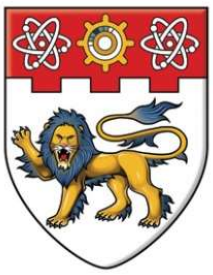


NANYANG TECHNOLOGICAL UNIVERSITY

Gene Expression Signatures of Hepatocellular Carcinoma Circulating Tumor Cells vary with Site of Dissemination

(BS6202: TECHNIQUES IN BIOMEDICAL DATA MINING)

GROUP 4

ZHANG ZHU

ABEL TAN

YE SHUCHEN

HAN WENHAO

LI GUANCHEN

Content

1. Introduction

Hepatocellular carcinoma (HCC) is the main subtype of liver cancer, which leads to high mortality worldwide[1]. One of the obstacles to combating it is hematogenous dissemination, namely the major route of HCC metastasis[2]. Multiple researches have revealed that circulating tumor cells (CTC) play an important role in HCC metastasis[3]. Therefore, the underlying mechanisms are urgently needed to improve the patients' cure rate and living quality. Here we obtained a gene expression dataset from a published study[4]. The dataset contains transcriptomic information of 113 single circulating tumor cells (CTC) from 4 different vascular sites, including the hepatic vein (HV), peripheral artery (PA), peripheral vein (PV), and portal vein (PoV).

The gist of this project is to find the genes differentially expressed along with the metastasis pathway and interrogate their function which equips CTC capability to overcome the complex microenvironment and finally colonize successfully.

2. Materials and Methodology

The workflow is summarized in Figure1.

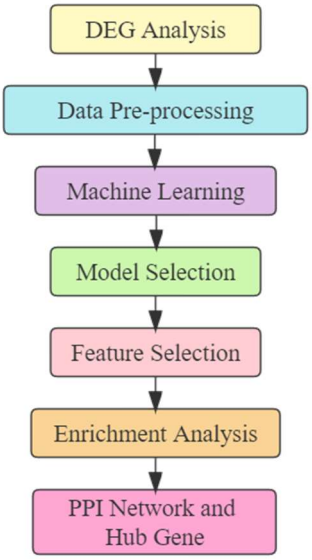


Figure 1. Workflow of project

2.1. Data set

As aforementioned, the dataset contains 113 samples from 4 vascular sites. The total number of the gene is 13013. Sample sizes in each location are 45 (HV), 12 (PA), 40 (PV), and 16 (PoV) respectively.

The expression missing gene name was removed, and the missing expression was filled with the mean of this gene. The total number of the genes is 13005 after data cleaning.

2.2. Differentially expressed gene analysis

4 methods were applied in DEG analysis, encompassing Deseq2, limma, and edgeR, the Wilcoxon sum rank test with quantile normalization, and Wilcoxon sum rank test with GFS normalization. DEseq2 perform its function individually, whereas the popular edgeR package is used to import, organize, filter and normalize the data, followed by the limma package with its voom method, linear modelling and empirical Bayes moderation to assess differential expressed gene. The purpose of utilizing different methods is to investigate whether the results are reproducible among these methods and the underlying reasons. To obtain the comparable results, the thresholds of each method are adjustable in a rational scope to ensure the number of DEG is around 100.

The threshold for each method (ordered in HV-PA, PA-PV and PV-PoV):

DEseq2: $p < 0.01$, abs (log2 fold change) >= 0.95; $p < 0.005$, abs (log2 fold change) >= 1.15; $p < 0.01$, abs (log2 fold change) >= 0.9.

Limma and edgeR: $p<0.01$ and abs (log2 fold change) > mean (log fold change) + 3*(standard deviation of log2 fold change)

Wilcoxon rank sum test with quantile normalization: P value <0.0005, P value < 0.0005 and P value <0.005

Wilcoxon rank sum test with GFS normalization: top 100 genes with smallest P value.

2.3. Machine Learning algorithms

For 4 different vascular sites, samples are labeled with 0, 1, 2 and 3 and be grouped into 3 pairs of neighbors that are directly connected. Then, binary classification was performed by using the five models, including decision tree, random forest, SVC (Support Vector Classifier), SVC-RFECV (Recursive Feature Elimination with Cross Validation), and XGBoost (Extreme Gradient Boosting). Before training, z-score normalization was implemented to convert the data into the same scope.

1) Dimension reduction and visualization

Principal Component Analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) are used to visualize the distribution of samples.

2) Tree-based model

The decision tree model is a popular supervised learning model which is usually applied in data mining tasks. In a decision tree structure, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. The Gini impurity and information gain (entropy and log loss), are the evaluation metrics for the decision tree. It can handle numerical data well.

The decision tree can be very non-robust. A small change in the training data can result in a large change in the tree. Therefore, random forest was applied to examine the robustness of the results, which is the ensemble of the decision trees. In order to get comparable result, the parameters of both models should be the same.

The XGBoost is a boosted tree model, which combines many CART (a binary tree) regression tree models to form a strong classifier. It uses many strategies (regularization, shrinkage, and column subsampling) to prevent overfitting. It has now become a very common feature extraction method in the field of bioinformatics.

For the tree-based model, the importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature.

3) SVC and RFECV-SVC

SVM is a supervised learning model that can be applied in feature selection. For the SVM classifier with linear kernel, the feature importance can be calculated based on the weights assigned to the features.

Just as its name implies, RFECV is a strategy performs recursive feature elimination with cross-validation to select features. The dataset is split into a sequence of train and test portions to perform cross-validation for the estimator. The estimator can be any classifier, in our study, we choose SVC as the estimator to make the comparison to the SVC classifier feature selection. For RFECV, we can check the importance of features based on the order of elimination.

Some other tactics are deployed for the sake of model performance. As Figure2, the class label in all three locations is imbalanced. Therefore, we applied two strategies to handle this problem. In the data splitting stage, we applied stratified splitting; In the stage of model training, we added class weight to each label to reduce the effect of imbalance.

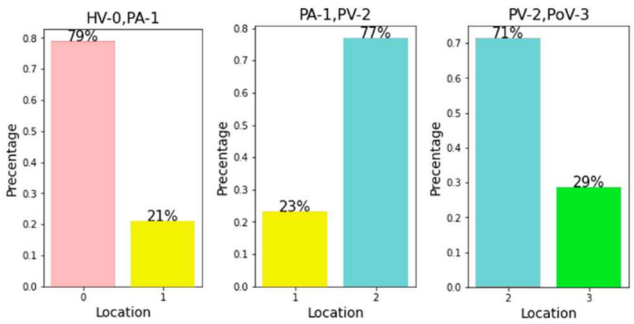


Figure2. The imbalance class of each binary classification task

2.4. Model evaluation

Learning curve was conducted to assess how the sample size influence the performance of model.

Different metrics were calculated to examine the robustness of the results. Apart from the accuracy, we also calculate the sensitivity and specificity to check if the model performs well on both positive and negative labels. The precision is used to check the false positive.

Receptor Operating Curve (ROC) and the confusion matrix were created to visualize the performance. 3 numeric metrics, area under curve (AUC), accuracy and f1 score in test set are regarded as metrics to determine the optimal model. For each location, the best 3 models are qualified to feature selection.

2.5. Feature selection

To find the genes contributing to model's discriminative capability mostly, we obtained the information of feature importance.

Among 5 models, feature importance information can be extracted directly from model's attributes. For SVC with linear kernel, the attribute 'coef_' represents the weight of each gene. For RFECV, the attribute 'feature_names_out' represent the selected gene. The rest of models have 'feature_importance_' attributes.

The union set for each location will be analyzed separately in enrichment analysis, and also the union set among 3 locations will participate in network analysis.

2.6. Enrichment analysis

To find out what biological processes and pathways that might be differentially regulated based on dissemination site, we conducted a Gene Set Enrichment Analysis on the Gene Ontology Database (GO) as well as the KEGG pathways database. For GO analysis, biological process (BP) is mainly discussed. The enrichment analysis is carried out in EnrichR[5].

2.7. Generation of protein-protein interaction network

The union set of important gene in all the locations are imported for network analysis. The protein–protein interaction (PPI) network was constructed using the String database (version 11.5)[6], for which genes and interactions were determined and designated as nodes and lines, respectively, in the network. To show the biological interactions among DEGs, their relationship network was visualized using Cytoscape (v3.9.1)[7]software. Moreover, the cytoHubba[8] plug-in was used to mine the hub genes using the Betweenness, Stress, Degree, Closeness, and ecCentricity calculation method to further extract the core sub-networks.

3. Result

3.1. Dataset description

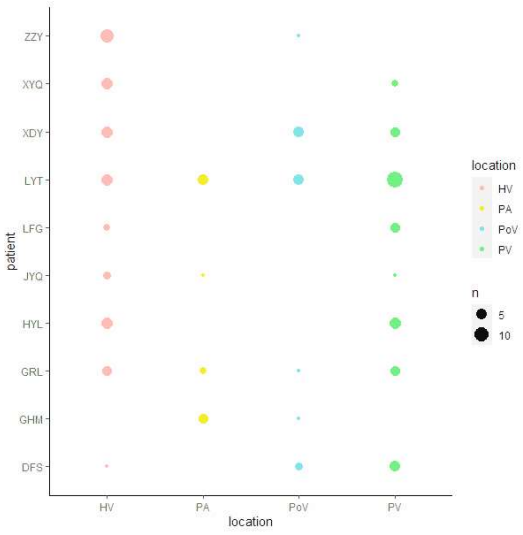
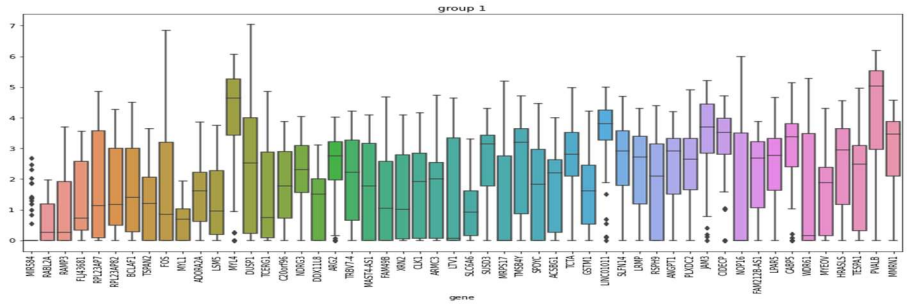
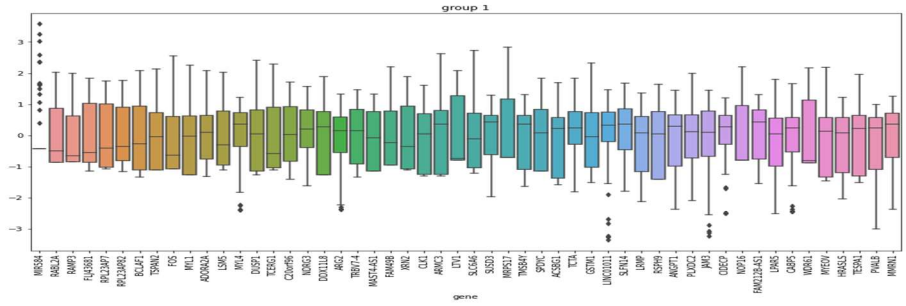


Figure3. The sample size in location and patient level



(a)



(b)

Figure4. The distribution of gene expression level before (a) and after (b) z-score normalization

3.2. DEG analysis

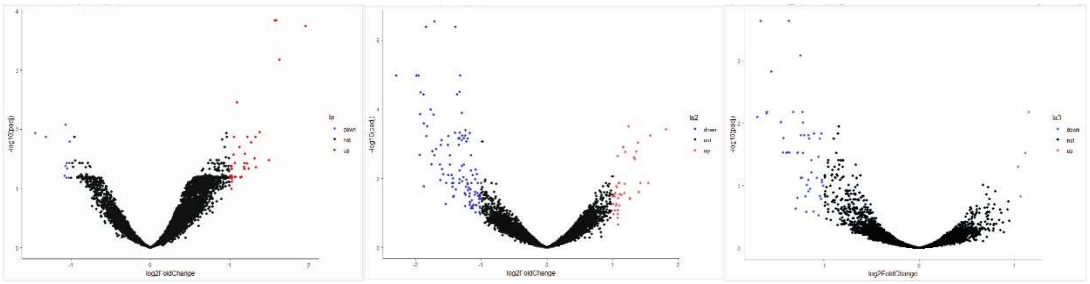


Figure5. The volcano plot of DEG recognized by DEsqe2

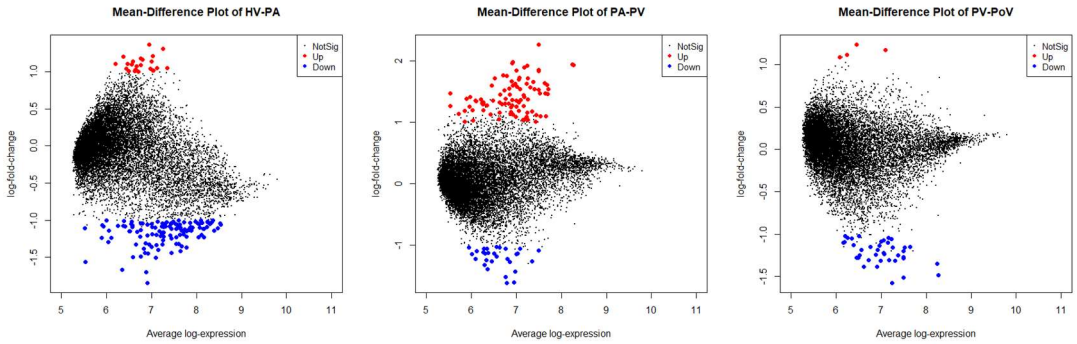
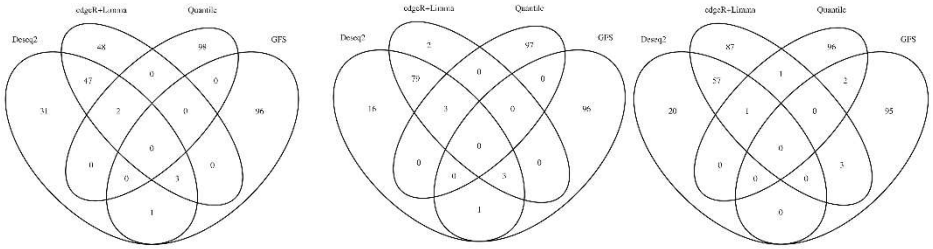


Figure6. The Mean-Difference Plot derived from edgeR and limma

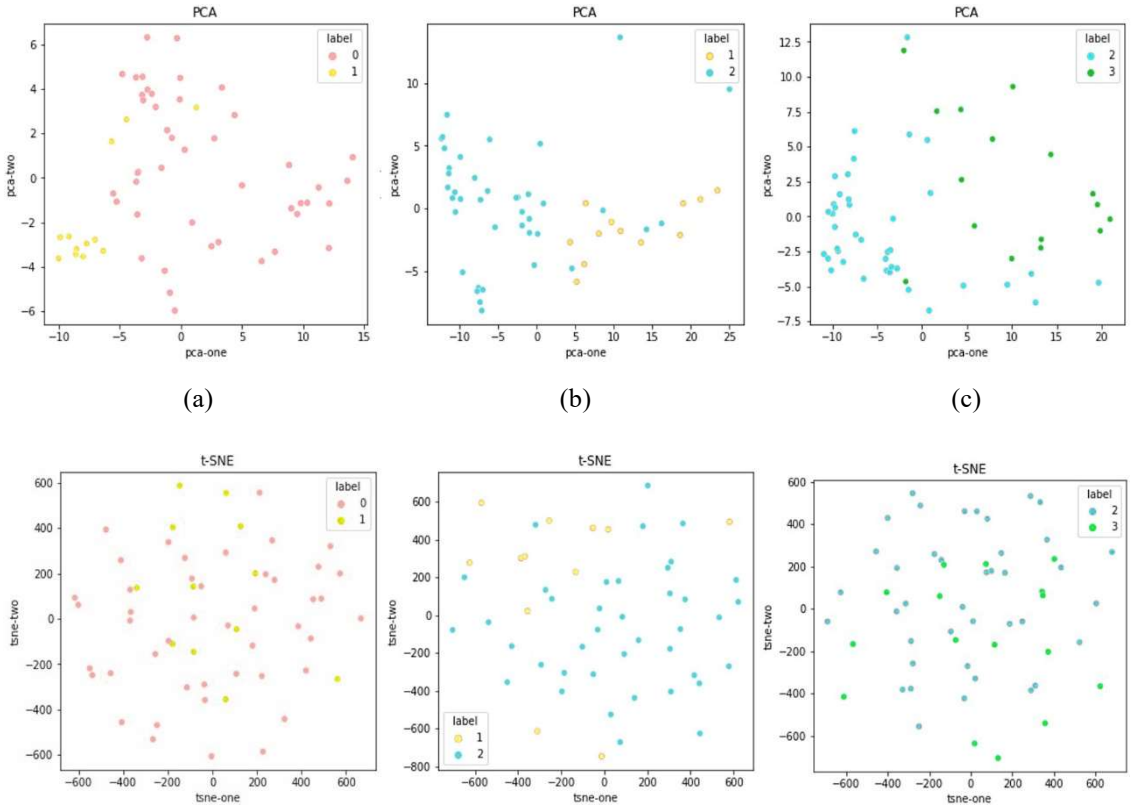


(a) (b) (c)

Figure7. The intersection of DEG among 4 methods in HV-PA (a), PA-PV (b) and PV-PoV (c)

According to above results, 2 results derived from R packages are similar with each other relatively. The intersection genes of DEseq2 and edgeR+ limma are selected for downstream analysis.

3.3. Dimension reduction



(a) (b) (c)

(d) (e) (f)

Figure8. PCA (a, b, c) and t-SNE (d, e, f) for 3 locations

3.4. Model selection

For ROC and Confusion Matrix in each location, please see the Supplementary Figure1-6 in supplementary materials.

Table1. Model performance in HV-PA test set

	DT	RF	SVC	RFECV	XGB
accuracy	0.96	1	0.83	0.79	0.67
f1	0.86	1	0.8	0.67	0.5
AUC	0.94	1	0.97	0.97	0.94
average	**0.92**	**1**	**0.87**	0.81	0.70

Table 2. Model performance in PA-PV test set

	DT	RF	SVC	RFECV	XGB
accuracy	0.4	0.73	1	1	0.78
f1	0.76	0.84	1	1	0.9
AUC	0.4	0.8	1	1	0.97
average	0.52	0.79	**1**	**1**	**0.88**

Table 3. Model performance in PV-PoV test set

	DT	RF	SVC	RFECV	XGB
accuracy	0.95	0.82	0.8	0.95	0.85
f1	0.89	0.75	0.67	0.89	0.73
AUC	0.93	0.82	0.9	0.95	0.92
average	**0.92**	0.80	0.79	**0.93**	**0.83**

Based on average score, the best 3 models are qualified to recognize important genes.

HV-PA: Decision Tree, Random Forest and SVC

PA-PV: SVC, RFECV-SVC, XGB

PV-PoV: Decision Tree, RFECV and XGB

3.5. Feature Selection

The union set of important gene derived from selected model are as follow.

HV-PA: TCERG1, FOS, ANGPT1, ADORA2A, LSM5, RABL2A, TRBV7-4, MIR584, C20orf96, GSTM1, NDRG3, WDR61, SUSD3, XRN2, CLK1, MYL1, TMSB4Y, RAMP3, RSPH9.

PA-PV: RPL23AP7, TSPAN7, AMBP, SNORD3B-1, SNORD3B-2, C20orf96, COMMD5, RNASE6, SNORD3D, FMO5, CHMP4C, CPB2, LUC7L3, AKR1C1, PLA2G16, ANXA1, S100A13, FABP1, MIR584, FADD, MT1E, APOC4, AKR1C2.

PV-PoV: ALDOB, MTCH2, FCER1A, ETFB, HRG, MRPL40, MT1H, PTPN6, DCXR, ANGPTL3, MRPL12, AKR1C4.

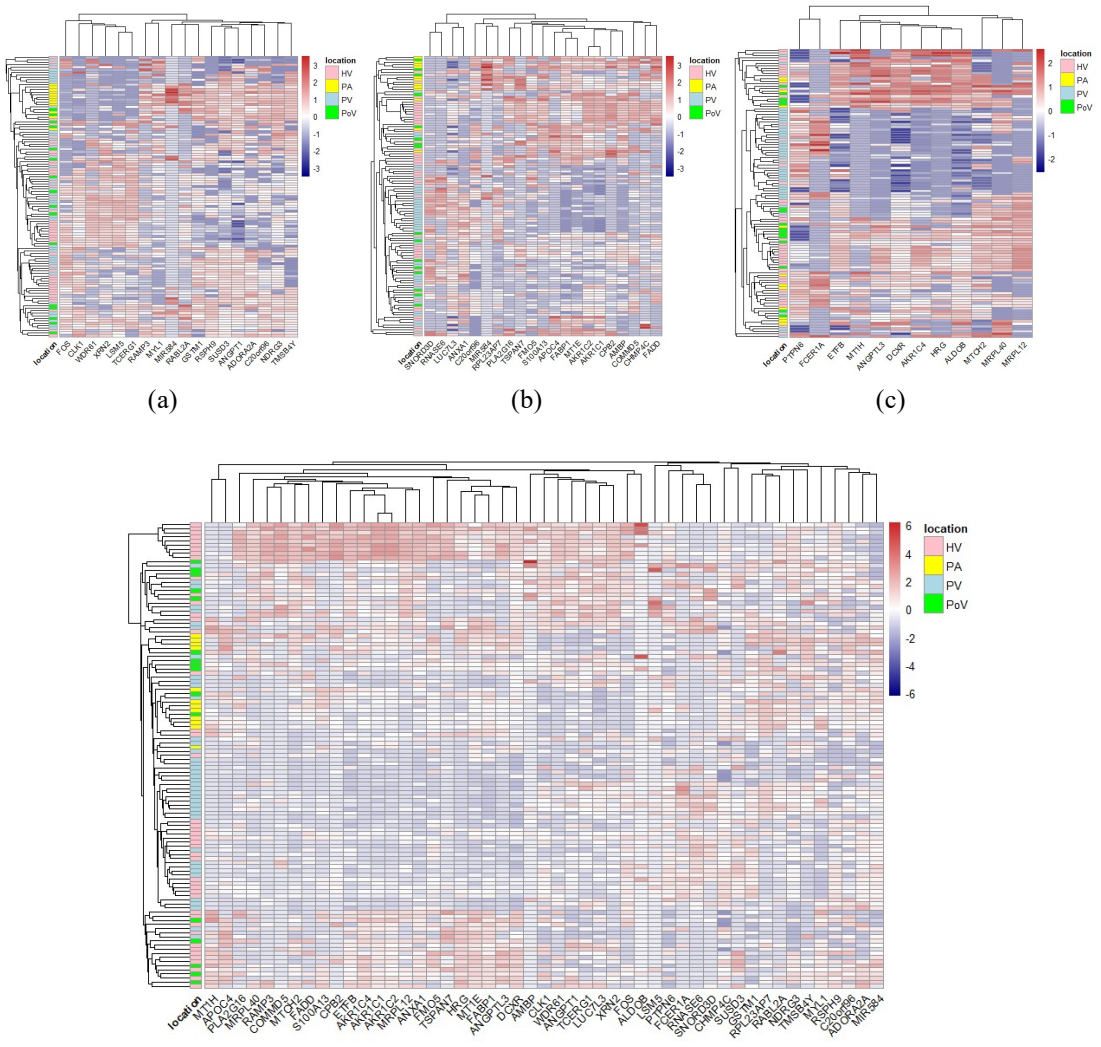


Figure 9. Clustered Heatmap in HV-PA (a), PA-PV (b) and PV-PoV (c) with their own important genes and heatmap with important gene in all locations (d). To some extent, the clustering tendency can be found in PA of (a), PA and PV of (b), PV of (c), PA and PV of (d), which indicate the phenotype of CTC might be unique especially in PA and PV.

3.6. Enrichment analysis

The gene sets produced as a result of these analysis reflect the pathways and gene ontology labels that occur in our most important gene features set significantly higher than can be expected by chance. A total of 249 GO categories with unadjusted p value below 0.05 were found, of these, only 22 had adjusted p values that were below 0.05. They are shown below.

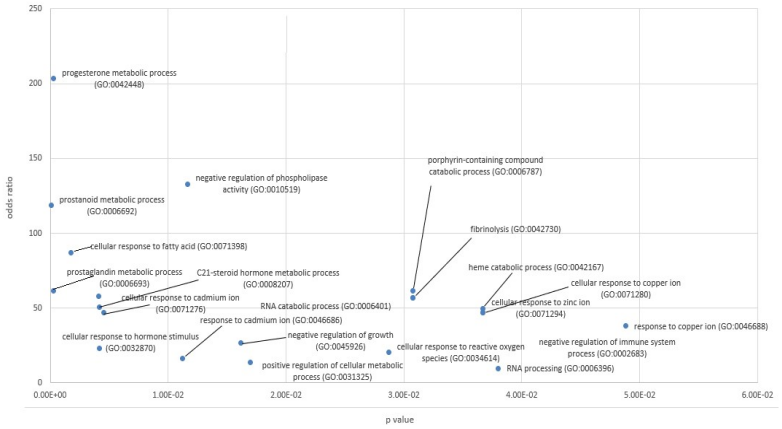


Figure10. Scatterplot of Top GO enriched categories for the union of the most important genes for all comparisons, with adjusted p value above 0.05

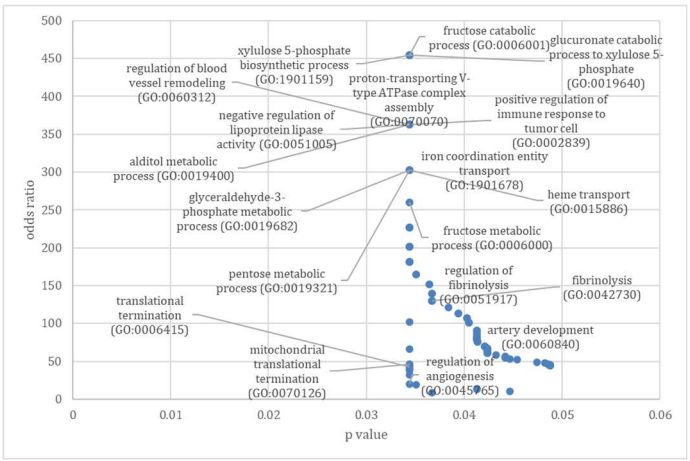


Figure 11. Scatterplot of odds ratio against p value for GO categories enriched comparing peripheral vein and portal vein

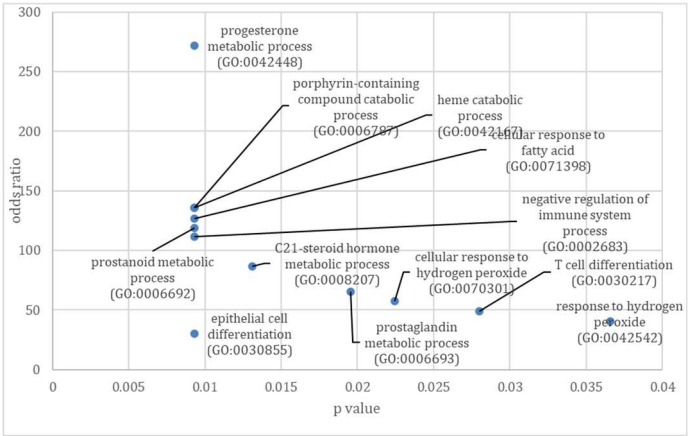
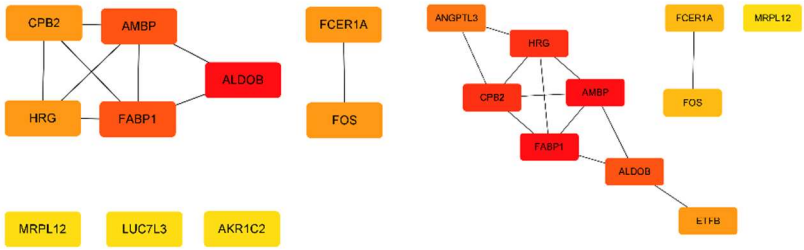


Figure 12. Scatterplot of GO categories enriched comparing peripheral artery and peripheral vein

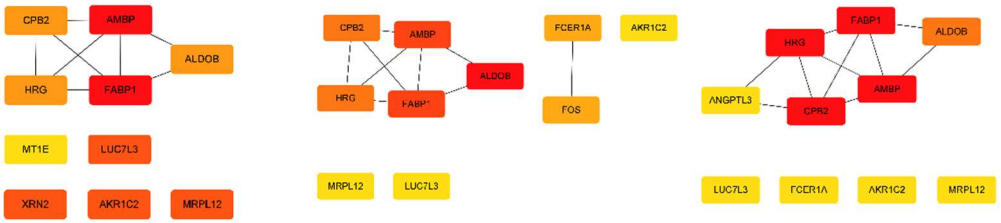
3.7. Network analysis

5 algorithms are introduced for hub gene analysis, including degree, betweenness, closeness, ecCentricity and stress. Top 10 genes are selected based on each algorithm.

The networks containing hub genes are as follow.



(a) Betweenness (b) Closeness



(c) ecCentricity (d) Stress (e) Degree

Figure13. The subnetworks with hub genes identified by 5 ranking algorithms

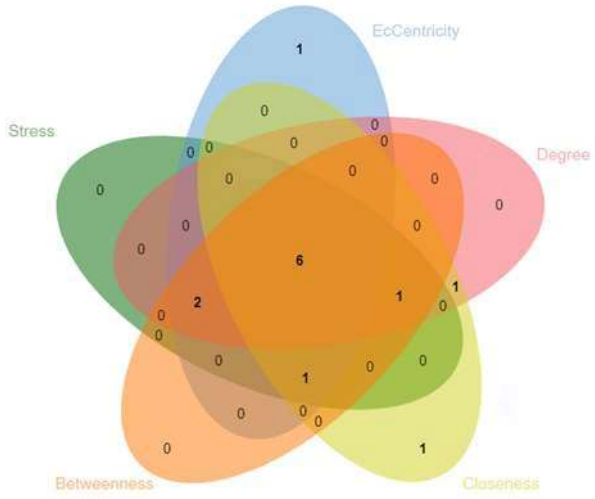


Figure14. The intersection of hub genes recognized by 5 algorithms

Totally 6 genes are regarded as hub gene by all algorithms. They are: ALDOB, AMBP, FABP1, CPB2, HRG and MRPL12.

4. Discussion

4.1. DEG analysis based on statistical methods

2 normalization methods, Gene Fuzzy Score (GFS) [9] and quantile normalization were applied.

Then, we performed the Wilcoxon Rank Sum Test and set the threshold to choose the genes who have lower p-values to do further analysis. Compared with quantile normalization, GFS is a bold normalization method which also makes use of gene expression ranks instead of absolute values, thus earning more robustness to batch effects. Further, we set two quantile thresholds, θ_1 and θ_2, to assign a fuzzified score to each gene in each patient. Ranks below θ_1 in a sample are all assigned with 0, those above θ_2 are given a score of 1, and intermediate ranks are interpolated to obtain a score between 0 and 1. In particular, let r (g_i, p_j) be the rank of gene expression of a gene g_i in patient p_j, and q (p_j, θ) be the rank corresponding to the upper θ_{th} quantile of gene expression in patient p_j. Then, the gene fuzzy score s(g_i, p_j) is given by the following function[9].

$$s\left(g_i, p_j\right) = \begin{cases} 1, & \text{if} \quad q(p_j, \theta_1) < r(g_i, p_j) \\ \frac{r(g_i,p_j)-q(p_j,\theta_2)}{q(p_j,\theta_1)-q(p_j,\theta_2)}, & \text{if} \quad q(p_j, \theta_1) > r(g_i, p_j) \geq q(p_j, \theta_2) \\ 0, & \text{otherwise} \end{cases}$$

Apart from the use of rank values in computing transformed scores, GFS also benefits from the fact that it allows for selection of quantile threshold. In this project, we fix θ_1 to 5% and θ_2 to 15%. The reason for using the nonparametric Rank Sum test instead of t-test was the failure of the normality test in the t-test. This can be due to the gene fuzzy scoring. The null hypothesis is that two sets of scores are samples from the same population, therefore they do not differ systematically. The steps are it first transform observed scores into ranks, then calculate W (Difference from the rank sum of A group to the rank sum of B group) and do calculation of the probability of the observed case or more extreme cases, the final step is calculating of probability[10]. After we done that, we chose 100 top ranked genes to do further machine learning analysis.

The second preprocessing methods is quantile normalization. As we mentioned previously, some standard normalization techniques may perform poorly, so we would like to try one of the techniques on our data set to compare the result with GFS method. Similarly, we performed Wilcoxon Rank Sum Test after quantile normalization and get 100 top ranked genes with lower p-values to do further analysis. We find that there is barely intersection in results of 2 methods. For GFS method, it fuzzified the expression ranks so that irrelevant fluctuations introduced by minor differences in ranks are alleviated, and noise from low-ranked genes is discarded. Since ranks are known to be relatively more robust to batch effects than absolute expression values, quantile normalization is expected to lead to better performance on datasets with batch effects. And the expression rank of low-expression genes is highly unstable. This may adversely affect the performance of a ranking-based normalization method such as quantile normalization. Therefore, the huge difference between 2 methods may be due to the nature of them, in the other word, the origin purpose of design the methods. Our data is from one resource, do not have obvious batch effect and there are some low-expression genes which may let quantile normalization perform bad. Therefore, there is no one-size-fit-all methods for data processing. The methods chosen for a specific task should strongly depend on the characteristic of data.

4.2. DEG analysis based on R packages

Basically, ordinary statistics use the t-test to compare the difference between two samples, and then use analysis of variance to determine whether there is a difference between the samples when the sample is large. Of course, the premise is that the sample comes from a normally distributed

population or random independent large sampling. For non-normal distribution, we use Wilcox instead of a t-test.

For the differential expression analysis of gene chips, since it is generally believed that the data are subject to normal distribution, the differential expression analysis is applying a t-test sum or variance analysis to each gene. High throughput finds many genes at one time, so multiple tests need to be corrected to control false positives. At present, Limma is the most used in gene chip analysis.

In this study, we tried 2 methods based on Deseq2, edgeR and limma. Deseq2 works individually, and the second methods applied both of edgeR and limma, in which edgeR package is used to import, organize, filter and normalize the data, followed by the limma package with its voom method, linear modelling and empirical Bayes moderation to assess differential expressed gene.

Deseq2 is devised based on the assumption that most genes are not DE. The DESeq2 scale factor for a given lane was calculated as the median of the ratio of the read counts for each gene to the geometric mean of all lanes. The basic idea is that non-DE genes should have similar read counts between samples, resulting in a ratio of 1. Assuming most genes are not DE, the median of this ratio for the lane provides an estimate of the correction factor that should be applied to all read counts of this lane to fulfill the hypothesis. After normalization, Deseq2 fits data to a negative binomial generalized linear models, and perform Wald test or Likelihood Ratio Test for DEG identification[11].

EdgeR is also based on the assumption that most genes are not DE. Calculate the TMM factor for each lane, where one lane is regarded as the reference sample and the other lanes are regarded as the test sample. For each test sample, TMM was calculated as the weighted average of the log ratios between that test and the reference, after excluding the most expressed genes and genes with the largest log ratios. Based on the assumption of low DE, this TMM should be close to 1. If not, its value provides an estimate of a correction factor that must be applied to library size (rather than raw counts) to fulfill the hypothesis. To obtain normalized read counts, these normalization factors were rescaled by the mean of normalized library sizes. Normalized read counts were obtained by dividing the raw read counts by these rescaled normalization factors. Limma plays the role of removing heteroscedasticity from count data[12]. In limma, linear modelling is carried out on the log-CPM values which are assumed to be normally distributed and thus the mean-variance relationship should be accommodated using precision weights calculated by the voom function. For DEG detected, limma fits a separate model to the expression values for each gene. Next, empirical Bayes moderation is carried out by borrowing information across all genes to obtain more precise estimates of gene-wise variability[13].

In terms of the similarities and differences among the 2 methods, DESeq2 and EdgeR provide different normalization algorithms, which are very similar in that both assume no differential expression of genes. DESeq2 uses a "geometric" normalization strategy, while EdgeR is a log-ratio-based weighted average method. Both initially normalize the data by computing a size/normalization factor.

The main differences between the 2 methods are the model fitting and DEG testing, which may be the underlying reason of their results are not absolutely identity with each other. According to the Venn plot, the intersection of 2 R packages is considerable, which indicates their result could be more persuasive.

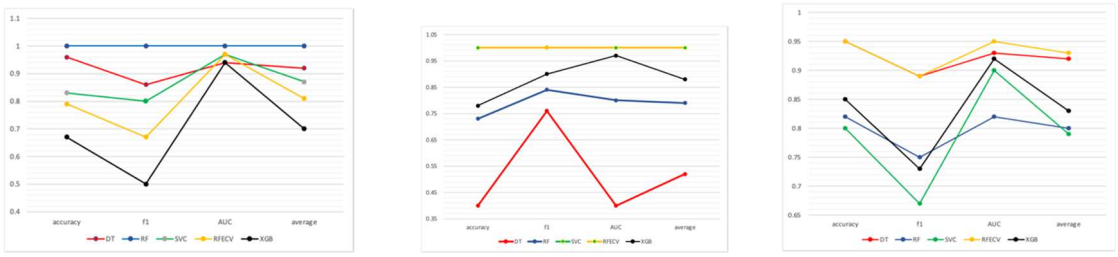
4.3. Machine learning

The discrimination capability is beyond our assumption. The paucity of sample size and the relative high dimension of the dataset didn't confine the learning performance and most of the AUC is close to 1. This can be attributed to the previous DEG analysis which narrow down the dimension greatly (from 13005 to less than 100) and extracted the useful information from redundant data set.

Since the classification task among 3 locations are similar in sample size, number of features and performance, the result in first location HV-PA can be representative and it is discussed for illustration.

4.3.1. Model selection metrics

The performance of the ML model in binary classification tasks usually evaluated based on the different metrics. Classification accuracy is a general metric for model performance, which is taken as one selection metric. Due to the nature of trade-off the of precision and recall, F1 score is one of the numeric metrics to appraise to model. In addition, AUC is not sensitive to the imbalanced class label which is also suitable for our dataset. In our study, three models with higher average scorers are selected as feature extractors. Figure15 displays the variety of evaluation scorers among different models, which guarantees the robustness of model selection.



(a) HV-PA (b) PA-PV (c) PV-PoV

Figure 15. Model selection metrics for five models in 3 locations

4.3.2. Tree-based model

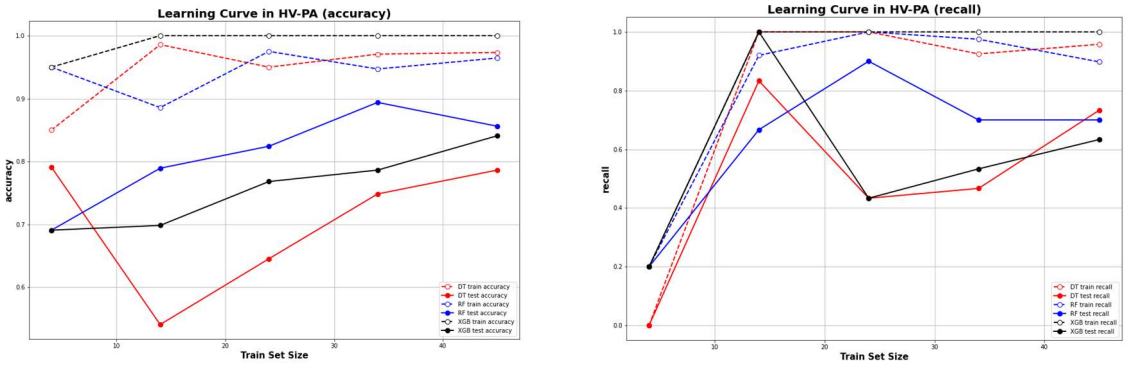
The table below shows the evaluation scores for tree models. According to the table, the XGBoost performed well when classifying HV, but didn't perform well on PA classification. This indicates that the performance of XGBoost may not be stable when performing feature selection. The random forest model is the most robust one which probably is the best tree model for feature selection for location HV- PA.

Table4. The evaluation of tree-based models

HV-PA						
model	**accuracy**	**precision**	**sensitivity**	**specificity**	**f1-score**	**AUC**
DT	0.96	1	0.92	1	0.96	0.73
RF	0.96	1	0.92	1	0.96	0.90
XGBoost	0.67	0.86	1	0.33	0.92	0.89

The graph below shows the learning curve (accuracy) of three tree models for location 1. According to the graph, the classification accuracy of the tree model is positively correlated with

the training set size. The classification accuracy of random forest and XGBoost are quite similar both of which are higher than the decision tree. This phenomenon indicates that the ensemble of trees could increase the performance of the model in some circumstance. However, among 3 locations, the performance of ensemble learning not always surpass the basic tree. The sensitivity of decision tree surpassed Random Forest and XGBoost. This may imply that the instability of tree-based models in such small dataset. Decision tree is a relative weak algorithm which can only learn the basic pattern of data, whereas ensemble learning is designed to learning from each weak classifier. When the pattern is easy, the complex algorithms could not perform better than we assumed.



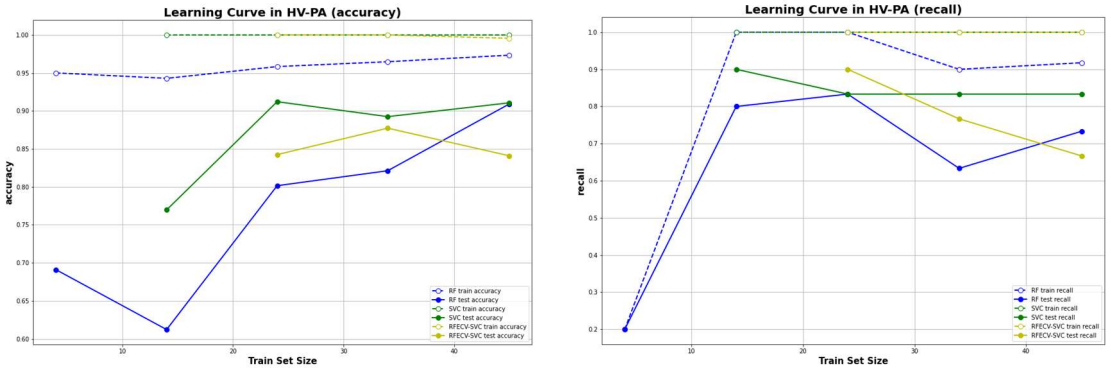
(a) accuracy (b) recall

Figure 16. Learning curve for tree-based model based on accuracy(a) and recall(b)

4.3.3. SVC and SVC-RFECV

The line graph (a) shows the learning curve (accuracy) of random forest, SVC and SVC-RFECV. The SVC and random forest perform better than the SVC-RFECV, while SVC performs better than the random forest when the sample size is small. But the gap of the performances between the two models decreased with the increase of the sample size. The performance of the ensemble model increased with the training set. Unlike other models, the performance of SVC-RFECV didn't increase with the growth of the sample size, the feature elimination strategy also didn't exhibit significant improvement.

The line graph (b) shows the learning curve (sensitivity) of the five models. The performances of SVC, SVC-RFECV, and random forest are better than the other two models on the positive sample classification. This also indicates the robustness of results in the premise of performing feature extraction.



(a) (b)

Figure 17. The learning curve of all models based on accuracy(a) and recall(b)

4.3.4. Dimension reduction

PCA and t-SNE are 2 widely-used dimension reduction methods in biomedical domain. We performed PCA and t-SNE for data visualization for the sake of comparison. Figure8 is the visualization of the data set after dimension reduction. The top part of the graphs shows the results of PCA on three groups, the majority of samples with the same labels gather together, which indicates that the principal components retain the majority of information of the dataset.

On the contrary, for the t-SNE, there is no perspicuous cluster tendency and the data point in different class mix with each other. The reason behind it is that the t-SNE is a t-distribution-based algorithm and it is non-linear. According to the last chapter, the SVC classifier with a linear kernel has the best performance, which means the data point is distributed on the two sides of a linear hyperplane. The non-linear transformation will ruin the distribution and finally lead to information loss.

4.4. Biological significance

The enrichment analysis aims to obtain insight of gene selected by DEG and machine learning. For Figure 11, Surprisingly, except for RNA processing and catabolism, and fibrinolysis, all the GO categories were generally related to metabolic processes. The genes with the top 5 lowest p values were exclusively dealing with fatty acid and lipid metabolism. Sorting the significant genes by odds ratio, again, the top 5 genes were exclusively related to fatty acid and lipid metabolism. The odds ratio defines the ratio in which the proportion of genes in the given set is higher than the proportion of genes with that particular GO term in the Gene Ontology. Thus, we can say that there are far more lipid metabolism (in these GO categories) genes important in determining CTC dissemination than expected, and this enrichment result is extremely unlikely to have occurred by chance (having an extremely low p value)

It could be possible that the metabolism related genes are differentially expressed as a result of nutrient and hypoxic related responses to the sites with different levels of oxygenation and nutrient contents. In particular, the hepatic portal vein is a direct conducting channel supplying nutrients from the colon to the liver, and all absorbed nutrients (with the exception of very large chain fatty acid, which get absorbed by the lymphatic system) flow through the portal vein. Thus, it could be possible that the enriched genes seen here are driven by the differential changes comparing CTCs disseminated to the portal vein compared to the other nutrient poorer sites.

In addition, there are several GO categories which are related to respiration, and these involve the response to ions and cofactors playing functions in or disruptors to the respiratory chain. These include copper, heme, porphyrin and reactive oxygen species (ROS). The enriched genes that we observed might occur as a result of CTC responses to differing levels of oxygenation comparing the different dissemination sites. In particular, the peripheral artery is highly oxygenated in comparison to the rest of the sites, and it is expected that the hepatic portal vein and hepatic vein might be poorer oxygenated than the peripheral vein due to the high oxygen consumption level of the intestines[14].

The gene regulatory response to ROS proceeds through hypoxia inducible factor (HIF), and this transcription factor might be responsible in producing the gene expression changes which result in ROS related genes being enriched in our GO analysis.

Additionally, genes that are responsive to changes in copper concentration might also be enriched as a result of changes in the local concentration of this metal[15], or by hypoxia, due to the differences in nutrient content and oxygen content between the sites. Long-term Ischemia causes changes in the concentrations of metals and minerals, including copper. This could result in changes in the expression of genes associated with HIF[16]. Finally, studies have shown increased uptake of copper in response to hypoxia[17, 18]. All these changes could result in an enrichment of copper response genes in our analysis.

For the KEGG pathways analysis, we did not observe any significantly enriched genes with adjusted p value below 0.05. However, the two pathways with p values close to 0.05, steroid hormone biosynthesis and RNA degradation, are also among the functions observed in our GO enrichment analysis.

To further investigate if indeed the differences in oxygenation level and nutrient content could be responsible for the enriched GO categories, we obtained the set of genes which had the highest feature importance in discriminating the Portal vein with the peripheral vein (to account for nutrient content), and the set of genes discriminating peripheral vein and peripheral artery. The top GO categories that had an adjusted p value of less than 0.05 comparing the portal vein with the peripheral vein is shown in Figure12.

Again, the list of GO categories is enriched with metabolic related categories, with the top 5 being related to carbohydrate metabolism. In comparison, the metal ion related gene function categories (in particular, copper) were missing from the comparison between peripheral artery and peripheral vein (Figure13). On the other hand, the GO categories relating to cofactors in the mitochondrial chain, heme, porphyrin, as well as the ROS hydrogen peroxide, were retained, suggesting that only cofactor and ROS related genes, and not metal ion, might be regulated by hypoxia.

As Figure13 illustrates, the comparison between peripheral artery and hepatic vein shows exclusively GO categories related to RNA processing. Interestingly RNA degradation features heavily in the enriched set, with all of the categories except for RNA processing being related to degradation.

It is known that due to a shortage of ATP, translation slows down[19] in hypoxic conditions. In addition, the stability of RNA transcripts is decreased in hypoxia, in a HIF dependent manner. Tiana et al[20] observed that the half-lives of RNA transcripts fell by around 20% in the presence of hypoxia. The dependence of this phenomenon on HIF, a transcription factor, suggests that differential gene regulation of RNA catabolism factors is possible cause for decreased RNA stability, even though the exact mechanism was not discovered by Tiana et al.

RNA degradation is not featured at all in the comparison between the peripheral artery and vein, suggesting that it could be a hepatic related phenomenon in contrast to a hypoxia dependent one. Alternatively, because of the higher oxygen consumption of the intestines (of which blood indirectly flows to the hepatic vein from the portal vein) and the liver, it could be possible that the levels of hypoxia in the peripheral vein were insufficient to trigger the hypoxia related RNA degradation processes.

4.5. Hub gene ranking algorithms

5 hub gene ranking algorithms are deployed. Except 6 genes detected by all the algorithms, the rest of genes have 6 origins: 2 of them not recognized by closeness, 1 not for degree, 1 not for ecCentricity, 1 only for ecCentricity, 1 only for Closeness and 1 only for degree and closeness. The controversy mainly focuses on closeness. To investigate what is the difference between closeness and others, the understanding of each algorithm is necessary.

Suppose a node v, the formula for each metric are as below[8].

Degree(v) is the number of neighbors of v.

$$Deg(v)=|N(v)|.$$

Closeness(v) is inversely proportional to the shortest path between nodes u and v, which u belongs to the component containing node v (component means a subnetwork containing only connected nodes). The longer the distance, the lower the closeness.

$$Clo(v)=\sum_{w\in V}\frac{1}{dist\,(v,w)}$$

ecCentricity(v) is inversely proportional to the max shortest path between nodes u and v. It has a coefficient, V(C(v)) is the collection of nodes in component containing v, and V is the total nodes in the network. As its name implies, it measures whether the node close to the center of network.

$$EC\,(\mathsf{v})=\frac{|V\,(C\,(v))|}{|V|}\times\frac{1}{\max\{dist(v,w)\colon w\in C\,(v)\}}$$

Stress(v) is the sum of σ_{st} (v), which is the number of shortest paths from node s to node t which use the node v. Both of s and t belong to component containing v. It can give us the insight of the ability of nodes v to connect others.

$$Str\,(v)=\sum_{s\neq t\neq v\in C(v)}\sigma_{st}\,(v)$$

Betweenness(v) is similar to stress. the only difference is the σ_{st}, which is the number of shortest paths from node s to node t.

$$BC\,(v)=\sum_{s\neq t\neq v\in C(v)}\frac{\sigma_{st}(v)}{\sigma_{st}}$$

Each formula shows different emphasis and they all aim to identify the hub gene in a network.

Based on the explanation of closeness, the controversy of it may imply that only distance of shortest path is not sufficient to assess the hub gene. And its analog, ecCentricity, which has a calibration term |V(C(v))| / |V|, performs more concordant with others. This phenomenon may imply that tuning the formula after devising it is necessary. And the 6 genes detected by all the algorithms may have strong credibility under the specific circumstance.

5. Conclusion

Under the background of single cell gene expression profile, we conduct a thorough bioinformatics analysis encompassing DEG analysis based on R package and statistical methods, feature selection based on different machine learning algorithms and evaluation metrics, enrichment analysis and network analysis with different hub gene ranking algorithms to identify

the important gene for hepatocellular metastasis. It's unsurprisingly that the result is strongly impacted by methods.

The different combination of methods may exhibit unexpected result, such as the outcome of different normalization methods with the same statistical methods scarcely similar, and machine learning performs good in small data set after using feature selected by R package. As for bioinformatics analysis is usually regarded as guideline for downstream wet lab experiments or clinical trials, the reliability or reproducibility cannot be underestimated. Therefore, a thorough, scrupulous and comprehensive work flow with rigorous verification is highly required. We can't say that a specific method is suitable for any sorts of data or purpose. Whether the result is convincing or not should be determined by comparison and evaluation among different methods.

As for the biological insight obtained from this study, we found that majority of the important genes are related to nutrient metabolism, inflammation, and respiration. It could be possible that the metabolism related genes are differentially expressed as a result of nutrient and hypoxic related responses to the sites with different levels of oxygenation and nutrient contents. Inflammatory pathways might be triggered by shear stress encountered by CTCs in arteries.

References

1. Zhou, J., et al., *Guidelines for Diagnosis and Treatment of Primary Liver Cancer in China (2017 Edition).* Liver Cancer, 2018. **7**(3): p. 235-260.
2. Poon, R.T., et al., *Tumor microvessel density as a predictor of recurrence after resection of hepatocellular carcinoma: a prospective study.* J Clin Oncol, 2002. **20**(7): p. 1775-85.
3. Mann, J., H.L. Reeves, and A.E. Feldstein, *Liquid biopsy for liver diseases.* Gut, 2018. **67**(12): p. 2204-2212.
4. Sun, Y.F., et al., *Dissecting spatial heterogeneity and the immune-evasion mechanism of CTCs by single-cell RNA-seq in hepatocellular carcinoma.* Nat Commun, 2021. **12**(1): p. 4091.
5. Chen, E.Y., et al., *Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool.* BMC Bioinformatics, 2013. **14**: p. 128.
6. Szklarczyk, D., et al., *The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets.* Nucleic Acids Res, 2021. **49**(D1): p. D605-d612.
7. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.
8. Chin, C.H., et al., *cytoHubba: identifying hub objects and sub-networks from complex interactome.* BMC Syst Biol, 2014. **8 Suppl 4**(Suppl 4): p. S11.
9. Belorkar, A. and L. Wong, *GFS: fuzzy preprocessing for effective gene expression analysis.* BMC Bioinformatics, 2016. **17**(Suppl 17): p. 540.
10. Kim, H.Y., *Statistical notes for clinical researchers: Nonparametric statistical methods: 1. Nonparametric methods for comparing two groups.* Restor Dent Endod, 2014. **39**(3): p. 235-9.
11. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biology, 2014. **15**(12): p. 550.
12. Law, C.W., et al., *RNA-seq analysis is easy as 1-2-3 with limma, Glimma, and edgeR.* F1000Res, 2016. **5**.
13. Smyth, G.K., *Linear models and empirical Bayes methods for assessing differential expression in microarray experiments.* Stat Appl Genet Mol Biol, 2004. **3**: p. Article3.

14. Rowell, L.B., J.R. Blackmon, and R.A. Bruce, *INDOCYANINE GREEN CLEARANCE AND ESTIMATED HEPATIC BLOOD FLOW DURING MILD TO MAXIMAL EXERCISE IN UPRIGHT MAN.* J Clin Invest, 1964. **43**(8): p. 1677-90.
15. González, M., et al., *Expression of copper-related genes in response to copper load.* The American Journal of Clinical Nutrition, 2008. **88**(3): p. 830S-834S.
16. Bell, E.L., et al., *Mitochondrial Reactive Oxygen Species Trigger Hypoxia-Inducible Factor-Dependent Extension of the Replicative Life Span during Hypoxia.* Molecular and Cellular Biology, 2007. **27**(16): p. 5737-5745.
17. *Hypoxia makes copper move.* Journal of Cell Science, 2009. **122**(9): p. e904-e904.
18. Zimnicka, A.M., et al., *Upregulated Copper Transporters in Hypoxia-Induced Pulmonary Hypertension.* PLOS ONE, 2014. **9**(3): p. e90544.
19. Chee, N.T., I. Lohse, and S.P. Brothers, *mRNA-to-protein translation in hypoxia.* Molecular Cancer, 2019. **18**(1): p. 49.
20. Tothill, R.W., et al., *An Expression-Based Site of Origin Diagnostic Method Designed for Clinical Application to Cancer of Unknown Origin.* Cancer Research, 2005. **65**(10): p. 4031-4040.

Member	Contribution
Zhang Zhu	Devise the framework and divide the work Perform GFS and Wilcoxon sum rank test Perform edgeR + limma Modify codes of Machine Learning Select model based on evaluation metrics Perform enrichment analysis Perform PPI Perform hub gene ranking Write introduction, methods, discussion of hub gene ranking algorithms and conclusion Modify everyone's report Compile and organize the report Compile and organize the slides
Ye Shuchen	Help clarify the direction of research Perform DESeq2 Comparing Different Methods for Differential Expression Analysis Modify the color of the paper picture Help modify the article format
Han Wenhao	Develop the strategy of ML task Perform dimension reduction and visualization using PCA and t-SNE Implement machine learning Implement the visualization of evaluation metrics Perform evaluation of ML task Help write the methods and discussion
Li Guanchen	DEG analysis: use statistical methods to detect DEGs

	Write the discussion of statistical methods
Abel Tan	Do literature review Analysis of the gene function of most important genes feature result