifestations, personality traits and health-related quality of life in cancer of unknown primary site.* Psychooncology, 2013. **22**(9): p. 2009-15.
5. Hudis, C.A., *Trastuzumab--mechanism of action and use in clinical practice.* N Engl J Med, 2007. **357**(1): p. 39-51.
6. Varadhachary, G.R., et al., *Carcinoma of unknown primary with a colon-cancer profile-changing paradigm and emerging definitions.* Lancet Oncol, 2008. **9**(6): p. 596-9.
7. Janick, S., et al., *Immunohistochemistry for Diagnosis of Metastatic Carcinomas of Unknown Primary Site.* Cancers, 2018. **10**(4): p. 108-110.
8. Kandalaft, P.L. and A.M. Gown, *Practical Applications in Immunohistochemistry: Carcinomas of Unknown Primary Site.* Archives of Pathology & Laboratory Medicine, 2015. **140**(6): p. 508-526.
9. Centeno, B.A., et al., *Hybrid Model Integrating Immunohistochemistry and Expression Profiling for the Classification of Carcinomas of Unknown Primary Site.* Journal of Molecular Diagnostics, 2010. **12**(4): p. 476-486.
10. Voigt, J.J., *Immunohistochemistry: a major progress in the classification of carcinoma of unknown primary.* 2008. **10**(12): p. 693-697.
11. Huebner, G., et al., *503 POSTER Comparative analysis of microarray testing and immunohistochemistry*

in patients with carcinoma of unknown primary – CUP syndrome. 2007. **5**(4): p. 90-91.

12. Fu, Z., et al., *Diagnosis of Primary Clear Cell Carcinoma of the Vagina by 18F-FDG PET/CT.* Clinical Nuclear Medicine, 2019. **44**(4): p. 493-494.
13. Kwee, T.C., et al., *FDG PET/CT in carcinoma of unknown primary.* European Journal of Nuclear Medicine and Molecular Imaging, 2010. **37**(3): p. 635-644.
14. Fencl, P., et al., *Prognostic and diagnostic accuracy of [18F]FDG-PET/CT in 190 patients with carcinoma of unknown primary.* European Journal of Nuclear Medicine and Molecular Imaging, 2007. **34**(11): p. 1783-1792.
15. Ambrosini, V., et al., *18F-FDG PET/CT in the assessment of carcinoma of unknown primary origin.* La Radiologia medica, 2006. **111**(8): p. 1146-1155.
16. Oien, K.A. and J.L. Dennis, *Diagnostic work-up of carcinoma of unknown primary: from immunohistochemistry to molecular profiling.* Annals of Oncology, 2012. **23**(suppl_10): p. x271-x277.
17. Li, Y., et al., *A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data.* BMC Genomics, 2017. **18**(1): p. 508-512.
18. Hoadley, K.A., et al., *Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin.* Cell, 2014. **158**(4): p. 929-944.
19. Küsters-Vandevelde, H.V.N., et al., *Copy number variation analysis and methylome profiling of a GNAQ-mutant primary meningeal melanocytic tumor and its liver metastasis.* Experimental & Molecular Pathology, 2017. **102**(1): p. 25-31.
20. Beroukhim, R., et al., *Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(50): p. 20007-20012.
21. Baudis, M., *Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data.* 2007. **7**(1): p. 226-0.
22. Sheffield, B.S., et al., *Personalized oncogenomics in the management of gastrointestinal carcinomas-early experiences from a pilot study.* Current Oncology, 2016. **23**(6): p. 68-73.
23. Marquard, A.M., et al., *TumorTracer: a method to identify the tissue of origin from the somatic mutations of a tumor specimen.* Bmc Medical Genomics, 2016. **8**(1): p. 58-59.
24. Lawrence, M.S., et al., *Discovery and saturation analysis of cancer genes across 21 tumour types.* Nature, 2014. **505**(7484): p. 495-501.
25. Dietlein, F. and W. Eschner, *Inferring primary tumor sites from mutation spectra: a meta-analysis of histology-specific aberrations in cancer-derived cell lines.* Human Molecular Genetics, 2014. **23**(6): p. 1527-1537.
26. Qu, K.Z., et al., *Molecular identification of carcinoma of unknown primary (CUP) with gene expression profiling.* Journal of Clinical Oncology, 2007.
27. Erlander, M.G., et al., *Molecular classification of carcinoma of unknown primary by gene expression profiling from formalin-fixed paraffin-embedded tissues.* 2004. **22**(14_suppl): p. 9545.
28. Hainsworth, J.D., et al., *Molecular Gene Expression Profiling to Predict the Tissue of Origin and Direct Site-Specific Therapy in Patients With Carcinoma of Unknown Primary Site: A Prospective Trial of the Sarah Cannon Research Institute.* Journal of Clinical Oncology, 2013. **31**(2): p. 217-223.
29. Gross-Goupil, M., et al., *Identifying the Primary Site Using Gene Expression Profiling in Patients with Carcinoma of an Unknown Primary (CUP): A Feasibility Study from the GEFCAPI.* Onkologie, 2012. **35**(1-2): p. 54-55.
30. Greco and F. A., *Cancer of Unknown Primary or Unrecognized Adnexal Skin Primary Carcinoma? Limitations of Gene Expression Profiling Diagnosis.* Journal of Clinical Oncology, 2013. **31**(11): p.

1479-1481.

31. Erlander, M.G., et al., *Performance and Clinical Evaluation of the 92-Gene Real-Time PCR Assay for Tumor Classification.* Journal of Molecular Diagnostics Jmd, 2011. **13**(5): p. 493-503.

32. Rosenwald, S., et al., *Validation of a microRNA-based qRT-PCR test for accurate identification of tumor tissue origin.* Mod Pathol, 2010. **23**(6): p. 814-23.

33. Bloom, G., et al., *Multi-platform, multi-site, microarray-based human tumor classification.* Am J Pathol, 2004. **164**(1): p. 9-16.

34. Kao, K.J., S.H. Cheng, and A.T. Huang, *Gene expression profiling for prediction of distant metastasis and survival in primary nasopharyngeal carcinoma.* 2006. **24**(18_suppl).

35. Sandri, M. and P. Zuccolotto. *Variable Selection Using Random Forests*. in *Data Analysis, Classification and the Forward Search*. 2006. Berlin, Heidelberg: Springer Berlin Heidelberg.

36. R, M. and R. Rohini, *LASSO: A feature selection technique in predictive modeling for machine learning*. 2016, 2016 IEEE International Conference on Advances in Computer Applications (ICACA): Rohini, R.

37. Malhi, A. and R. Gao, *PCA-Based Feature Selection Scheme for Machine Defect Classification.* Instrumentation and Measurement, IEEE Transactions on, 2005. **53**(26): p. 1517-1525.

38. Breiman, L., *Random Forests.* Machine Learning, 2001. **45**(1): p. 5-32.

39. Meyer, J.G., et al., *Learning Drug Function from Chemical Structure with Convolutional Neural Networks and Random Forests.* Journal of Chemical Information and Modeling, 2019.

40. Waardenberg, A.J., et al., *Erratum to: 'CompGO: an R package for comparing and visualizing Gene Ontology enrichment differences between DNA binding experiments'.* BMC Bioinformatics, 2016. **17**(1): p. 179-185.

41. Ye, J., et al., *WEGO: a web tool for plotting GO annotations.* Nucleic acids research, 2006. **34**(12): p. 293-312.

42. Nota, B., *Gogadget: An R Package for Interpretation and Visualization of GO Enrichment Results.* Molecular informatics, 2016. **36**.

43. Yu, G., et al., *clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters.* Omics : a journal of integrative biology, 2012. **16**: p. 284-7.

44. Ma, X.J., et al., *Molecular classification of human cancers using a 92-gene real-time quantitative polymerase chain reaction assay.* Arch Pathol Lab Med, 2006. **130**(4): p. 465-73.

45. Melissa K McConechy, J.D., Janine Senz, Winnie Yang, Nataliya Melnyk, Alicia A Tone, Leah M Prentice, Kimberly C Wiegand, Jessica N McAlpine, Sohrab P Shah, Cheng-Han Lee, Paul J Goodfellow, C Blake Gilks & David G Huntsman *Ovarian and endometrial endometrioid carcinomas have distinct CTNNB1 and PTEN mutation profiles.* Modern Pathology, 2014. **27**(1): p. 128-134.

46. Kuhnen, C., et al., *APC and β-catenin in alveolar soft part sarcoma (ASPS) - immunohistochemical and molecular genetic analysis.* Pathology - Research and Practice, 2000. **196**(5): p. 0-304.

47. Birnbaum, D.J., et al., *Expression Profiles in Stage II Colon Cancer According to APC Gene Status.* Translational Oncology, 2012. **5**(2): p. 72-76.

48. Cui, D., et al., *R132H mutation in IDH1 gene reduces proliferation, cell survival and invasion of human glioma by downregulating Wnt/β-catenin signaling.* The International Journal of Biochemistry & Cell Biology, 2016. **73**(2): p. 72-81.

49. Pieper, R.O., S. Ohba, and J. Mukherjee, *MUTANT IDH1-DRIVEN CELLULAR TRANSFORMATION INCREASES RAD51-MEDIATED HOMOLOGOUS RECOMBINATION AND TEMOZOLOMIDE (TMZ) RESISTANCE.* Cancer research, 2014. **74**(17): p. 4836-44.

50. Ohno, M., et al., *Secondary glioblastomas with IDH1/2 mutations have longer glioma history from preceding lower-grade gliomas.* Brain Tumor Pathology, 2013. **30**(4): p. 224-232.

51. Ohka, F., et al., *A novel all-in-one intraoperative genotyping system forIDH1-mutant glioma.* Brain Tumor Pathology, 2017. **34**(2): p. 91-97.

52. Vihko, P.T., et al., *Prostatic acid phosphatase (PAP) is PI(3)P-phosphatase and its inactivation leads to change of cell polarity and invasive prostate cancer.* Cancer Research, 2005. **65**(10): p. 62-78.

53. Maatman, T.J., M.K. Gupta, and J.E. Montie, *The Role of Serum Prostatic Acid Phosphatase as a Tumor Marker in Men with Advanced Adenocarcinoma of the Prostate.* Journal of Urology, 1984. **132**(1): p. 58-60.

54. Drago, J.R., et al., *Relative value of prostate-specific antigen and prosttic acid phosphatase in diagnosis and management of adenocarcinoma of prostate Ohio State University Experience.* Urology, 1989. **34**(4): p. 187-192.

55. Makhlouf, A.M., et al., *Identification of CHEK1, SLC26A4, c-KIT, TPO and TG as new biomarkers for human follicular thyroid carcinoma.* Oncotarget, 2016. **7**(29): p. 45776-45788.