## Abstract:

Among the common cancers in women are breast cancer, cervical cancer, endometrial cancer, and ovarian cancer. These cancers have become the number one killer of women in the world today. To explore the role of integrating molecular data with clinical variables in the promotion of survival prediction in these cancers, we used different molecular data from 1861 samples of the four cancer types in the TCGA project, including DNA methylation, mrna, miRNA, and Protein expression for retrospective prediction of patient survival. Using the c-index index to test the performance of the predictive model, it was found that the combination of molecular data (methylation, mRNA, miRNA) and clinical data showed a statistically significant change in the predictive power of the model constructed by cancer CESC (c-index increased by 8.73%-15.03%).

乳腺癌、宫颈癌、子宫内膜癌、卵巢癌等是在女性中常见的癌症，这些癌症成为了当今世界女性的头号杀手。为了探索在这些癌症中，整合分子数据与临床变量对生存预测的促进作用，我们使用了来自TCGA项目中这四种癌症类型的1861个样品的不同分子数据，包括DNA methylation、mRNA、miRNA、和protein expression 来进行回顾性预测患者存活率。使用c-index指数来检验预测模型性能，发现将分子数据（methylation, mRNA, miRNA）与临床数据结合后，癌症CESC所构建模型的预测能力有统计学上显著性改变（c-index提高了8.73%-15.03%）。

## Keywords:

Cox Proportional Hazards Regression Model, Random Survival Forest model, LASSO

## Introduction

Breast cancer is the most common female malignancy and the second leading cause of cancer death. Breast cancer is most common between the ages of 20 and 39. In underdeveloped countries, the average age of women diagnosed with breast cancer is even lower than that of developed countries by about 10 years. There is evidence that the incidence of breast cancer in young women is increasing[1-3]. The increase in the incidence of breast cancer every year can lead to a disease burden.

乳腺癌是最常见的女性恶性肿瘤，同时也是癌症死亡的第二大原因。乳腺癌在20到39岁之间最常见，在不发达国家，被诊断患有乳腺癌的妇女平均年龄甚至比发达国家低约10年，有证据表明，年轻女性乳腺癌的发病率正在增加[1-3]。每年乳腺癌发病率的增加可能会导致疾病负担。

Cervical cancer is the leading cause of morbidity and cancer deaths in women around the world. Nearly 500,000 women worldwide suffer from the disease each year, and approximately two-thirds of patients are diagnosed with locally advanced cervical cancer[4, 5]. Almost all cases were caused by the oncogenic human papilloma virus (HPV), which continues to infect 15 of the genotypes. Each genotype of HPV is independently infected and has a different cancer risk[6].

宫颈癌是世界各地妇女发病和癌症死亡的主要原因。全世界每年有近50万妇女患上这种疾病，大约有三分之二的患者被诊断为局部晚期宫颈癌[4-5]。几乎所有的病例都是持续感染15种基因型之一的致癌性性人乳头瘤病毒（HPV）导致的。HPV的每种基因型都是独立感染，有着不同的致癌风险[6]。

Endometrial cancer is the most common malignant tumor of the female reproductive tract in developed countries. Endometrial cancer mainly affects postmenopausal women， however, 15-25% of cases are diagnosed before menopause[7].

子宫内膜癌是发达国家女性生殖道最常见的恶性肿瘤，子宫内膜癌主要影响绝经后妇女；然而15-25%的病例在绝经前被诊断出[7]（这里引用文献不太确定）。辅助以积极的治疗，晚期患者5年生存率仍然低于20%。

Ovarian cancer is the fifth most common tumor among women in the world[7]. Its incidence rate has been increasing year by year, and mortality has ranked first among gynecologic tumors. In recent years, the occurrence of ovarian cancer has become younger[8-10].

卵巢癌是世界女性中第五大常见肿瘤[7]，其发病率逐年上升，死亡率在妇科肿瘤中名列第一。近年来卵巢癌的发生呈现年轻化的趋势[8-10]。

Using survival analysis to study time-event results has important clinical and statistical reasons. In clinical research, the main goal of survival analysis is to find the factors that can predict the survival rate of patients in specific clinical situations. Ideally, we can build an accurate prognosis model.

使用生存分析来研究时间-事件结果有重要的临床和统计学原因。在临床研究中，生存分析的主要目标是找到能够预测特定临床情况下患者存活率的因素。理想情况下，我们能够构建一个准确和精确的预后模型[1]。

Hugo Gómez-Rueda integrated the mRNA, miRNA, CNA, and DNA somatic mutation data of OV, GBM, LUAD, and BRCA, respectively, to produce a model with slightly higher performance, using three different algorithms for predictive model selection[11]. José M. Lezcano-Valverde uses machine learning methods to develop Random Survival Forest (RSF) and independently validate prediction models for death probability of rheumatoid arthritis (RA), which can provide evidence for further external validation[12]. Bin Zhu performed pan-cancer prognostic evaluation and proposed a multi-omic core machine learning method to systematically quantify the prognostic power of high-throughput genome, epigenome, and transcriptome features[13]. Treppmann T et al. used the Bayesian method to integrate the copy number variation data into the gene expression-based survival prediction model of glioblastoma (GBM) patients, and studied the behavior and predictive performance of the model under different conditions[14]. Yuan yuan et al. used RSF and Cox methods to construct models for GBM, KIRC, OV, and LUSC, and conducted in-depth analysis of well-performing models to obtain important biological insights[15].

Hugo Gómez-Rueda等人分别对OV、GBM、LUAD、BRCA的mRNA、miRNA、CNA、DNA 体细胞变异数据整合产生了具有稍高性能的模型，使用三种不同算法用于预测模型选择[24]；José M. Lezcano-Valverde等人使用机器学习方法随机生存森林（RSF）开发和独立验证类风湿性关节炎（RA）死亡概率预测模型，可以为进一步的外部验证工作提供证据[25]；Bin Zhu等人进行了泛癌预后评估，并提出了多组学核心机器学习方法，以系统地量化高通量基因组、表观基因组和转录组特征的预后能力[26]；Treppmann T等人使用贝叶斯方法将拷贝数变异数据整合到基于基因表达的胶质母细胞瘤（GBM）患者生存预测模型中，研究模型在不同情况下的行为和预测表现[27]。

Yuan yuan等人使用RSF与Cox方法分别对GBM、KIRC、OV、LUSC构建模型，对表现好的模型进行深入分析，得到重要的生物见解[13]。

The Cox proportional hazards model is the most important and commonly used survival analysis model. It is a semi-parametric model proposed by the British statistician D.R.Cox in 1972. It is mainly used for the prognosis of tumors and other chronic diseases. It can also be used to explore the causes of cohort studies. Its main advantages are (1) multi-factor analysis methods; (2) disregarding the distribution of survival time; and (3) the use of truncated data.

Cox比例风险（Cox’s proportional hazards regression model）模型是生存分析最重要也是最常用的模型，是一种半参数模型，该模型由英国统计学家D.R.Cox于1972年提出，主要用于肿瘤和其他慢性病的预后分析，同时也可用于队列研究的病因探索。其主要的优点有（1）多因素分析方法；（2）不考虑生存时间分布；（3）可以利用截尾数据。

The RSF (random surviving forest) model is also a commonly used survival prediction analysis model. The RSF is a method of fusing survival trees. It inherits the advantages of random forest, which can anti-noise, prevent over-fitting, and handle non-linear correlation. It can be used for high Dimensional data analysis and variable screening.

RSF（随机生存森林）模型也是进行生存预测分析较为常用的一个模型，RSF 方法是一个组合生存树方法， 继承了随机森林抗噪声、防止过拟合、可处理非线性相关等优点， 可用于高维数据分析和变量筛选。

其计算原理与随机森林相似，即用自助法（bootstrap）从原始数据中有放回的随机抽取N个自助样本，对每个样本都建立一个二元递归生存树。（用于方法里面介绍）

The TCGA project generated genome, transcriptome, epigenomic, and proteomic data from a sample of patients with many types of cancer. The TCGA project aims to evaluate the value of these large-scale, multidimensional analyses of the molecular characteristics of human cancer and provide data to the research community.

TCGA项目通过从许多癌症类型的患者样本中产生基因组、转录组、表观基因组和蛋白质组数据，TCGA项目旨在评估这些人类癌症分子特征（molecular characteristics）大规模多维分析（large-scale multidimensional analysis）的价值，并向研究界提供数据。

We collected DNA methylation, miRNA, mrna, and protein data from four cancers in TCGA. At the same time, we added SNF subtypes to each data type[16]. LASSO+cox and RSF models were used to predict survival of these data. , Use cindex to test the predictive power of the model. We expect to establish a good survival prediction model and find prognostic factors that influence the performance of the prognostic model to provide a basis for clinical practice.

我们采集TCGA中四种癌症的DNA甲基化、miRNA、mrna、protein数据，同时对每一种数据类型中加入SNF亚型[11]，分别采用LASSO+cox和RSF模型针对这些数据进行生存预测，使用cindex检验模型的预测能力。我们期望建立良好的生存预测模型，同时找到影响预后模型性能的预后因子，为临床实践提供依据。

Survival prediction analysis of cancer can help to identify important prognostic molecules. Hugo Gómez-Rueda and Emmanuel Martínez-Ledesma used three different algorithm(based on constrained particle swarm optimization (CPSO), network feature selection(NFS), and least absolute shrinkage and selection operator (LASSO)) to selection features and established survival prediction models for four cancers(BRCA、LUAD、OV、GBM) using the Cox proportional hazard model. It was found that the integration of genomic data produced a survival model that was slightly more than the single genome data[11].

对癌症的生存预测分析有助于发现重要预后分子，Hugo Gómez-Rueda, Emmanuel Martínez-Ledesma等人使用基于粒子群优化（CPSO）、网络特征选择和最小绝对收缩和选择算子（LASSO）三种不同的特征选择算法和Cox proportional hazard model对四种癌症（BRCA、LUAD、OV、GBM）建立生存预测模型，发现基因组数据的整合产生的生存模型在性能上略高于单一基因组数据[12]。

José M. Lezcano-Valverde, Fernando Salazar, and others established a survival prediction model for RA() using RSF. And they have identified potentially modifiable mortality risk factors. In addition, they established a model that could identify subtypes with higher mortality risk two years after RA diagnosis[12].

José M. Lezcano-Valverde、Fernando Salazar等人通过使用RSF对RA（）建立生存预测模型，确定了潜在的可修改的死亡率危险因素（modifiable mortality risk factors），为在疾病早期阶段对疾病进行彻底控制提供依据。另外他们建立的模型，可以在RA诊断两年后确定死亡风险较高的亚组[25]。

Zaixiang Tang and Yueping Shen propose new Bayesian hierarchical generalized linear models, called group spike-and-slab lasso GLMs, for predicting disease outcomes and detecting associated genes by incorporating large-scale molecular data and group structures[17].

## Results

### 评估各分子预后能力

For four types of cancer, breast cancer (BRCA), cervical cancer (CESC), endometrial cancer (UCEC), and ovarian cancer (OV), we obtained their clinical data from the TCGA database and downloaded four molecular data from firehose (Including DNA methylation, miRNA expression data, mrna expression data, RPPA data). In order to preserve a statistically sufficient samples size, we did not include the protein sample of CESC. A final set was established based on five types of cancer (four for CESC) data, in which BRCA contained 620 samples, CESC contained 291 samples, UCEC contained 425 samples, and OV contained 527 samples.

针对四种癌症，乳腺癌（BRCA），宫颈癌（CESC），子宫癌内膜癌（UCEC），卵巢癌（OV），我们从TCGA数据库中获取他们的clinical数据， 从firehose下载四种分子数据（包括DNA methylation、 miRNA表达数据、mrna表达数据、RPPA数据），为保持统计上足够多的样本，我们没有包含CESC的protein样本。分别根据癌症的五种（CESC是四种）数据建立最终集，其中BRCA包含620个样本，CESC包含291个样本，UCEC包含425个样本，OV包含527个样本。

**Table 1 Overview of Four Cancer Samples**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **cancer** | **overall survival** | **Methy** | **miRNA** | **mRNA** | **RPPA** | **result set** |
| BRCA | 1081 | 791 | 1079 | 1093 | 887 | 620 |
| CESC | 294 | 307 | 307 | 304 |  | 291 |
| UCEC | 544 | 444 | 550 | 557 | 440 | 425 |
| OV | 577 | 592 | 578 | 547 | 426 | 527 |

For each final set, we use 5 folds cross-validation to train the datasets 100 times and use the concordance index (C-index) to test the model prediction performance. C-index is between 0.5-1. C-index equal to 0.5 indicates no prediction ability, and 1 indicates complete prediction ability[18]. For the model, the first method chooses cox+LASSO to build the model. Because the cox model has no feature selection function, the LASSO method is used for feature selection, and then the cox proportional hazards model is used for modeling. The second method chooses the random survival forest (RSF) establishment (using the R package "randomForestSRC"). Because the RSF inherits the characteristics of the random forest, feature selection is performed automatically during model building.

对于每个最终集，我们采用5折交叉验证（5 fold cross-validation）分别将数据集训练100次，并使用concordance index（C-index）检验模型预测性能。C-index在0.5-1之间，c-index等于0.5表示没有预测能力，1表示有完全的预测能力[15]。对于模型，第一种方法选用cox+LASSO建立模型，由于cox模型没有特征选择的功能，使用LASSO方法进行特征选择,然后使用cox比例风险模型建模；第二种方法采用random survival forest（RSF）建立（使用R包“randomForestSRC”），由于RSF继承了RF的特点，在模型建立过程中自动进行特征选择。

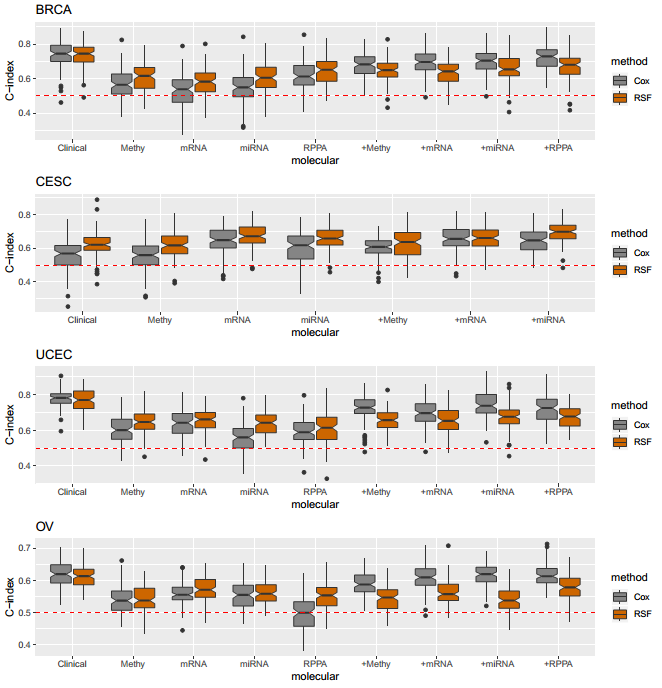


figure1 对每一种癌症（BRCA、CESC、UCEC、OV），根据临床变量、分子数据或者整合临床变量与分子数据训练模型获得的c-index，不同的颜色代表不同的模型；对于每一个癌症，将样本的80%用于训练模型，剩下的20%作为测试集，计算c-index。

In order to study whether the model built with the integration of molecular data and clinical variables was superior to the model constructed using only clinical variables, we combined the molecular data for each cancer separately with the clinical variables to built the model and calculated the c-index using the test data set. We found that proteomics and genomics in the CESC promoted the prognostic ability of the model. The model established by BRCA, UCEC, and OV using clinical variables has significant predictive power compared to the models established using only molecular data. For the CESC, a model built by integrating clinical variables and molecular data has a higher predictive power than models built using only clinical variables. In CESC, the predicted c-index of the cox model constructed by DNA methylation+clinical was 8.73% higher than that of the classical cox model (0.612 vs. 0.562, Wilcoxon signed rank test, P=1.95e-4.). In addition, the predicted c-index of the cox model constructed by miRNA+clinical increased by 15.03% (0.647 versus 0.562, Wilcoxon signed rank test, pvalue = 3.536e-10). The predicted c-index of the rsf model constructed by miRNA+clinical was increased by 11.3% (0.695 vs. 0.625, Wilcoxon signed rank test, pvalue = 1.913e-09) compared to the rsf model constructed using only the clinical data. The predicted c-index of the cox model constructed by mRNA+clinical increased by 16.6% (0.656 vs. 0.562, Wilcoxon signed rank test, pvalue = 4.506e-10) compared to the cox model using only the clinical data. The predicted c-index of the rsf model constructed by mRNA+clinical was increased by 5.9% (0.662 versus 0.625, Wilcoxon signed rank test, pvalue = 0.001019) compared to the rsf model constructed using only clinical.

为了研究整合分子数据与临床变量所构建的模型是否优于仅使用临床变量构建的模型，我们分别对每种癌症的分子数据分别于临床变量结合，构建模型并使用测试数据集计算c-index。我们发现蛋白组学与基因组学在CESC中，对于模型的预后能力有促进作用。BRCA、UCEC、OV使用临床变量建立的模型相对于仅适用分子数据建立的模型，有显著的预测能力。而对于CESC来说，整合临床变量与分子数据建立的模型比仅使用临床变量建立的模型有更高的预测能力。在CESC中，DNA methylation+clinical构建的cox模型预测的c-index相对于clinical构建的cox模型提高了8.73%（0.612对0.562，Wilcoxon符号秩和检验（Wilcoxon signed rank test，pvalue=1.95e-4）。miRNA+clinical构建的cox模型预测的c-index相对于仅使用clinical构建的cox模型，提高了15.03%（0.647对0.562，Wilcoxon符号秩和检验，pvalue= 3.536e-10），而miRNA+clinical构建的rsf模型的c-index相对于仅使用clinical构建的rsf模型，均值提高了11.3%（0.695对0.625，Wilcoxon符号秩和检验，pvalue= 1.913e-09）； mRNA+clinical构建的cox模型预测的c-index相比仅使用clinical构建的cox模型，提高了16.6%（0.656对0.562，Wilcoxon符号秩和检验，pvalue= 4.506e-10）。mRNA+clinical构建的rsf模型预测的c-index相比仅使用clinical构建的rsf模型，提高了5.9%（0.662对0.625，Wilcoxon符号秩和检验，pvalue= 0.001019）。

我们对比了各种情况构建模型的c-index，发现除CESC以外，BRCA, UCEC和OV的仅使用临床数据性能最高。BRCA的clinical的cox模型的c-index中位数为0.745158， rsf模型的c-index中位数为0.745261。UCEC的clinical的cox模型中位数为0.782898, rsf模型的中位数为0.770794. OV的clinical的cox模型c-index的中位数为0.619679. 如图2.1, 2.2, 2.3所示。

Fig 2.1

Fig 2.2

Fig 2.3

与仅使用分子数据构建模型相比，整合了临床数据的模型性能看起来更好。最显著的是UCEC（下图）。C-index提升了0.054-0.178，提升显著。

我们接着从UCEC的clinical+miRNA的cox模型中对miRNA进行筛选，得到以下结果：

|  |  |  |
| --- | --- | --- |
| **miRNA** | **Hazard ratio** | **wald's test Pvalue** |
| hsa-mir-106a | 1.00E+00 | 6.87E-03 |
| hsa-let-7g | 9.99E-01 | 3.74E-02 |
| hsa-mir-10b | 1.00E+00 | 7.97E-02 |
| hsa-mir-101-2 | 9.74E-01 | 2.73E-02 |
| hsa-mir-101-1 | 1.00E+00 | 1.28E-02 |
| hsa-mir-103-2-as | 5.84E+09 | 3.73E-06 |

### Biological insights from good prognostic models

For the model with prominent prognostic ability in Figure 1, we further studied the important molecular features included in each model to gain some mechanistic insights. The prognostic power of the clinical+miRNA model in cancer cesc was significantly higher than that of the clinical variable model alone. The important influencing factors with high predictive power in this model are shown in Table 2. Isobe T et al. concluded in the study that miR-142 affects the proliferation and invasion of cervical cancer cells and enhances apoptosis by directly targeting HMGB1 expression[19]. In the review of HILDA JIMÉNEZ-WENCES, miR-100 was down-regulated in patients with cervical cancer (cesc)[20]. miR-100, miR-99a showed a correlation with cervical cancer, and their down-regulation made the target Loss of function during cervical carcinogenesis[21, 22]. According to Longwen Shu's study, miR-204 plays an important role in the migration and invasion of cervical cancer by targeting TCF12[23].

对于图1中有突出预后能力的模型，我们进一步研究了每个模型中包含的重要分子特征以获得一些机制的见解。Cesc中clinical+miRNA模型的预后能力明显高于仅临床变量模型的预后能力，在该模型中具有高预测能力的重要影响因子如表2所示。Isobe T等人在研究中得出结论，miR-142通过直接靶向HMGB1的表达，影响宫颈癌细胞的增殖和侵袭能力，增强细胞凋亡[16];在HILDA JIMÉNEZ-WENCES等人的review中显示，miR -100在宫颈癌（cesc）患者样品中下调（Downregulated）[17]；miR -100、miR-99a显示与宫颈癌的相关性，他们的下调使得靶标在宫颈癌变过程中功能丧失[18-19]；Longwen Shu等人研究显示，miR-204通过靶向TCF12，在宫颈癌的迁移和侵袭中发挥重要作用[20]。

**表格2 CESC中基于clinical+miRNA所建立模型中重要影响因子**

|  |  |  |
| --- | --- | --- |
| **miRNA** | **Hazard ratio** | **wald's test Pvalue** |
| hsa-mir-142 | 0.9995 | 0.000048 |
| hsa-mir-100 | 1.0001 | 0.02327 |
| hsa-mir-147b | 1.096 | 0.001456 |
| hsa-mir-99a | 0.9996 | 0.0205 |
| hsa-mir-204 | 0.9098 | 0.03109 |
| hsa-mir-3074 | 0.9362 | 0.052164 |
| SNFtype | 1.373341 | 0.049652 |

For the clinical+methylation model, probes with high predictive methylation are shown in Table 3.1, and the corresponding gene list is obtained by predicting the gene corresponding to the probe in the NCBI file GPL13534, as shown in Table 3.2.

对于clinical+methylation模型，具有高预测能力methylation的探针如表3.1所示，通过NCBI中文件GPL13534预测探针对应的gene，获得相应gene list，如表3.2所示。

**表格3.1 CESC中基于clinical+methylation所建立模型中重要影响因子**

|  |  |  |
| --- | --- | --- |
| **methylation** | **Hazard ratio** | **wald's test Pvalue** |
| cg24424615 | 1.57275 | 0.001086 |
| cg13517138 | 1.60804 | 0.00025 |
| cg00146334 | 1.358303 | 0.069773 |
| cg03312426 | 0.701846 | 0.03852 |
| cg03552151 | 0.63272 | 0.04305 |
| cg02485922 | 1.51231 | 0.02181 |
| cg03629577 | 0.75602 | 0.0984 |
| cg17128799 | 1.37687 | 0.08092 |
| cg12283819 | 0.672003 | 0.007098 |
| cg25003147 | 0.59892 | 0.005881 |
| cg25951582 | 1.86005 | 0.00296 |
| cg02924534 | 1.50572 | 0.0163 |
| cg09052108 | 0.65515 | 0.011495 |
| SNFtype | 1.67518 | 0.00428 |

**表格3.2 methylation重要影响因子对应gene list**

|  |
| --- |
| **gene list** |
| PCDHB8 |
| PANK4 |
| RNF25 |
| C7orf61 |
| ATP6V1E1 |
| GPR157 |
| C8orf77 |
| RAPH1 |
| ZNF252 |
| DYNLRB1 |
| DLGAP5 |
| STK36 |
| DTWD1 |
| RRAS |
| C15orf33 |
| TBC1D10B |

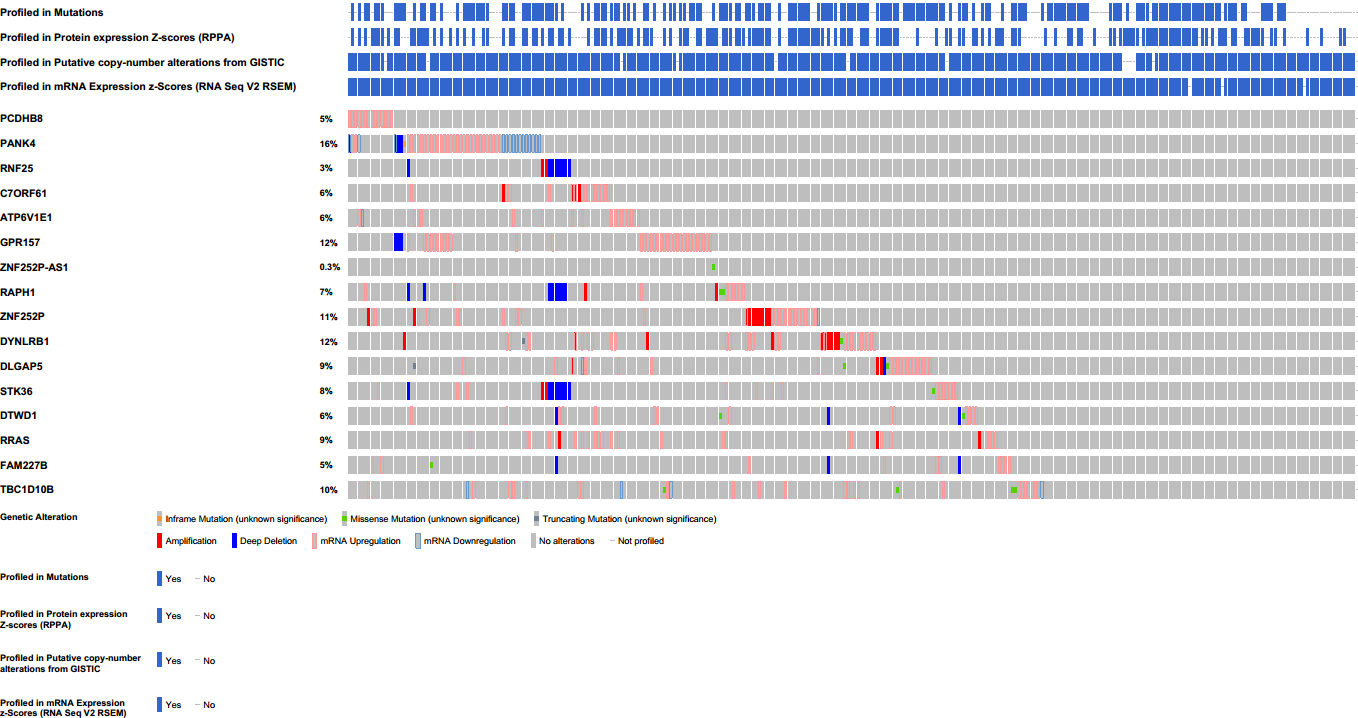
Get Disease as shown in Table 3.3:

**表3.3 甲基化相关基因富集分析Disease结果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ID** | **Name** | **Source** | **FDR B&H** | **Genes in Annotation** |
| 1 | C0545080 | Composite Lymphoma | 0.00222 | 1 |
| 2 | C0235598 | HODGKIN'S-LIKE | 0.002589 | 1 |
| 3 | C1301361 | Post-transplant lymphoproliferative disorder, polymorphic | 0.003328 | 1 |
| 4 | C0034888 | Rectal Prolapse | 0.003697 | 1 |
| 5 | C0265493 | Cat eye syndrome | 0.004067 | 1 |
| 6 | C1333171 | Primary Cutaneous Follicle Center Lymphoma | 0.004436 | 1 |
| 7 | C1709656 | Primary Cutaneous Diffuse Large B-Cell Lymphoma, Leg Type | 0.004804 | 1 |
| 8 | C0920296 | Developmental reading disorder | 0.005173 | 1 |
| 9 | C1519214 | Secondary Glioblastoma | 0.005542 | 1 |
| 10 | C0334581 | Gemistocytic astrocytoma | 0.006279 | 1 |

The methylation-related genes were analyzed using cBio, the cancer research type was selected as the squamous cell carcinoma and endocervical adenocarcinoma, the Patient/Case Set was selected for all types, and the gene list was input to obtain the gene list mutation map in the cancer CESC (Fig. 2). These genes have been mutated in 69% of patients. The results of mutual exclusivity were also shown in Table 3.5.

对甲基化相关基因使用cBio进行分析，癌症研究类型选择cervical squamous cell carcinoma and endocervical adenocarcinoma，Patient/Case Set选择所有类型，将gene list输入，得到gene list在癌症CESC中突变图谱（图2），69%病人中这些gene发生了突变。同时得到mutual exclusivity结果如表3.5所示。

****

**图2 CESC中甲基化相关基因的癌症图谱**

**表格3.5 甲基化相关基因mutual exclusivity结果**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Gene A** | **Gene B** | **Neither** | **A Not B** | **B Not A** | **Both** | **Log Odds Ratio** | **p-Value** | **Adjusted p-Value** | **Tendency** |
| RNF25 | STK36 | 285 | 0 | 15 | 10 | >3 | <0.001 | <0.001 | Co-occurrence |
| RNF25 | RAPH1 | 286 | 3 | 14 | 7 | >3 | <0.001 | <0.001 | Co-occurrence |
| RAPH1 | STK36 | 272 | 13 | 17 | 8 | 2.287 | <0.001 | 0.007 | Co-occurrence |
| C7ORF61 | RRAS | 273 | 10 | 20 | 7 | 2.257 | <0.001 | 0.022 | Co-occurrence |
| PANK4 | GPR157 | 237 | 35 | 24 | 14 | 1.374 | <0.001 | 0.067 | Co-occurrence |
| DYNLRB1 | RRAS | 254 | 29 | 18 | 9 | 1.477 | 0.002 | 0.269 | Co-occurrence |

In the clinical+mRNA model, important influencing factors with high predictive power are shown in Table 4. In Rong Zhang’s study, the expression of the exo-matrix gene ANGPTL4 in cervical cancer cells was significantly reduced[24]. The protein iASPP encoded by PPP1R13L was considered as an oncogene that not only inhibits the transcriptional activity of p53 on promoters of downstream genes, but also promotes carcinogenesis through p53-independent mechanisms[25]. The SHP-2 gene encoded by the gene PTPN11 may play an important role in HPV infectious diseases such as cervical cancer[26].

在模型clinical+mRNA中，具有高预测能力的重要影响因子如表4所示。在Rong Zhang等人研究中表明，外基质基因ANGPTL4在宫颈癌细胞中表达显著降低[21]；由PPP1R13L编码的蛋白质iASPP 被认为不仅抑制下游基因启动子上p53转录活性的癌基因，而且还通过p53非依赖性机制促进癌发生[22]（iASPP was considered as an oncogene that not only inhibits the transcriptional activity of p53 on promoters of downstream genes (Figure ​(Figure1a),1a), but also promotes carcinogenesis through p53-independent mechanisms）;由基因PTPN11编码的SHP-2基因在宫颈癌等HPV感染性疾病中可能发挥重要作用[23]。

**表格4 CESC中基于clinical+mRNA所建立模型中重要影响因子**

|  |  |  |
| --- | --- | --- |
| **mRNA** | **Hazard ratio** | **wald's test Pvalue** |
| mRNA\_ANGPTL4|51129 | 1.5530815 | 0.00169 |
| mRNA\_PPP1R13L|10848 | 1.342358 | 0.0934 |
| mRNA\_PTPN11|5781 | 1.515392 | 0.00305 |
| mRNA\_FAM23A|653567 | 1.20567 | 0.0369 |
| mRNA\_ERGIC3|51614 | 1.353676 | 0.0978 |

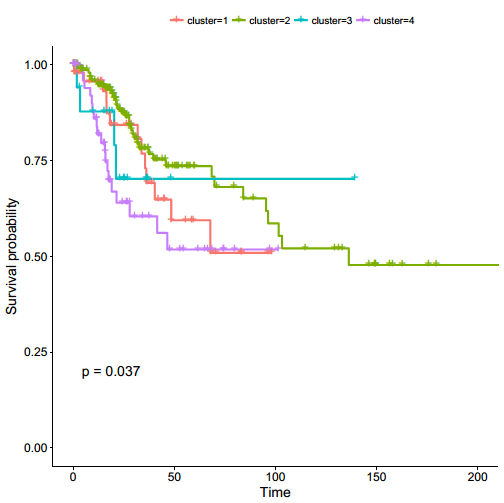


Figure3a CESC患者通过SNF亚型聚类的Kaplan-Meier图

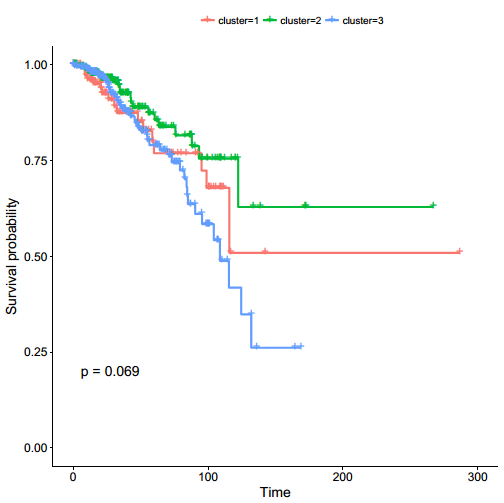


Figure3b BRCA患者通过SNF亚型聚类的Kaplan-Meier图

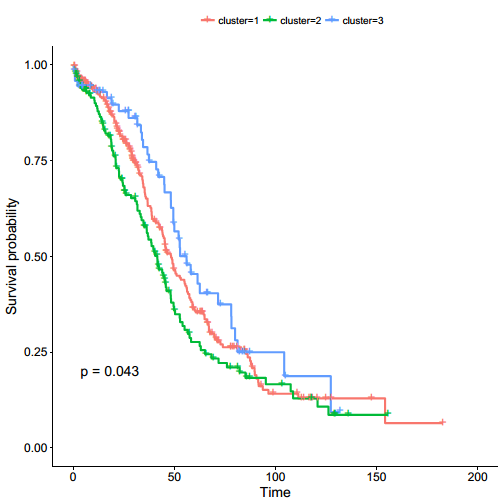


Figure3c OV患者通过SNF亚型聚类的Kaplan-Meier图

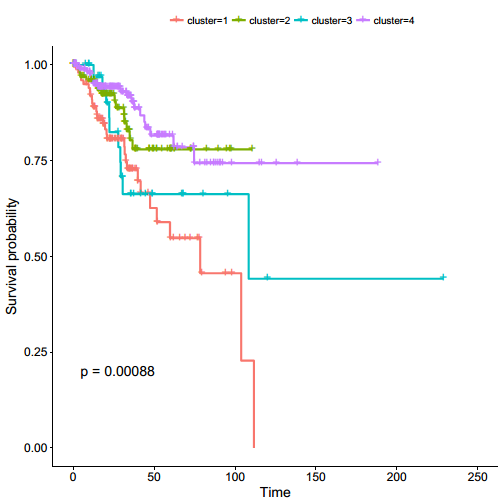


Figure3d UCEC患者通过SNF亚型聚类的Kaplan-Meier图

It is noteworthy that the SNFtools subtypes in cervical cancer models are important factors. In fact, SNF subtypes showed different survival patterns in cervical cancer (log-rank test, Fig. 3a, P=0.037), while SNF subtypes also showed different survival patterns in other cancer types (BRCA, Fig. 3b, log-rank test, P=0.069; OV, Fig. 3c, log-rank test, P=0.043; UCEC, FIG. 3d, log-rank test, P=0.00088).

值得注意的是，SNFtools亚型在癌症CESC中所建立的clinical+miRNA和clinical+methylation是有重要影响的预测因子。事实上，SNF亚型在CESC中显示出不同的存活模式（log-rank检验，图2a，P=0.037），同时，SNF亚型在其他癌症类型中也显示出不同的存活模式（BRCA，图2b，log-rank检验，P=0.069；OV，图2c，log-rank检验，P=0.043；UCEC，图2d，log-rank检验，P=0.00088）。

## DISCUSSION

Compared with previous studies driven by a single cancer type or a single data type, we systematically evaluated survival predictions for four cancer patients from different molecular data types, while integrating SNF subtype data into molecular data and extracting potentially important predictors are described in well-performing models. These potentially important predictors contribute to clinical application.

与之前由单一癌症类型或单一数据类型驱动的研究相比，我们系统地评估了来自不同分子数据类型的四种癌症患者的生存预测，同时将SNF亚型数据整合到分子数据中，并且提取了在性能良好模型中潜在重要预测因子加以描述，这些潜在重要预测因子有助于进行临床应用。

By establishing survival prognostic models of different molecular data types for cancers that are frequently occurring in the female population (including breast cancer, cervical cancer, endometrial cancer, and ovarian cancer), we determined several models have better performance than models using only clinical data. It is worth noting that, based on the model established using only clinical data alone, CESC has the lowest average performance of c-index (0.562), which is related to the least sample (294).

通过对在女性人群中常发的癌症（包括乳腺癌、宫颈癌、子宫内膜癌、卵巢癌）建立不同分子数据类型的生存预后模型，我们确定了几个相比仅使用临床数据所建立模型性能更佳的例子（如CESC的clinical+methylation，clinical+mRNA，clinical+miRNA）。值得注意的是，单就仅使用临床数据所建立模型来看，CESC拥有c-index平均最低的性能（0.562），这与其拥有最少的样例（294）存在一定关系。

而BRCA、OV、UCEC中，clinical+分子的模型与仅使用分子数据的模型相比，普遍拥有更好的预后能力，这是因为clinical数据是更高阶的特征，是更重要的特征，在clinical的参与下，模型会产生更好的性能。

同时我们还发现，当仅使用分子数据构建模型的时候，RSF的性能一般要高于cox模型。而使用整合分子数据与临床数据构建模型时，RSF的性能一般低于cox。这是因为仅使用分子数据时，没有经过临床数据的相关性选择，RSF模型本身继承随机森林的特点，对分子特征的选择有随机过程，所以在这种情况下有优势。而clinical参与的情况下，与clinical相关的分子数据会更倾向于保留，在这种情况下cox更加有优势。

We further studied the model with better performance and extracted important prognostic factors in the established cox model. We found that these important prognostic factors were strongly correlated with the corresponding cancer in the previous literature. This shows that the data information we used to build the models is useful information. With the identification of these important influence genes, prospective genomic analysis may provide individualized treatment options for patients with metastatic or local disease.

我们对性能更佳模型更进一步研究，对所建立cox模型过程LASSO选择预后因子进行提取，发现这些重要预后影响因子在之前文献中与所对应的癌症存在很强相关关系，这说明我们构建模型过程所使用数据信息是有用信息。随着鉴定出这些重要影响力基因，前瞻性基因组分析可能会为转移性或局部疾病患者提供个体化的治疗方案。

Although our study provides insights that use biological data for survival prediction to translate into clinical use, it still has some limitations. First, we simply use data mining methods to model, while ignoring some of the candidate features that can be driven by prior knowledge. second, because of the extensive collinearity of large-scale biological data, effectively combining multiple types of molecules data still has challenges. Therefore, an important research direction in the future is to establish a prognostic model that effectively integrates multiple types of data. Finally, in order to conduct comprehensive molecular analysis, the patient samples used are derived from multiple source sites, which may cause heterogeneity.

尽管我们的研究提供了利用生物数据进行生存预测以转化为临床使用的见解，但它依然存在一些局限性。首先，我们单纯地采用数据挖掘的方法进行建模，而忽略了一些可以通过先验知识驱动获得的候选特征；其次，因为大规模生物数据存在广泛的共线性，有效地结合多种类型的分子数据仍然存在挑战。因此，未来一个重要的研究方向是建立有效融合多种类型数据的预后模型；最后，为了进行全面的分子分析，所使用的的患者样本来源于多种来源站点，这可能会引起异质性。

## Method

### 样本集的获取

We obtained DNA methylation, miRNA, MRNA, and RPPA data from four cancers: BRCA, CESC, UCEC, and OV from firehose (http://firebrowse.org). At the same time, clinical data of four cancer patients were obtained from the TCGA Data Portal (https://portal.gdc.cancer.gov/repository). The sources of these data are shown in the table.

|  |  |  |
| --- | --- | --- |
| cancer | molecular data | platform |
| BRCA | DNA methylation | Illumina Infinium Human DNA Methylation 450K |
| miRNA | Illumina Genome Analyzer/HiSeq 2000 miRNA sequencing platform |
| mRNA | Illumina HiSeq 2000 RNA Sequencing V2 |
| RPPA | MD Anderson Reverse Phase Protein Array (RPPA) Core platform |
| CESC | DNA methylation | Illumina Infinium Human DNA Methylation 450K |
| miRNA | Illumina Genome Analyzer/HiSeq 2000 miRNA sequencing platform |
| mRNA | Illumina HiSeq 2000 RNA Sequencing V2 |
| UCEC | DNA methylation | Illumina Infinium Human DNA Methylation 450K |
| miRNA | Illumina Genome Analyzer/HiSeq 2000 miRNA sequencing platform |
| mRNA | Illumina HiSeq 2000 RNA Sequencing V2 |
| RPPA | MD Anderson Reverse Phase Protein Array (RPPA) Core platform |
| OV | DNA methylation | Illumina Infinium Human DNA Methylation 27K |
| miRNA | Agilent 8 × 15K Human miRNA-specific microarray platform |
| mRNA | Agilent 244K Custom Gene Expression G4502A |
| RPPA | MD Anderson Reverse Phase Protein Array (RPPA) Core platform |

For DNA methylation, the platform is Illumina Infinium Human DNA Methylation 450K (BRCA, CESC and UCEC) and Illumina Infinium Human DNA Methylation 27K (OV). We have retained the most negatively related probes for gene expression according to firebrowse. For miRNA, the platform is Illumina Genome Analyzer/HiSeq 2000 miRNA sequencing platform (BRCA, CESC and UCEC) and Agilent 8 × 15K Human miRNA-specific microarray platform (OV). For mRNA, the platform is Illumina HiSeq 2000 RNA Sequencing V2 (BRCA, CESC and UCEC) and Agilent 244K Custom Gene Expression G4502A (OV). For RPPA, the platform is Agilent 244K Custom Gene Expression G4502A (BRCA, UCEC and OV). For each cancer, we took each molecular data and clinical data according to their samples to obtain a result data set for each type of data.

### SNF亚型的获取

The SNF method uses a sample network as a basis for integration and consists of two steps: building a sample similarity network for each data type and integrating these networks into a single similarity network using a non-linear combination approach. We used SNFtools (http://compbio.cs.toronto.edu/SNF/SNF/Software.html) to establish SNF subtypes for four cancers and then added them to molecular data to build a survival prediction model. 、

### 模型介绍

（ref：<http://www.sthda.com/english/wiki/cox-proportional-hazards-model>）

（Cox, David R. "Regression models and life-tables." Breakthroughs in statistics. Springer, New York, NY, 1992. 527-541.）

Cox模型是由D.R.Cox提出的回归模型，它可以同时评估几个因素对生存的影响。该模型预测变量通常称为协变量。Cox模型由危险函数h(t)表示，h(t)可以解释为在时间t死亡的风险，它的表达式如下：

h(t)=h0(t)×exp(b1x1+b2x2+...+bpxp)

其中，t表示生存时间，h(t)是由一组因变量(x1, x2, …, xp)确定的危险函数，偏系数(b1, b2, …, bp)估计因变量对协变量的影响， h(0)称为baseline hazard，它对应与所有xi等于0时的危险率。

### Model training and c-index calculation



For each final data set, Cox proportional hazards and RSF algorithms were used to train. First, the samples were randomly divided into two groups, one group accounting for 80% as training data, another one group accounting for 20% as test data. We use two models to train the training data set. The first method is LASSO+Cox, which first uses the LASSO method in the “glmnet” package in the R software for feature selection, and then uses the coxph function in the "survival" package to train the cox model. The second method is random survival forest uses the “randomForestSRC” package in the R software to train the RSF model. We set the number of leaf nodes to 1000, so as to produce a stable result. Then, we use these two models to predict the test data set and calculate the c-index. For each result set, we performed 100 trainings and c-index calculations. Finally, we selected a model that performed well for each molecular data and clinical data. Wilcoxon signed rank test was used to calculate P values ​​to compare differences between C-indexes.

In order to integrate clinical variables and molecular data to build survival prediction models, we combined patient clinical data with DNA methylation, miRNA, mRNA, and RPPA data, and then constructed LASSO+cox and RSF models as before, and calculated the corresponding C-index.

## Reference

1. Merlo, D.F., et al., *Breast cancer incidence trends in European women aged 20-39 years at diagnosis.* Breast Cancer Res Treat, 2012. **134**(1): p. 363-70.

2. Ghiasvand, R., et al., *Higher incidence of premenopausal breast cancer in less developed countries; myth or truth?* BMC Cancer, 2014. **14**: p. 343.

3. Peccatori, F.A., et al., *Biology, staging, and treatment of breast cancer during pregnancy: reassessing the evidences.* Cancer Biol Med, 2018. **15**(1): p. 6-13.

4. Waggoner, S.E., *Cervical cancer.* Lancet, 2003. **361**(9376): p. 2217-25.

5. Marquina, G., A. Manzano, and A. Casado, *Targeted Agents in Cervical Cancer: Beyond Bevacizumab.* Curr Oncol Rep, 2018. **20**(5): p. 40.

6. Sales, K.J., *Human papillomavirus and cervical cancer*, in *Cancer and Inflammation Mechanisms: Chemical, Biological, and Clinical Aspects*. 2014, John Wiley & Sons. p. 165-180.

7. Siegel, R., et al., *Cancer statistics, 2014.* CA: a cancer journal for clinicians, 2014. **64**(1): p. 9-29.

8. Lee, J.Y., et al., *Trends in gynecologic cancer mortality in East Asian regions.* J Gynecol Oncol, 2014. **25**(3): p. 174-82.

9. Wu, Q.J., et al., *Cancer incidence among adolescents and young adults in urban Shanghai, 1973-2005.* PLoS One, 2012. **7**(8): p. e42607.

10. Kim, K., et al., *Current status of gynecological cancer in China.* J Gynecol Oncol, 2009. **20**(2): p. 72-6.

11. Gomez-Rueda, H., et al., *Integration and comparison of different genomic data for outcome prediction in cancer.* BioData Min, 2015. **8**: p. 32.

12. Lezcano-Valverde, J.M., et al., *Development and validation of a multivariate predictive model for rheumatoid arthritis mortality using a machine learning approach.* Sci Rep, 2017. **7**(1): p. 10189.

13. Zhu, B., et al., *Integrating Clinical and Multiple Omics Data for Prognostic Assessment across Human Cancers.* Sci Rep, 2017. **7**(1): p. 16954.

14. Treppmann, T., K. Ickstadt, and M. Zucknick, *Integration of Multiple Genomic Data Sources in a Bayesian Cox Model for Variable Selection and Prediction.* Comput Math Methods Med, 2017. **2017**: p. 7340565.

15. Yuan, Y., et al., *Assessing the clinical utility of cancer genomic and proteomic data across tumor types.* Nat Biotechnol, 2014. **32**(7): p. 644-52.

16. Wang, B., et al., *Similarity network fusion for aggregating data types on a genomic scale.* Nat Methods, 2014. **11**(3): p. 333-7.

17. Tang, Z., et al., *Group spike-and-slab lasso generalized linear models for disease prediction and associated genes detection by incorporating pathway information.* Bioinformatics, 2018. **34**(6): p. 901-910.

18. Harrell, F.E., Jr., K.L. Lee, and D.B. Mark, *Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.* Stat Med, 1996. **15**(4): p. 361-87.

19. Jiang, D., et al., *MiR-142 inhibits the development of cervical cancer by targeting HMGB1.* Oncotarget, 2017. **8**(3): p. 4001-4007.

20. Jimenez-Wences, H., O. Peralta-Zaragoza, and G. Fernandez-Tilapa, *Human papilloma virus, DNA methylation and microRNA expression in cervical cancer (Review).* Oncol Rep, 2014. **31**(6): p. 2467-76.

21. Chen, D., et al., *MicroRNA-99 family members suppress Homeobox A1 expression in epithelial cells.* PLoS One, 2013. **8**(12): p. e80625.

22. Granados-Lopez, A.J., et al., *Use of Mature miRNA Strand Selection in miRNAs Families in Cervical Cancer Development.* Int J Mol Sci, 2017. **18**(2).

23. Shu, L., Z. Zhang, and Y. Cai, *MicroRNA-204 inhibits cell migration and invasion in human cervical cancer by regulating transcription factor 12.* Oncol Lett, 2018. **15**(1): p. 161-166.

24. Zhang, R., et al., *E6/E7-P53-POU2F1-CTHRC1 axis promotes cervical cancer metastasis and activates Wnt/PCP pathway.* Sci Rep, 2017. **7**: p. 44744.

25. Dong, P., et al., *Reactivating p53 functions by suppressing its novel inhibitor iASPP: a potential therapeutic opportunity in p53 wild-type tumors.* Oncotarget, 2015. **6**(24): p. 19968-75.

26. Tao, X.H., et al., *Significance of SHP-1 and SHP-2 expression in human papillomavirus infected Condyloma acuminatum and cervical cancer.* Pathol Oncol Res, 2008. **14**(4): p. 365-71.