Advances in Applied Mathematics 应用数学进展, 2023, 12(7), 3398-3413
Published Online July 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.127338

Hans汉斯

使用特征选择流程和机器学习分类模型预测肾细胞癌亚型

魏嘉怡

青岛大学数学与统计学院，山东 青岛

收稿日期：2023年6月25日；录用日期：2023年7月19日；发布日期：2023年7月28日

摘　要

肾细胞癌是常见且致命的疾病，占肾癌的绝大多数。肾细胞癌是一种异质性的复杂疾病，主要由三种组织学亚型组成，存在不同的生物学和临床差异。如今，科技的发展能够得到肾细胞癌的分子亚型和生物标志物。在这项研究中，我们将三种不同的特征选择技术进行组合，即mRMR、Lasso、Boruta，利用投票法的思路从TCGA多个单组学数据集中选择最显著的特征，并将其作为基础机器学习模型的输入，用于肾细胞癌组织学亚型分类。我们评估了六种不同的分类模型，包括逻辑回归(LR)、随机森林(RF)、支持向量机(SVM)、朴素贝叶斯(NB)、k-最近邻(KNN)和XGBoost。结果表明，基于应用本文的新特征选择流程，miRNA成熟链表达RNAseq数据集提供的特征在准确性方面优于其他分类方法，在逻辑回归模型下能达到0.9779的准确率与0.9834的AUC，取得了最高性能。因此，我们改进和细化的特征选择和分类提供了诊断标志物，可能有助于提高诊断的准确性，从而帮助设计早期治疗策略，提高肾细胞癌患者的生存率。

关键词

肾细胞癌，亚型分类，特征选择，机器学习

Using Feature Selection Process and Machine Learning Classification Models to Predict Subtypes of Renal Cell Carcinoma

Jiayi Wei

School of Mathematics and Statistics, Qingdao University, Qingdao Shandong

Received: Jun. 25th, 2023; accepted: Jul. 19th, 2023; published: Jul. 28th, 2023

文章引用: 魏嘉怡. 使用特征选择流程和机器学习分类模型预测肾细胞癌亚型[J]. 应用数学进展, 2023, 12(7): 3398-3413. DOI: 10.12677/aam.2023.127338

Abstract

Renal cell carcinoma is a common and fatal disease, accounting for the majority of kidney cancers. There are three basic histological subtypes of renal cell carcinoma, each of which has unique biological and clinical characteristics, which represents a complex and heterogeneous ailment. The availability of molecular subtypes and biomarkers for renal cell carcinoma is made possible by modern technological advancements. In this study, we combined three different feature selection techniques, namely mRMR, Lasso, and Boruta, using the idea of voting method to select the most significant features from multiple single-omics datasets of TCGA and use them as input to a base machine learning model for histological subtype classification of renal cell carcinoma. We evaluated six classification models, including logistic regression (LR), random forest (RF), support vector machine (SVM), naive Bayes (NB), k-nearest neighbor (KNN), and XGBoost. The results demonstrate that the features from the miRNA mature strand expression RNAseq dataset outperformed other classification methods based on the application of the new feature selection process in this paper, achieving the highest performance with an accuracy of 0.9779 and an AUC of 0.9834 under the logistic regression model. As a result, the enhanced and refined feature selection and categorization offer diagnostic indicators that could help increase the accuracy of diagnosis, aid in the development of early treatment plans, and enhance the survival of patients with renal cell carcinoma.

Keywords

Renal Cell Carcinoma, Subtype Classification, Feature Selection, Machine Learning

1. 引言

肾细胞癌(RCC)是肾小管上皮细胞的异质性癌症，占肾脏癌症的 90%以上，一种常见且致命的疾病[1] [2]。根据世界卫生组织的最新数据，每年肾癌新增约 430,000 例，死亡近 180,000 例[3]。肾细胞癌由许多不同的癌症亚型组成，每个亚型都有不同的组织学、不同的遗传和分子改变、不同的临床病程以及不同的治疗反应[4]。RCC 组织学亚型主要包括以下三类：肾透明细胞癌(ccRCC)、乳头状肾细胞癌(pRCC)和嫌色细胞癌(chRCC)，其中肾透明细胞癌是最为常见的亚型，约占肾细胞癌的 75%，乳头状肾细胞癌占 15%~20%，嫌色肾细胞癌约占 5% [2]。组织病理学评估是诊断 ccRCC、pRCC 和 chRCC 的金标准。精确判别并了解各个亚型之间的差异对于改进患者的管理和治疗至关重要。

随着基因组学和转录组学技术的快速发展，越来越多的基因组数据被收集起来供人研究。癌症基因组图谱(TCGA, https://tcga-data.nci.nih.gov/tcga/)是一个里程碑式的癌症基因组学项目，它旨在描述驱动不同癌症的遗传特征[5]。TCGA 包含不同类型的组学数据，包括转录组数据(mRNA、lncRNA 和 miRNA)、基因组数据(突变、CNV)、表观基因组数据(DNA 甲基化)、蛋白质组数据和临床信息数据[5] [6]。许多研究仅采用全向组学数据进行癌症分类研究，最常见的是基因表达数据[7]。TCGA 中包含三个关于肾脏癌症的项目，并分配了四个字母的代码以反映其组织学定义：KICH (肾嫌色细胞癌)、KIRC (肾透明细胞癌)和 KIRP (肾乳头状细胞癌)。

在这项研究中，我们旨在充分利用 TCGA 数据，获得能够准确区分肾细胞癌进行亚型的特征，以确

定共有的和亚型特异性分子特征，这将为开发 RCC 的疾病特异性治疗方法和预后生物标志物提供基础。我们改进的分类可以带来更准确的诊断和更适合患者的护理，本文流程也可以应用于其他癌症或疾病。

2. 材料与方法

2.1. 患者数据

肾细胞癌组学数据来自 TCGA 中的肾嫌色细胞癌、肾透明细胞癌以及肾乳头状细胞癌三个队列的组合(KICH，KIRC，KIRP，https://tcga.xenahubs.net)。为了对肾细胞癌进行分类，我们考虑了包括基因级拷贝数变异、外显子表达 RNA 测序、甲基化数据、基因表达 RNA 测序和 miRNA 成熟链表达 RNA 测序在内的五种单组学数据集的数据来进行探究。我们在下文中用 copynumber、exon RNA-seq、methylation、gene RNA-seq 和 miRNA 的缩写来指这五个数据集。生存分析数据由表型数据集提供。

Table 1. Original samples and feature counts for each data set of renal cell carcinoma
表 1. 肾细胞癌各个数据集原始样本与特征数

队列	数据集	样本量	特征量
KICH	copynumber	66	24,776
	exon RNA-seq	91	239,322
	methylation	66	485,577
	gene RNA-seq	91	20,530
	miRNA	89	1917
KIRC	copynumber	528	24,776
	exon RNA-seq	606	239,322
	methylation	480	485,577
	gene RNA-seq	606	20,530
	miRNA	311	2048
KIRP	copynumber	288	24,776
	exon RNA-seq	323	239,322
	methylation	321	485,577
	gene RNA-seq	323	20,530
	miRNA	321	2114

五种单组学数据各自进行亚型赋值，新创建“subtype”列并将“KICH”、“KIRC”、“KIRP”三种亚型分别赋予 0、1、2。然后我们再对其进行数据合并，得到表 1 所述的五个合并后的数据集。接下来分别对其进行去除缺失值、均值插补和标准化，以确定合并数据是否有利于肾细胞癌的组织学亚型分类。其中，每个数据集的 20%的数据被留作独立的测试集，80%的数据被用于训练。

2.2. 数据预处理与特征选择

当数据规模有限但维度较高时，过拟合等问题容易导致分类精度降低。因此，选择合适的特征选择方法将使我们更有效地分析数据。在本文中，首先删除缺失超过 30%的列。接下来每个数据集的低方差的特征(methylation450k 的方差 < 0.05，copynumber、miRNA 与 gene RNA-seq 的方差 < 0.5，exon RNA-seq 的方差 < 1)被剔除，因为它们对分类的作用不大。我们同时考虑了各个特征与组织学亚型的相关性，研究发现，去除亚型相关性低的变量(所有数据集的 spearman 相关系数的绝对值 < 0.1)对本文的研究产生了

些许的提升。

接下来对数据集进行数据标准化，为了特征选择的鲁棒性，我们同时选择 LASSO [8]、mRMR (Max-Relevance and Min-Redundancy) [9]和 Boruta [10]进行特征选择。当某一特征在一次交叉验证中被三种特征选择方法同时选中两次或两次以上时，则认为该特征对肾细胞癌的亚型分类有意义。5 折交叉验证经过上述程序后都能够选出的特征即是最终的特征，并将其输入到多个机器学习模型中，判断哪个数据集最能将肾细胞癌的亚型精确分类。虽然原始特征的数量和每个数据集所包含的信息量差别很大，但为了更加明确地看出哪个数据集最能对肾细胞癌亚型分类提供信息，我们特意固定了交叉验证中每次 mRMR 选择的特征数量(所有数据集均选择 50 个特征)。每个数据集的最终特征显示在表 3 中。

2.3. 机器学习模型

如上所述，我们考虑将交叉验证重叠的特征作为机器学习模型的输入。在本文中，我们将平均准确率和平均 AUC 作为判断亚型预测好坏的标准。其中，准确率是通过预测结果与表型数据集中的表达亚型的一致性来评估的，AUC 定义为 ROC 曲线下坐标轴所包围的区域。观察不同交叉验证下的准确率和 AUC 的标准差，有助于判断该输入特征下的机器学习模型是否稳健。

在这项研究中，我们选择六种机器学习模型：逻辑回归(LR)、随机森林(RF)、支持向量机(SVM)、朴素贝叶斯(NB)、k 近邻(KNN)和 XGBoost。这些基础的机器学习模型被应用于根据选定的特征估计肾细胞癌的组织学亚型，以确定本文的特征选择方法是否有助于癌症亚型的分类。所有代码都可以通过 python 轻松实现。

3. 数据可用性

支持本研究结果的样本数据由 TCGA 中的 KICH，KIRC，KIRP 中公开提供(https://tcga.xenahubs.net)。

4. 结论

4.1. 特征选择

将 KICH、KIRC、KIRP 三种亚型合并，分别得到 copynumber、exon RNA-seq、methylation、gene RNA-seq 和 miRNA 五个总数据集，具体样本与特征数目如表 2 所示，此处只选择肾细胞癌亚型分类准确率与 AUC 最高的一组数据进行详细说明，即 miRNA 数据集。我们得到 721 个样本，共有 1847 个特征，是合并后的五个数据集中样本与特征数量最少的数据集。其中每个亚型的数量虽有一定差别，但仍较为平衡，不需要考虑过采样或欠采样：89 个 KICH，311 个 KIRC，321 个 KIRP。基于训练集与测试集的划分比例，在 5 折交叉验证中，每次交叉验证产生 577 个训练集和 144 个测试集。为了发现对准确划分亚型提供帮助的特征，近年来广泛采用三种特征选择技术：mRMR、LASSO 和 Boruta。

Table 2. Merged datasets
表 2. 合并后的数据集

数据集		样本量		特征量
copynumber	KICH	66	882	24,776
	KIRC	528		
	KIRP	288		
exon RNA-seq	KICH	91	1020	239,322
	KIRC	606		
	KIRP	323		

Continued

methylation	KICH	66	867	485,577
	KIRC	480		
	KIRP	321		
gene RNA-seq	KICH	91	1020	20,530
	KIRC	606		
	KIRP	323		
miRNA	KICH	89	721	1847
	KIRC	311		
	KIRP	321		

使用方法章节中讨论过的特征选择过程，我们最终从 miRNA 数据集中得到了 77 个特征，如表 3 所示。其余数据集所得特征与数量也于表 3 所示。各数据集选择出的特征的表达水平以及热力图由附录中的附图 1 与附图 2 展示，我们可以看出，使用本文特征选择流程挑选出的特征在三种亚型中差异明显，也可以反映出特征对肾细胞癌亚型分类有益处。

Table 3. Features extracted from each dataset
表 3. 各数据集提取出的特征

数据集	特征	数量	原特征数量
copynumber	/	0	24,776
exon RNA-seq	chr13:28959688-28964241: −, chr19:8145887-8146003: −, chr15:57976600-57977560: +, chr19:55672008-55672143: −, chr19:46093025-46095955: −, chr1:79355451-79356901: −, chr1:79357209-79357377: −, chr1:79358783-79358874: −, chr1:79383319-79383385: −, chr17:74010507-74010657: −, chr18:12658738-12658772: +, chr4:55944427-55946330: −, chr5:138784245-138784597: −, chr1:163112098-163117293: −, chr13:28874483-28877505: −, chr5:138784767-138784865: −, chr12:57630107-57630464: −, chr12:57629532-57629592: −, chr16:83828628-83830197: +, chr5:52243175-52243291: +, chr6:32002291-32002374: +, chr6:5998235-5999437: −, chr9:1049858-1050775: +, chr9:1051535-1052138: +	24	239,322
methylation	cg22078451, cg24922129, cg23856138, cg11840467, cg11191368, cg22301128, cg04935121, cg09645818, cg05141014, cg05256605, cg05425699, cg09548780, cg00204802, cg09468836, cg08897188, cg06428163, cg24499411, cg08559364, cg08435683, cg08223003, cg06756211, cg06786219, cg22024120, cg07479030, cg02670123, cg15431659, cg01020475, cg01881182, cg19009372, cg03830585, cg26971042, cg19062108, cg16929739, cg16026813, cg15266205, cg15077193, cg20176142, cg14078335, cg13740187, cg12785694, cg17444849	41	485,577
gene RNA-seq	LDB2, ECSCR, MAPRE2, CDH13, PLVAP, CDH5, MYCT1, ERG, PPP1R3C, RGS5, ELTD1, PYY, PRICKLE2, EDNRA, CXorf36, GPR4, NR5A2, CYYR1, LRRTM1, PCSK6, CD34, KDR, GGT1, AKR1C3, GALNT14, PCDH12, FLT1, TMEM176B, PCDH17, BCL6B, C11orf53, KLK15, MAFA, TCF4, KRT6C, C3orf70, GIMAP6, C6orf223, C8orf71, PSKH2	40	20,530

Continued

miRNA	MIMAT0000063, MIMAT0004748, MIMAT0004571, MIMAT0000461, MIMAT0004585, MIMAT0004589, MIMAT0000449, MIMAT0000445, MIMAT0000444, MIMAT0004598, MIMAT0000437, MIMAT0004599, MIMAT0000435, MIMAT0004601, MIMAT0000432, MIMAT0004613, MIMAT0000430, MIMAT0004615, MIMAT0003150, MIMAT0004569, MIMAT0004568, MIMAT0000681, MIMAT0001341, MIMAT0002874, MIMAT0002820, MIMAT0002809, MIMAT0003321, MIMAT0002174, MIMAT0004482, MIMAT0001343, MIMAT0004485, MIMAT0004553, MIMAT0000763, MIMAT0004494, MIMAT0004499, MIMAT0004500, MIMAT0004514, MIMAT0000689, MIMAT0004552, MIMAT0004703, MIMAT0000423, MIMAT0015045, MIMAT0000084, MIMAT0016895, MIMAT0017982, MIMAT0000095, MIMAT0000089, MIMAT0017985, MIMAT0000087, MIMAT0000086, MIMAT0017990, MIMAT0005951, MIMAT0000077, MIMAT0000076, MIMAT0000075, MIMAT0018073, MIMAT0019731, MIMAT0022925, MIMAT0026472, MIMAT0000242, MIMAT0002888, MIMAT0004911, MIMAT0005796, MIMAT0000279, MIMAT0004809, MIMAT0004909, MIMAT0000270, MIMAT0000245, MIMAT0004945, MIMAT0000265, MIMAT0000262, MIMAT0005825, MIMAT0000255, MIMAT0000254, MIMAT0000318, MIMAT0000250, MIMAT0004902	77	1847

4.2. 机器学习模型结果

为了评估本文的数据整合和特征选择过程是否有助于肾细胞癌的亚型分类，我们对合并后的各个数据集进行 5 折交叉验证，将上述特征选择流程得到的重要特征放入六个基础的机器学习模型中，并观察其最终的准确率与 AUC。

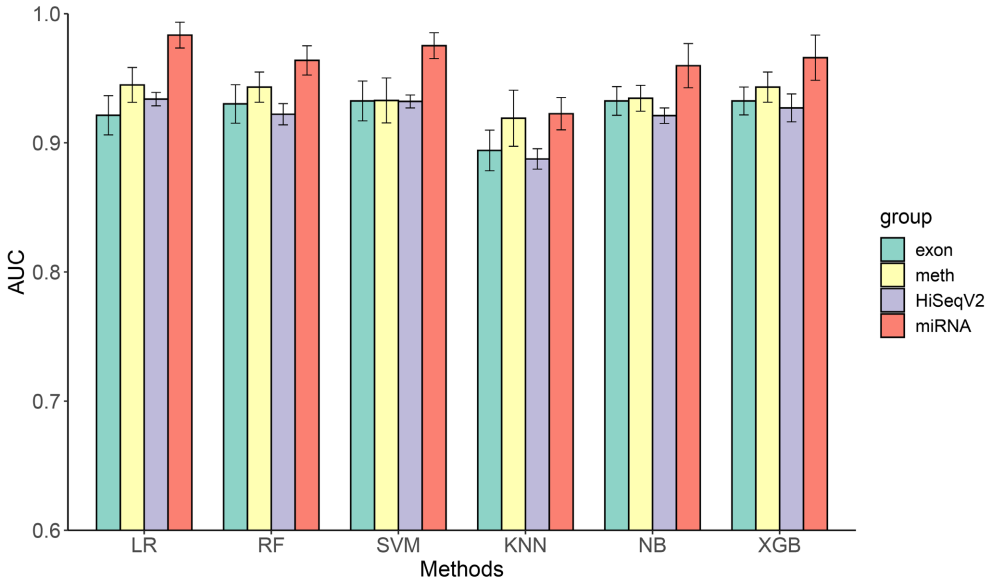


Figure 1. Mean AUC of each dataset
图 1. 各数据集的平均 AUC

肾细胞癌亚型分类的结果如图 1 所示。根据结果，我们可以清楚地看到，来自 miRNA 数据集的特征在所有数据中取得了最高的 AUC，这与表 2 的结果一致，miRNA 原始的特征最少，但共挑选出了最多的 77 个特征。来自 gene RNA_seq 数据集的特征对肾细胞癌亚型准确分类的贡献最小，但从图 1 中可以看出标准差很小，这表明 gene RNA_seq 可以为癌症亚型分类提供更稳定的特征。相比之下，其他三个数据集的标准差较大，在每个机器学习模型下的稳定性相对较差。

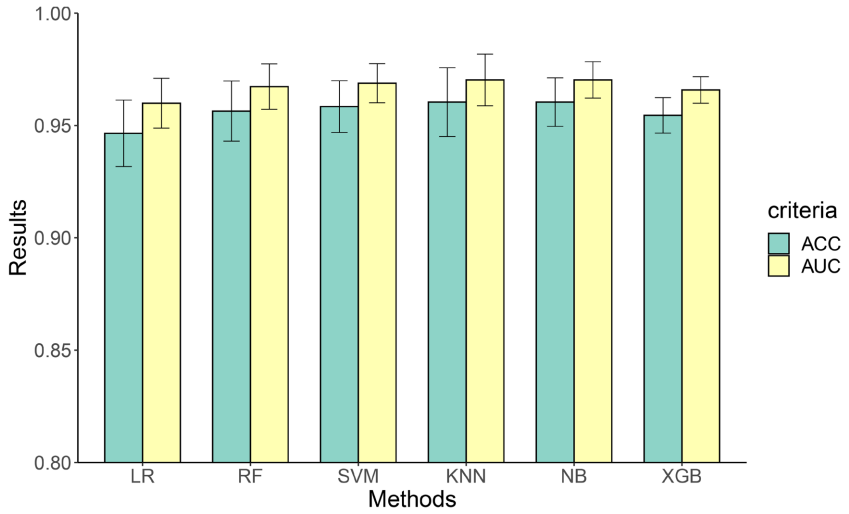


Figure 2. Mean accuracy and AUC of miRNA dataset
图 2. miRNA 数据集平均准确率和 AUC

我们选择图 1 中的最优结果进行更深一步的调查，即 miRNA 数据集的结果。图 2 和表 4 充分表明，大多数机器学习方法在这种组合下产生的特征可以获得 0.95 甚至更高的 AUC，最高甚至超过 0.98。除了 KNN 和 XGB 这两种机器学习方法有偏大的标准差外，所有方法得到的准确率和 AUC 都比较稳定，特别是 LR 与 SVM 方法。在所有方法中，LR 的结果最好，ACC 为 0.9779，AUC 为 0.9834。

Table 4. Results of using the new feature selection process with the machine learning classifier
表 4. 使用新特征选择流程配合机器学习分类器的结果

机器学习方法	准确率	准确率的标准差	AUC	AUC 的标准差
LR	0.9779	0.0134	0.9834	0.01
RF	0.9517	0.0151	0.9638	0.0113
SVM	0.9669	0.0134	0.9752	0.01
KNN	0.9462	0.0228	0.9597	0.0171
NB	0.9007	0.0167	0.9225	0.0125
XGB	0.9545	0.0223	0.9659	0.0175

5. 讨论

肾细胞癌是一种高度异质性的疾病，不同组织学亚型适合不同的治疗方案，且存在相差较大的预后。本文所述的研究对肾细胞癌至关重要。

本研究的创新之处在于，我们使用了多个 TCGA 单组学数据进行比较，应用了更全面的特征选择流程，采用多种类型的机器学习方法对肾细胞癌亚型进行分类，并评估每类的生存意义。在我们的研究中，

最核心的内容是提出了一个同时使用 mRMR、lasso 和 boruta 的特征提取器，用于基于 The Cancer Genome Atlas (TCGA)组学数据的肾细胞癌组织学亚型分类，利用六个基础的机器学习模型产生的准确率与 AUC 来判断最终亚型分类结果的优劣。虽然使用单一的特征选择方法提取特征很简单快捷，但结果很可能不尽人意。因此，我们决定采用更加鲁棒的投票法，结合三种特征选择技术，选择有利于肾细胞癌亚型分类的特征。最后，我们选择了 miRNA 数据集中的 77 个特征，使用六种机器学习模型评估特征选择技术的性能：逻辑回归(LR)、随机森林(RF)、支持向量机(SVM)、朴素贝叶斯(NB)、k 近邻(KNN)和 XGBoost，且每个机器学习模型都能获得出色的结果。其中，LR 在肾细胞癌亚型分类中能够产生最高和最稳定的准确率和 AUC。

图 1 中显示，miRNA 数据集为肾细胞癌亚型分类提供了最强的鉴别力，即交叉验证下各个机器学习方法均获得最高的平均结果，methylation 数据集的分类性能次之，gene RNA_seq 数据集的结果虽不名列前茅，但最为稳定。

为了验证本文方法的稳健性，我们选择了肺腺癌和乳腺癌的 TCGA 数据进行验证。我们发现，虽然肾细胞癌多组学数据没有明显提高亚型分类的准确率与 AUC，但使用多组学数据时肺腺癌的 AUC 高达 0.97，乳腺癌的 AUC 也可以达到 0.9 以上，相较单组学数据皆存在显著提升，这也反映出本文提出的特征提取方法较为鲁棒，可以适用于各种癌症。

虽然我们的方法在肾细胞癌组织学亚型分类中取得了良好的效果，但也有一些局限性：1) 在我们的研究中，我们只研究了肾细胞癌，且选取了肺腺癌与乳腺癌进行验证，我们的方法是否可以扩展到研究其他癌症的亚型，还需要进一步探讨。2) 虽然合并这五个数据集没有对肾细胞癌组织学亚型分类产生更优的结果，但我们没有使用更全面的多组学数据进行亚型分类，目前还不知道是否有更好的 TCGA 数据组合。

参考文献

[1] Eble, J.N. (2004) World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of the Urinary System and Male Genital Organs, 68-69.

[2] Hsieh, J.J., Purdue, M.P., Signoretti, S., Swanton, C., Albiges, L., Schmidinger, M., *et al.* (2017) Renal Cell Carcinoma. *Nature Reviews Disease Primers*, **3**, Article No. 17009. https://doi.org/10.1038/nrdp.2017.9

[3] Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A. and Bray, F. (2022) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, **71**, 209-249. https://doi.org/10.3322/caac.21660

[4] Linehan, W.M. (2012) Genetic Basis of Kidney Cancer: Role of Genomics for the Development of Disease-Based Therapeutics. *Genome Research*, **22**, 2089-2100. https://doi.org/10.1101/gr.131110.111

[5] Tomczak, K., Czerwińska, P. and Wiznerowicz, M. (2015) Review the Cancer Genome Atlas (TCGA): An Immeasurable Source of Knowledge. *Contemporary Oncology/Współczesna Onkologia*, **19**, A68-A77. https://doi.org/10.5114/wo.2014.47136

[6] Liu, J., Lichtenberg, T., Hoadley, K.A., Poisson, L.M., Lazar, A.J., Cherniack, A.D., *et al.* (2018) An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell*, **173**, 400-416. https://doi.org/10.1016/j.cell.2018.02.052

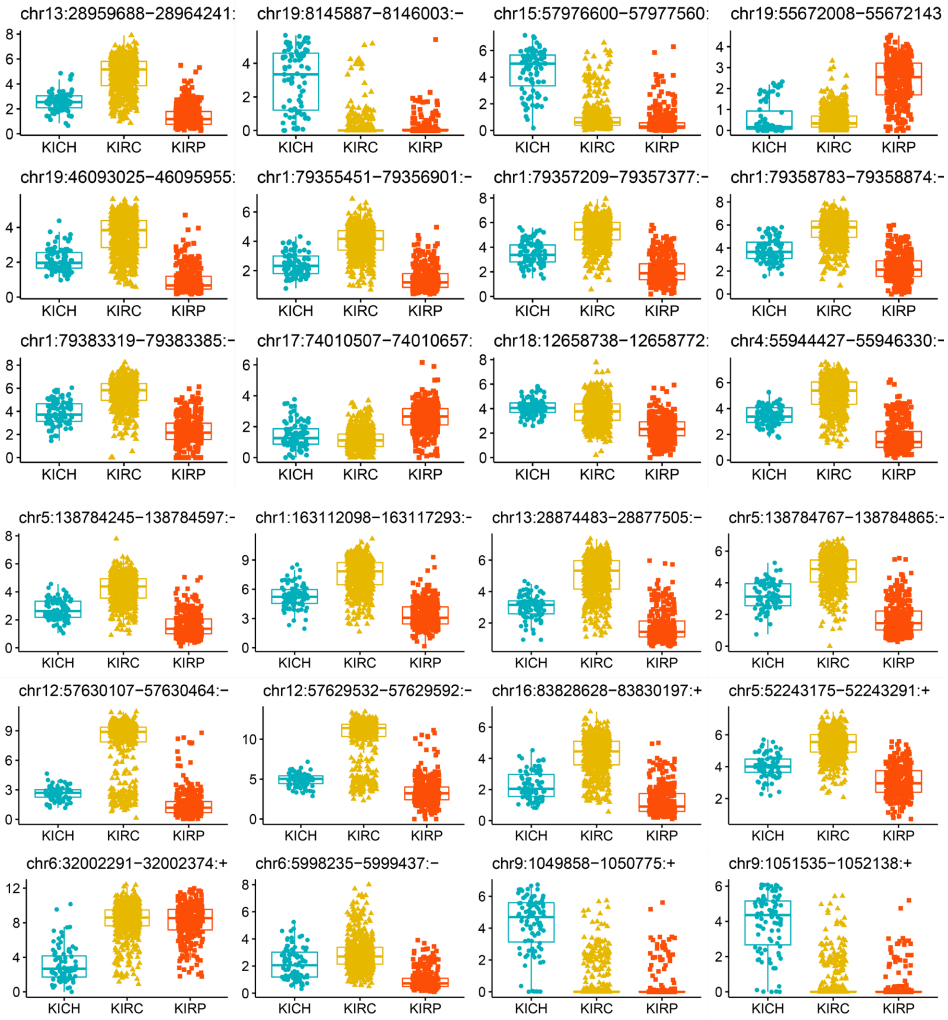
[7] Zhong, L., Meng, Q., Chen, Y., Du, L. and Wu, P. (2021) A Laminar Augmented Cascading Flexible Neural Forest Model for Classification of Cancer Subtypes Based on Gene Expression Data. *BMC Bioinformatics*, **22**, Article No. 45. https://doi.org/10.1186/s12859-021-04391-2

[8] Tibshirani, R. (1997) The Lasso Method for Variable Selection in the Cox Model. *Statistics in Medicine*, **16**, 385-395. https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3

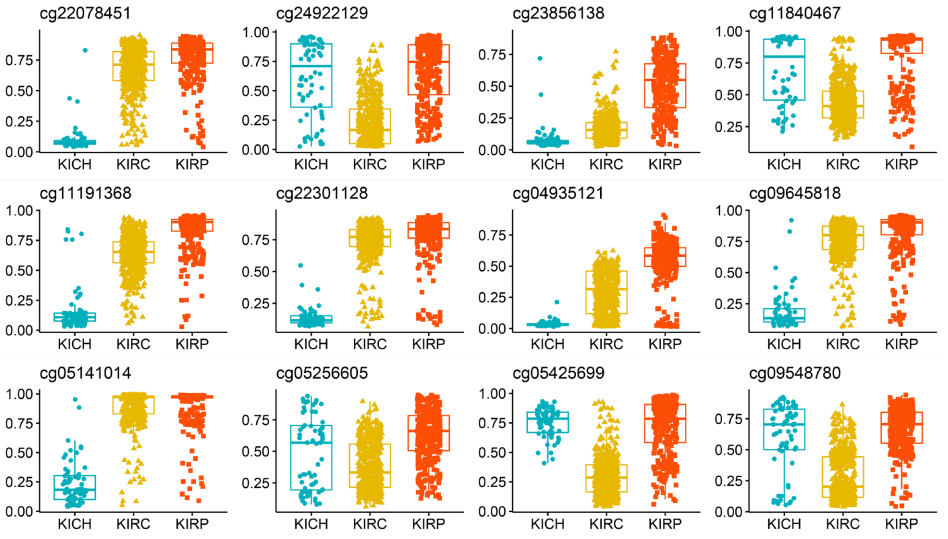
[9] Peng, H., Long, F. and Ding, C. (2005) Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, 1226-1238. https://doi.org/10.1109/TPAMI.2005.159

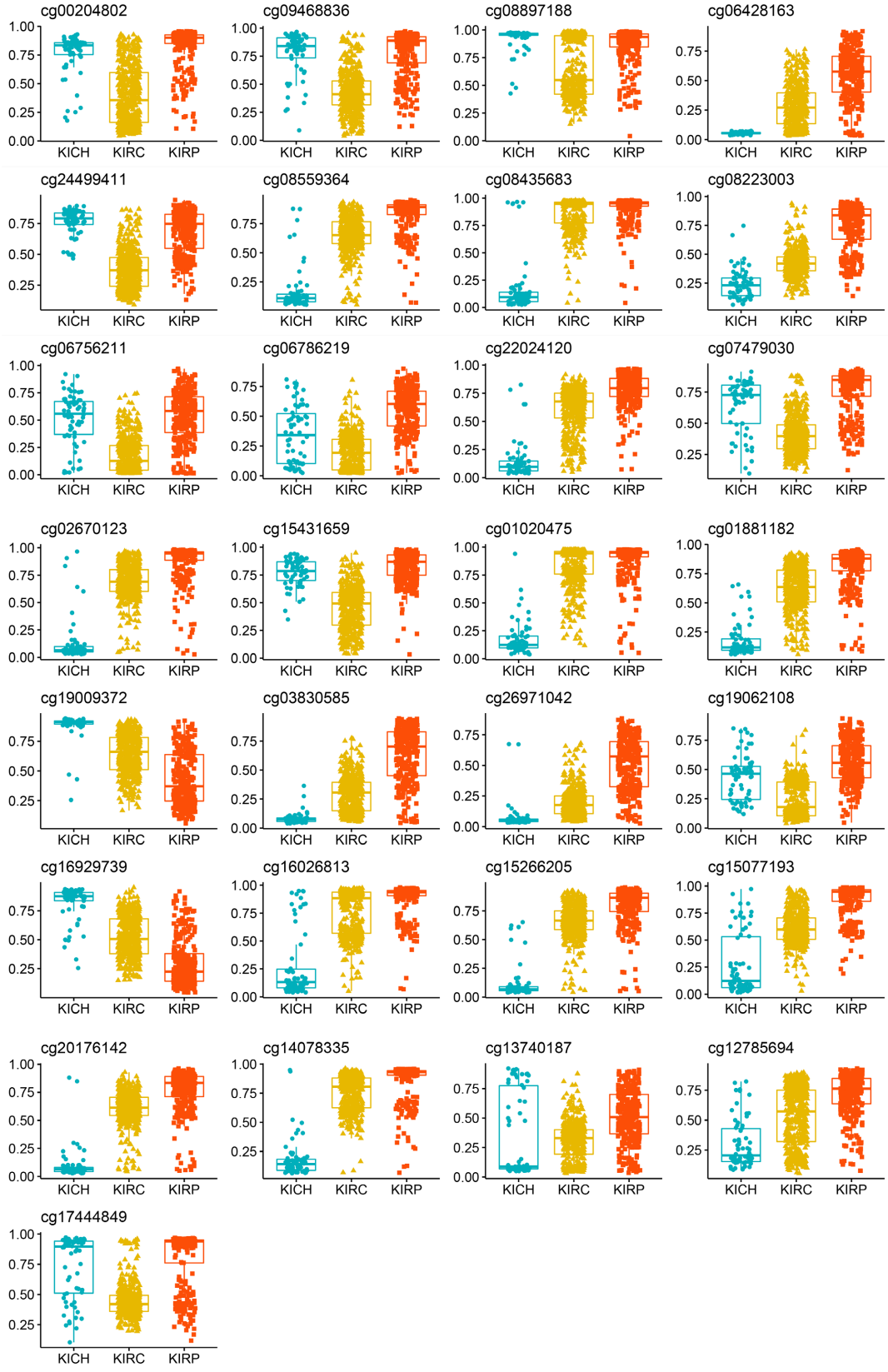
[10] Kursa, M.B. and Rudnicki, W.R. (2010) Feature Selection with the Boruta Package. *Journal of Statistical Software*, **36**, 1-13. https://doi.org/10.18637/jss.v036.i11

附　录

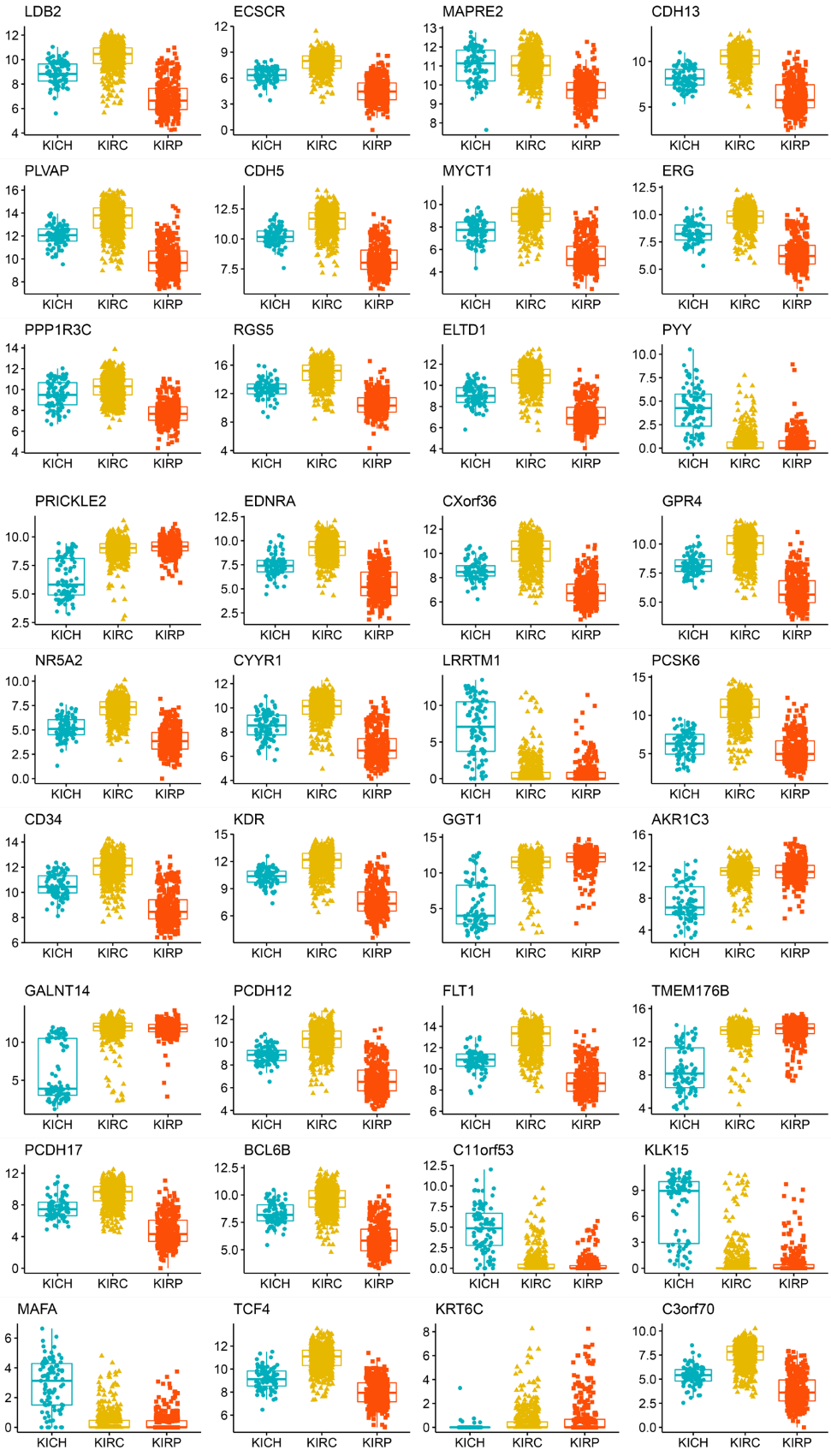

chr13:28959688−28964241:
chr19:8145887−8146003:−
chr15:57976600−57977560:
chr19:55672008−55672143
chr19:46093025−46095955:
chr1:79355451−79356901:−
chr1:79357209−79357377:−
chr1:79358783−79358874:−
chr1:79383319−79383385:−
chr17:74010507−74010657:
chr18:12658738−12658772:
chr4:55944427−55946330:−
chr5:138784245−138784597:−
chr1:163112098−163117293:−
chr13:28874483−28877505:−
chr5:138784767−138784865:−
chr12:57630107−57630464:−
chr12:57629532−57629592:−
chr16:83828628−83830197:+
chr5:52243175−52243291:+
chr6:32002291−32002374:+
chr6:5998235−5999437:−
chr9:1049858−1050775:+
chr9:1051535−1052138:+
KICH
KIRC
KIRP

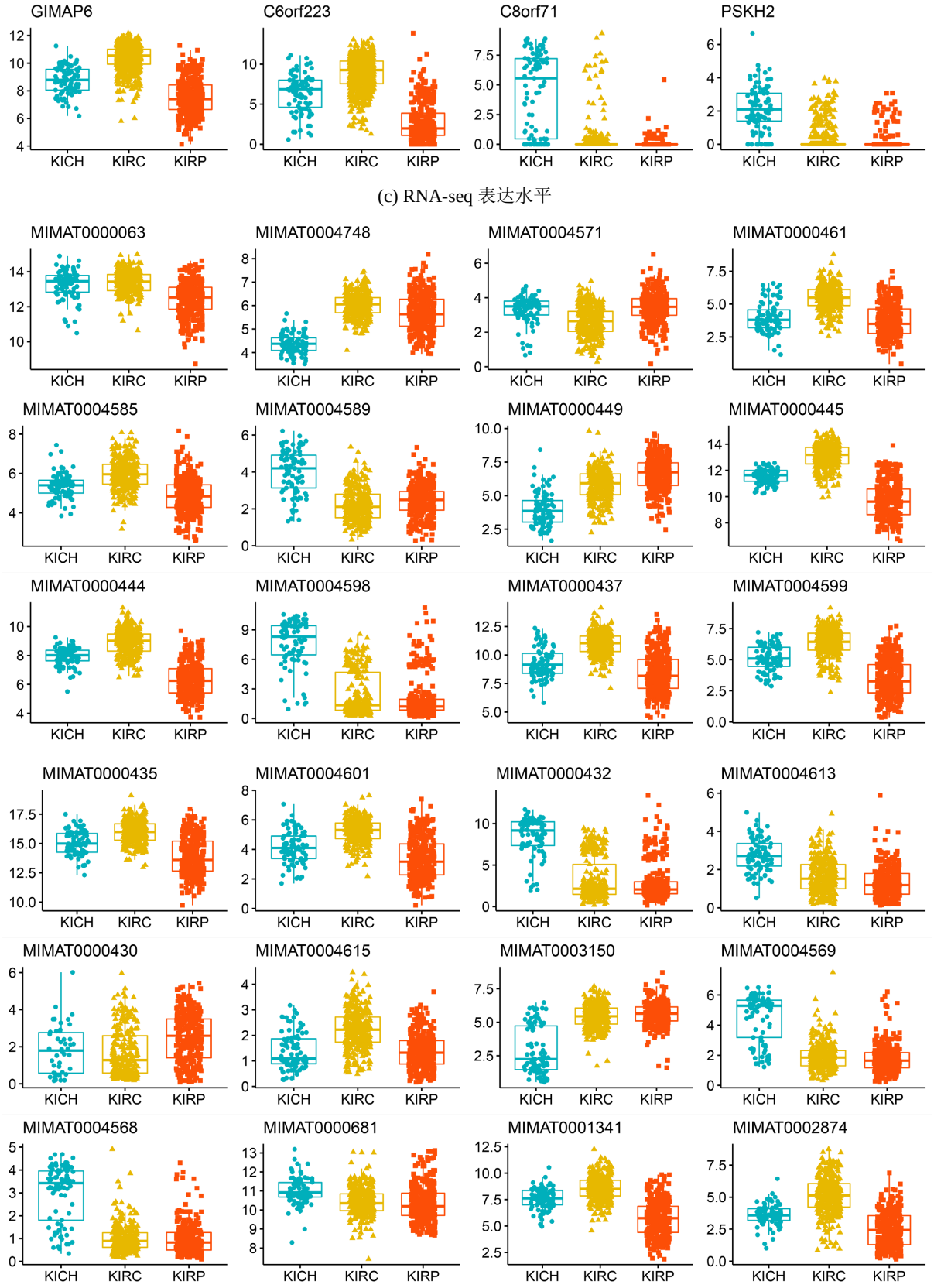
(a) 外显子表达水平


cg22078451
cg24922129
cg23856138
cg11840467
cg11191368
cg22301128
cg04935121
cg09645818
cg05141014
cg05256605
cg05425699
cg09548780
KICH
KIRC
KIRP

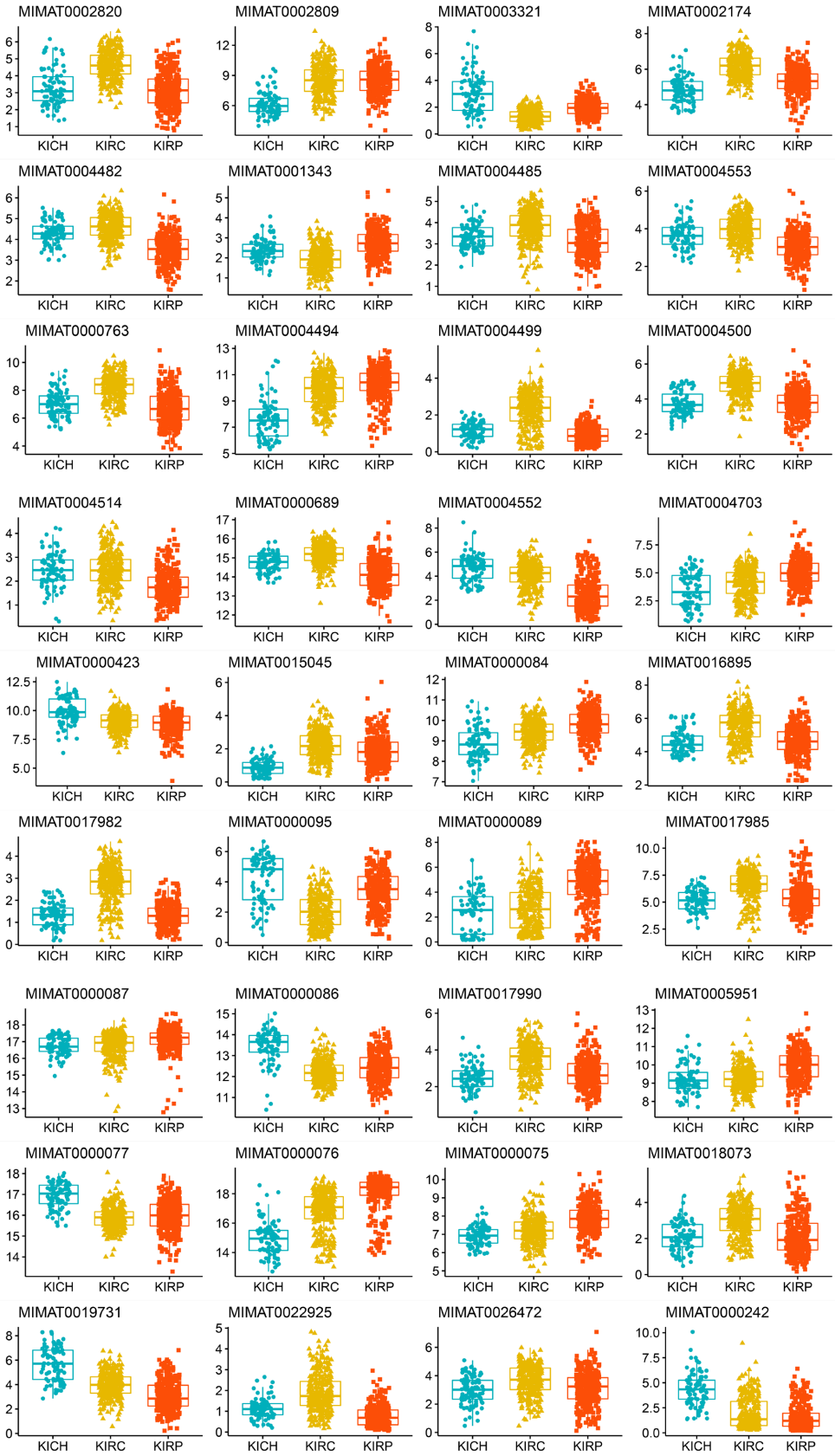

cg00204802
cg09468836
cg08897188
cg06428163
cg24499411
cg08559364
cg08435683
cg08223003
cg06756211
cg06786219
cg22024120
cg07479030
cg02670123
cg15431659
cg01020475
cg01881182
cg19009372
cg03830585
cg26971042
cg19062108
cg16929739
cg16026813
cg15266205
cg15077193
cg20176142
cg14078335
cg13740187
cg12785694
cg17444849
KICH
KIRC
KIRP

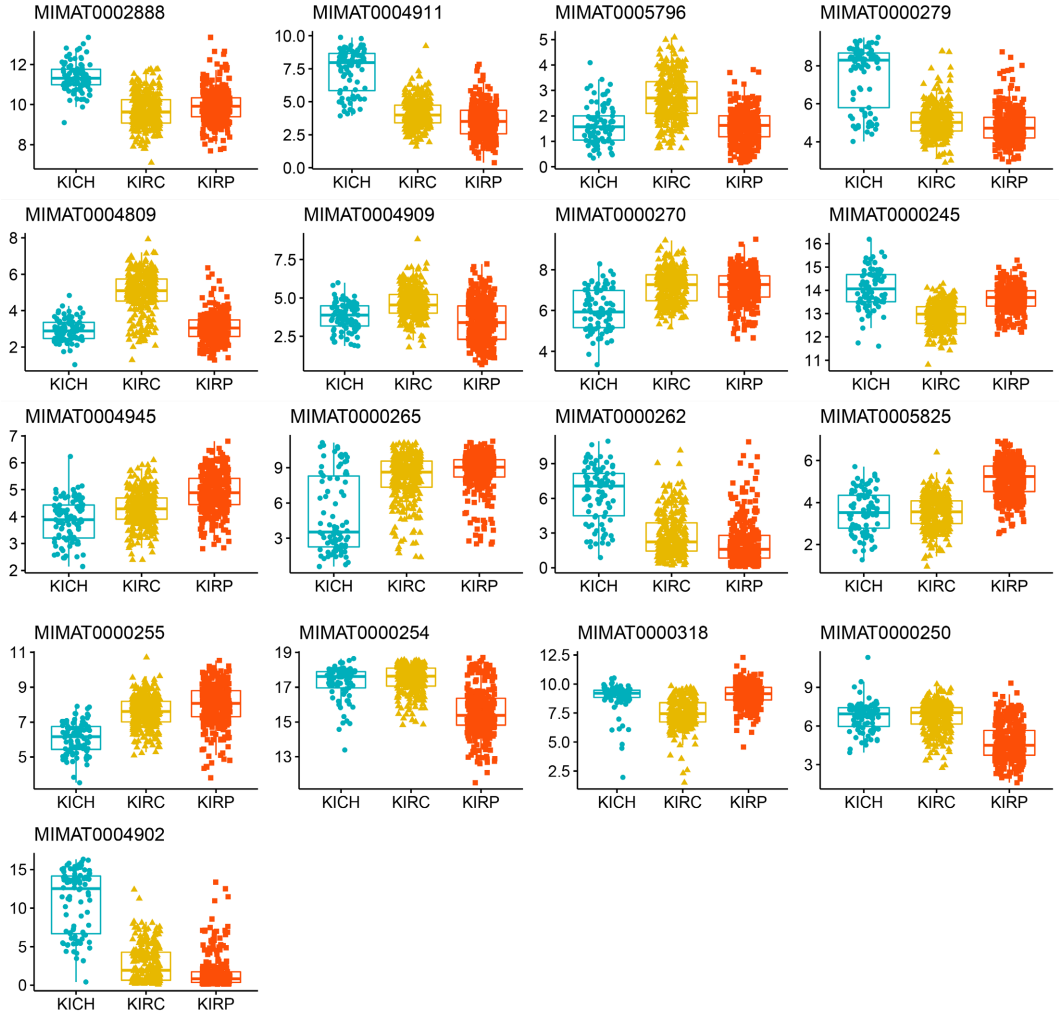
(b) 甲基化水平


LDB2
ECSCR
MAPRE2
CDH13
PLVAP
CDH5
MYCT1
ERG
PPP1R3C
RGS5
ELTD1
PYY
PRICKLE2
EDNRA
CXorf36
GPR4
NR5A2
CYYR1
LRRTM1
PCSK6
CD34
KDR
GGT1
AKR1C3
GALNT14
PCDH12
FLT1
TMEM176B
PCDH17
BCL6B
C11orf53
KLK15
MAFA
TCF4
KRT6C
C3orf70
KICH
KIRC
KIRP



(c) RNA-seq 表达水平


MIMAT0002820
MIMAT0002809
MIMAT0003321
MIMAT0002174
MIMAT0004482
MIMAT0001343
MIMAT0004485
MIMAT0004553
MIMAT0000763
MIMAT0004494
MIMAT0004499
MIMAT0004500
MIMAT0004514
MIMAT0000689
MIMAT0004552
MIMAT0004703
MIMAT0000423
MIMAT0015045
MIMAT0000084
MIMAT0016895
MIMAT0017982
MIMAT0000095
MIMAT0000089
MIMAT0017985
MIMAT0000087
MIMAT0000086
MIMAT0017990
MIMAT0005951
MIMAT0000077
MIMAT0000076
MIMAT0000075
MIMAT0018073
MIMAT0019731
MIMAT0022925
MIMAT0026472
MIMAT0000242
KICH
KIRC
KIRP

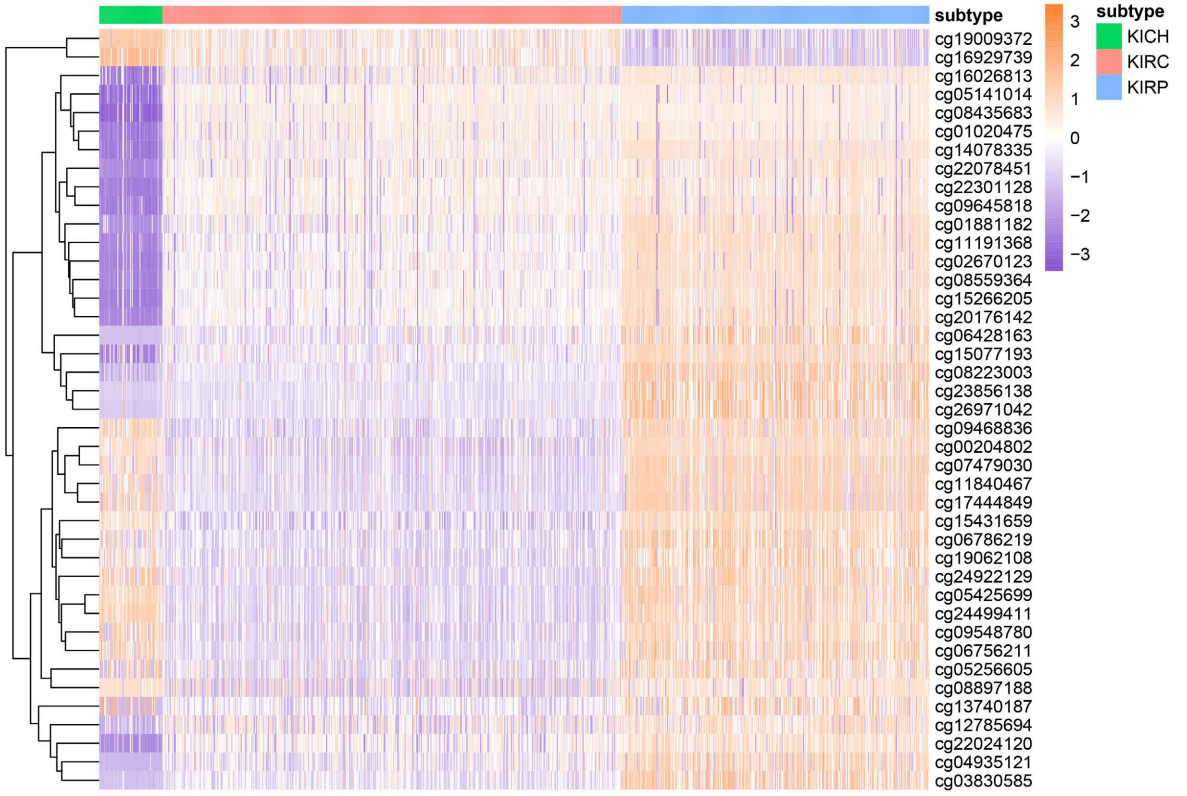


(d) miRNA 成熟链表达水平

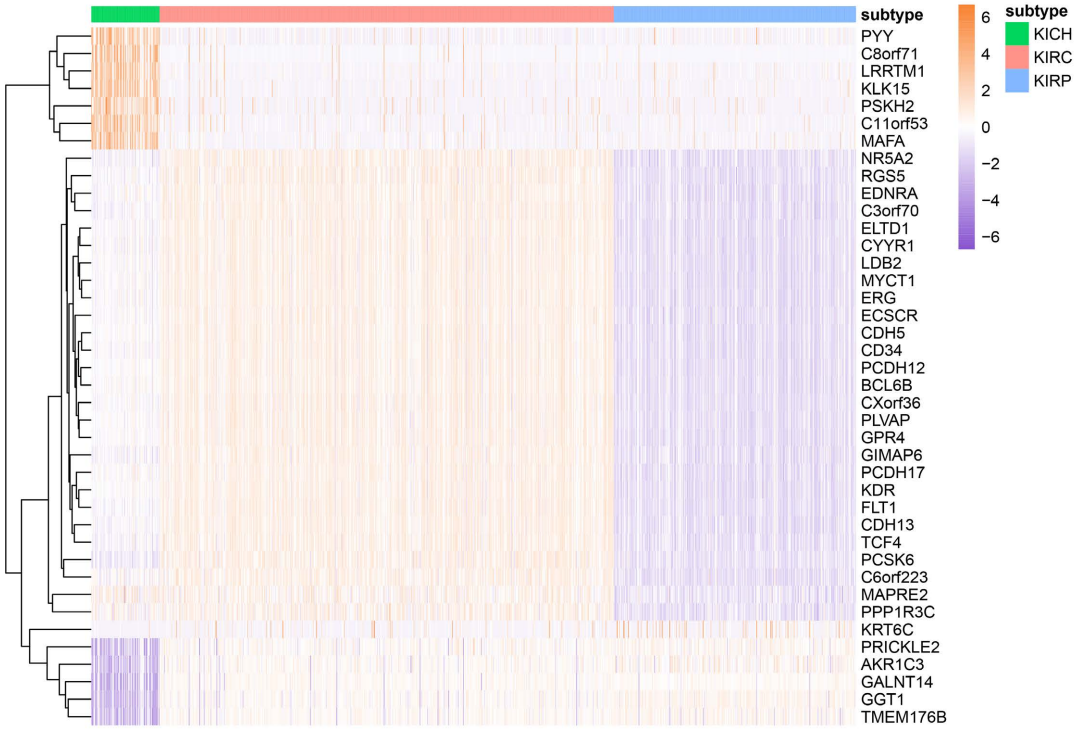
Figure S1. Box plots of expression levels of selected important features: exon expression level (a), methylation level (b), RNA-seq expression level (c), miRNA mature strand expression level (d)

附图 1. 挑选出的重要特征的表达水平箱线图：外显子的表达水平(a)，甲基化水平(b)，RNA-seq 的表达水平(c)，miRNA 成熟链表达水平(d)

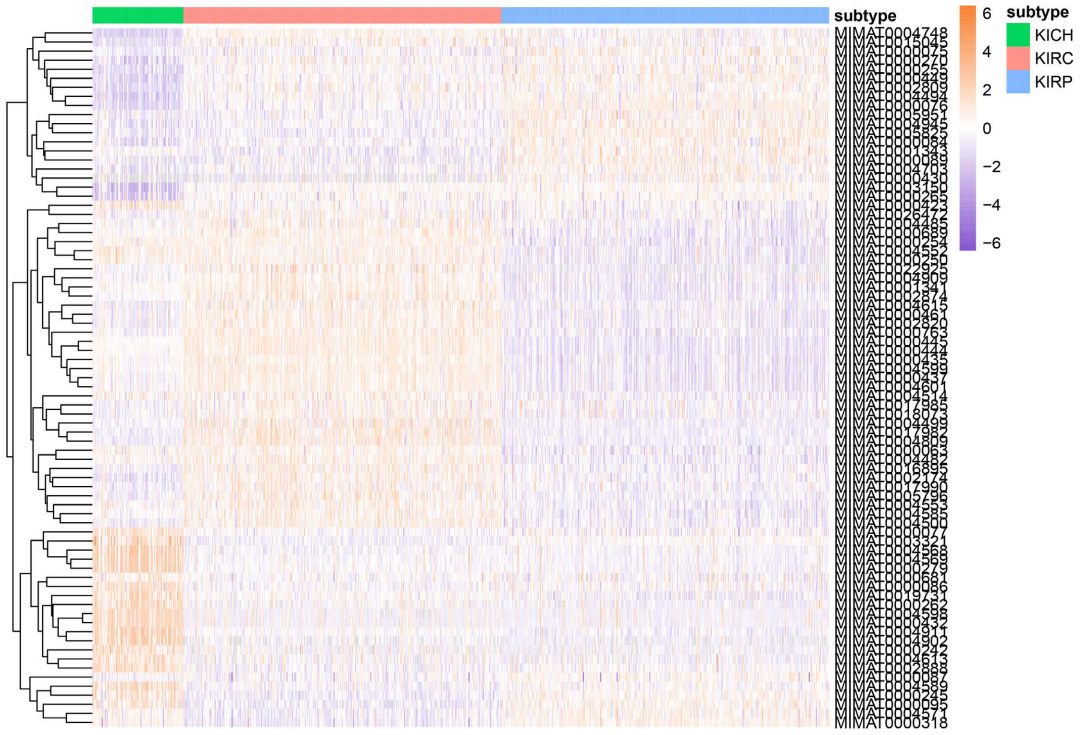
(a) exonRNA-seq 重要特征热力图



(b) methylation 重要特征热力图



(c) geneRNA-seq 重要特征热力图



(d) miRNA 特征热力图

Figure S2. Heatmap of selected important features: exon expression level (a), methylation level (b), RNA-seq expression level (c), miRNA mature strand expression level (d)

附图 2. 挑选出的重要特征的热力图：外显子的表达水平(a)，甲基化水平(b)，RNA-seq 的表达水平(c)，miRNA 成熟链表达水平(d)