引用格式：刘清瑞, 刘婕, 岳振宇, 等. 癌症基因组框内突变功能注释及计算方法比较分析 [J]. 电子科技大学学报, 2025, 54(2): 311-320.
LIU Q R, LIU J, YUE Z Y, et al. Comparative analysis of functional annotation and computational methods for in-frame InDels in cancer genome[J]. Journal of University of Electronic Science and Technology of China, 2025, 54(2): 311-320.

癌症基因组框内突变功能注释及计算方法比较分析

刘清瑞[1]，刘　婕[1]，岳振宇[2]，夏俊峰[1*]

（1. 安徽大学 物质科学与信息技术研究院，合肥 230601；2. 安徽农业大学 信息与人工智能学院，合肥 230036）

摘要：框内突变是编码区插入缺失突变的一种常见类型，与癌症的发生发展密切相关。然而，计算方法在癌症驱动框内突变预测方面的有效性尚缺乏明确共识。首先，系统地比较和评估了 8 种计算方法，证实了它们在识别癌症驱动框内突变的适用性及可靠性。然后，选用其中 4 种表现优异的计算方法，进一步挖掘了癌症基因组中潜在的驱动框内突变，并探究了这些突变作为癌症驱动突变的合理性。最终，构建了一个用户访问友好、集成多种预测方法及注释信息的线上数据库 dbCCID，旨在为研究人员提供便利。这些工作为癌症框内突变预测方法的选择和开发提供了理论支撑。

关键词：框内突变；癌症；性能评估；数据库构建
中图分类号：TP391　　**文献标志码：**A　　DOI: 10.12178/1001-0548.2024061

Comparative analysis of functional annotation and computational methods for in-frame InDels in cancer genome

LIU Qingrui[1], LIU Jie[1], YUE Zhenyu[2], and XIA Junfeng[1*]

(1. Institutes of Physical Science and Information Technology, Anhui University, Hefei 230601, China;
2. College of Information and Artificial Intelligence, Anhui Agricultural University, Hefei 230036, China)

Abstract: In-frame InDel is a common type of insertion and deletion mutations in coding regions, which are closely associated with the occurrence and development of cancer. However, there is currently a lack of clear consensus on the efficacy of computation methods for predicting cancer driver in-frame InDels. In this paper, eight computational methods are comprehensively and systematically compared and evaluated, confirming their applicability and reliability of these methods in identifying cancer driver in-frame InDels. Then, four computational methods with outstanding performance are selected to mine potential driver in-frame InDels in the cancer genome and explore the rationality of these mutations as cancer driver InDels. Finally, a user-friendly online database dbCCID that integrates multiple prediction methods and annotation information is constructed to create convenience for researchers. It is expected this work will provide a theoretical support for the selection and development of in-frame InDel prediction methods for cancer.

Key words: in-frame InDel; cancer; performance evaluation; database construction

癌症已成为危害人类健康的首要疾病，其发病率和死亡率逐年增长[1-2]。癌症的发生发展是一个极为复杂的过程，其中基因突变扮演着关键角色。插入缺失突变（insertion-deletion mutation, InDel）是基因突变中发生频次最多的突变之一，其涉及一对或多对碱基的插入或删除，常导致较长片段的蛋白质序列发生变化，进而诱发诸如癌症等严重影响生物体正常生理功能的疾病[3-4]。作为一个重要的癌症驱动因素，插入缺失突变可能发生在基因的编码区和非编码区。编码区是基因中负责编码蛋白质的区域，包含了丰富的基因功能信息，因此该区域的突变对基因表达的影响尤为显著。框内突变是编码区插入缺失突变的一种常见类型，其特点是涉及的碱基对（base pair, bp）数目为 3 的整倍数。这意味着

收稿日期：2024−03−19
基金项目：国家自然科学基金（U22A2038）
作者简介：刘清瑞，主要从事生物信息学方面的研究。
*通信作者 E-mail: jfxia@ahu.edu.cn

在框内突变中，氨基酸序列的改变是以一个或多个完整的氨基酸为单位进行的，这也是该突变呈现出致病性的生物学本质。

在癌症基因组中，致病突变可能涉及肿瘤抑制基因或原癌基因等关键基因，赋予癌细胞选择性生长优势，能够直接促进癌细胞的增长。因此，这类突变也被称为驱动突变（driver mutation）。与此同时，癌症基因组中也存在着一类被称为乘客突变（passenger mutation）的中性突变。这类突变虽存在于癌细胞中，但本身与癌症的发生发展没有直接联系[5]。对癌症基因组框内突变而言，判断其中癌症驱动突变实质上就是确定潜在的致病突变。相较于传统繁琐耗时的生物学实验，计算模型在预测致病突变方面提供了一种高效且节省资源的替代方案[6]。一般流程如下：首先使用致病突变预测方法对癌症中突变数据进行预测获得致病分值，然后将致病分值与该方法的致病阈值进行比较即可判断此突变是否致病。若此突变为致病突变，则说明该突变为潜在的癌症驱动突变。

近年来，已有研究表明编码区框内突变与癌症的发生和发展有密切的关联[7]。然而，在初期的调研阶段，通过“in-frame InDel”“pathogenicity prediction”等关键词进行文献检索，本文发现针对癌症框内突变方面的预测方法的对比工作尚未展开。因此，开展针对框内突变预测方法的对比研究分析，首先可以明确现有方法对癌症中框内突变预测的有效性和可靠性，为后续癌症特异性突变预测方法的开发及选择提供指导。其次，尽管现有方法一定程度上提高了致病突变预测的准确性，在对于不同长度或发生在不同基因等情况下的框内突变的预测上仍可能存在一些不足。基于收集整理的突变基准数据集，进行全面的评估分析，有利于进一步揭示这些不足[8]，并可作为未来基于深度学习或者大数据的智能模型的研究方向。最后，随着下一代测序技术的发展[9-10]，涌现了以 COSMIC[11] 为代表的大量未被解析的癌症基因组框内突变数据。结合对比研究的结果并选择合适的方法，可以更加准确地挖掘这些突变中潜在的癌症驱动突变。

目前，有基于大数据学习模型的深度学习方法用于预测突变，然而并非针对框内突变进行致病预测或功能效应研究[12]。同时，现有方法在癌症中框内突变预测方面的有效性尚未明确。众多突变数据库和基因组分析工具过于分散，致使研究者获取全面的框内突变信息面临挑战。为此，本文首先系统地比较了 8 种现有致病突变预测方法，证实了这些方法在识别癌症驱动框内突变的适用性和可靠性。其中，VEST-InDel[13]（简写为 VEST-I）和 CAPICE[14] 是性能表现最突出的两个方法，在多个基准数据集上的 AUC 和 AUPR 指标均超过了 0.8。随后，基于 4 种性能稳定的方法对癌症基因组框内突变进行预测，并与已知的驱动突变进行相似性对比，证实了其作为驱动突变的合理性。最后，构建了一个用户访问友好、集成多种方法预测结果及注释信息的线上数据库 dbCCID（http://dbccid.xialab.info/），旨在为研究人员提供便利。综上，本文工作为面向癌症的框内突变的相关研究提供了理论支撑。

1 材料与方法

1.1 数据收集

为比较致病突变预测方法在识别癌症驱动框内突变时的表现，分别从 HGMD[15] 和 gnomAD[16] 中收集驱动框内突变（正样本）和乘客框内突变（负样本），以构建一个框内突变基准数据集。为尽可能收集多的正样本数据，首先从 HGMD 中选取标签为引起疾病或疑似引起疾病的框内突变，同时通过关键字查询筛选出与癌症相关的突变。其次，从 gnomAD 中收集等位基因频率大于 1% 的框内突变作为负样本，并去除正样本中已存在的样本数据。然后，采用 Close-by 的原则构建数量平衡数据集，即对于每一条正样本，从负样本中不放回的选取一条与其处于同一条染色体且突变位置最接近的突变。最后，整理得到了一个包含正负样本各 203 条的基准数据集，记为数据集 I。基于该数据集，本文从多方面评估了不同预测方法的性能。

为探究癌症基因组中未被解析的框内突变数据的功能效应，分别从 COSMIC[11] 提供的 GRCh38 和 GRCh37 这 2 个参考版本下的全基因组数据中收集框内突变数据。考虑到部分致病突变预测方法只能进行特定版本下的突变预测，使用 LiftOver[17] 对收集的框内突变进行基因组坐标转换并取交集。通过这一过程，最终得到了 2 个版本下均存在的 6 142 条癌症基因组框内突变，记为数据集 M。基于数据集 M 的功能注释结果，本文挖掘了癌症基因组中潜在的驱动框内突变数据。

1.2 方法简述

考虑到是否提供线上预测网页、是否有本地源码等多方面因素，本文研究了 8 种现有的致病突变预测方法。其中，能够预测包括框内突变在内的多

种插入缺失突变的广谱性方法有 6 种，分别为 CADD[18]、CAPICE、SIFT-InDel[19]（简写为SIFT-I）、VEST-I、MutationTaster2021[20]（简写为 MT2021）和 FATHMM-InDel[21]（简写为 FATHMM-I）。仅能预测框内突变这一种插入缺失突变的特异性方法有 2 种，分别为 MutPred-InDel[22]（简写为 MutPred-I）和 MetaRNN-InDel[23]（简写为 MetaRNN-I）。这些预测方法的基本信息如表 1 所示。

表 1　框内突变预测方法基本信息

方法	输入格式	线上服务	基础模型	参考基因	中性阈值	预测阈值	长度限制/bp
CADD	VCF	+	回归分类器	37/38	< 20	⩾ 20	<50
CAPICE	TSV	−	决策树	37/38	< 0.02	⩾0.02	NA
SIFT-I	CSV	−	决策树	37/38	NA	NA	NA
VEST-I	VCF	+	随机森林	37/38	< 0.5	⩾ 0.5	NA
MT2021	VCF	−	随机森林	37	NA	NA	⩽40
FATHMM-I	VCF	+	支持向量机	37	< 0.5	⩾ 0.5	NA
MutPred-I	FASTA	+	神经网络	37/38	< 0.672	⩾ 0.672	30～30 000
MetaRNN-I	VCF	−	Transformer	38	< 0.5	⩾ 0.5	⩽40

需要说明的是，表 1 中第一列“线上服务”表示是否提供了预测网页。对于提供了的方法（表中用“+”表示），直接使用其网页进行预测。反之（表中用“−”表示），采用本地复现的方式完成预测。“参考基因”表示可预测的突变版本，37 表示仅能预测 GRCh37 下的突变，38 表示仅能预测 GRCh38 下的突变，37/38 表示两者皆可。MutPred-I 预测前需将突变转为氨基酸序列，突变碱基变化长度（简称长度）限制针对于该氨基酸序列。部分工具用于区分致病和中性突变的阈值来自文献[24] 的研究工作。表中用“NA（not available）”表示方法中未明确的相关信息。

特别地，表 1 中呈现的 8 种致病预测方法所采用的基础模型均为常见的机器学习模型或者深度学习模型。对广谱性方法而言，CAPICE 和 SIFT-I 采用决策树作为基础模型，VEST-I 和 MT2021 基于随机森林构建预测模型，CADD 和 FATHMM-I 则分别采用回归分类器和支持向量机作为预测模型的核心组成结构；对特异性方法而言，MetaRNN-I 采用 Transformer 作为基础模型，MutPred-I 则基于神经网络进行了构建。可以看出，在现有致病突变预测方法中，尤其是特异性方法中，深度学习技术被广泛运用，以处理更加复杂的突变识别任务和更为多样化的生物学数据。

2　实验结果与分析

2.1　已有预测方法性能评估

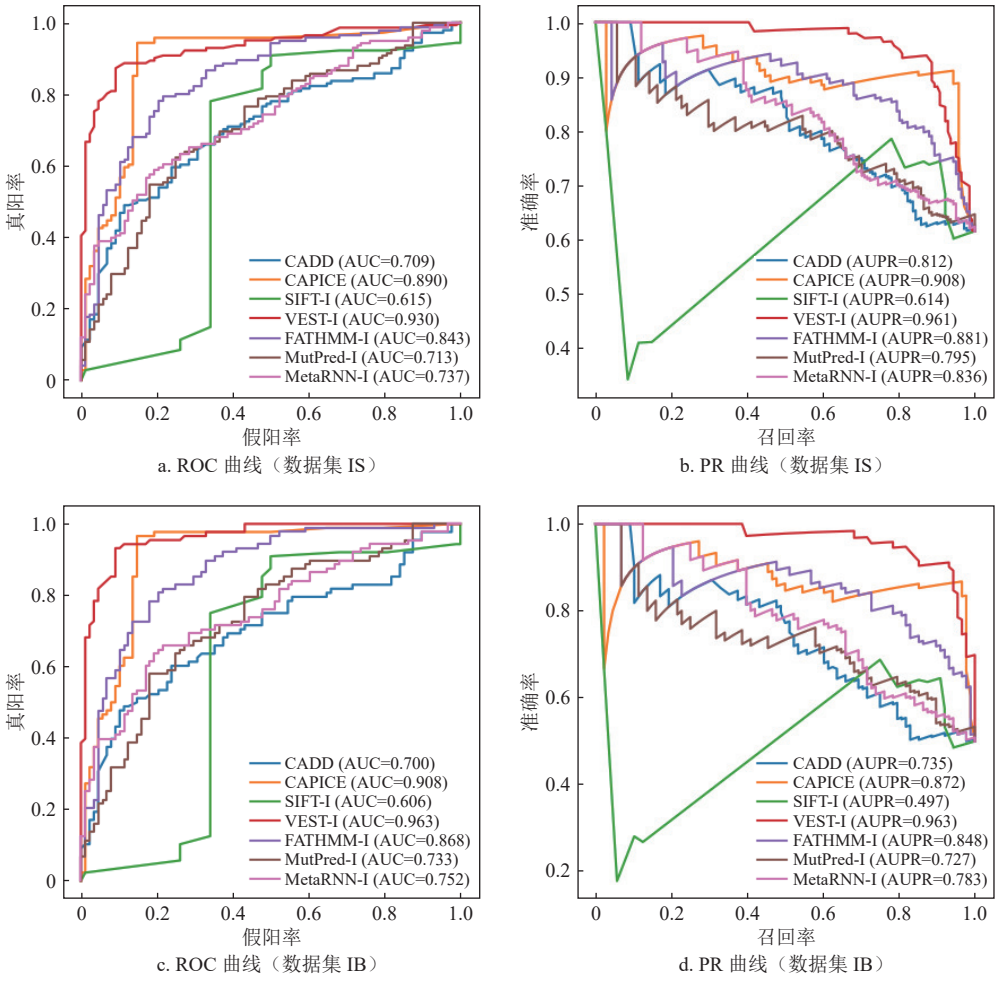
为明确现有致病突变预测方法在识别癌症驱动框内突变时的适用性和可靠性，分别从预测性能指标、预测一致性和致病阈值等方面系统地评估了这些方法的性能。

2.1.1　性能指标分析

本研究采用表 1 中的方法对基准数据集 I 中框内突变进行致病预测，并统计了每个方法致病分数的缺失情况。统计结果表明，CAPICE 不存在分数缺失问题，CADD 等其他广谱性方法存在少量的分数缺失，这表明这些方法在预测癌症驱动框内突变方面具有较好的适用性。所有突变数据中，同时具备所有方法致病分数的正负样本分别有 141 和 88 条，记为基准数据集 IS（score）。值得注意的是，框内突变特异性方法的致病分数缺失数量普遍多于广谱性方法。这一现象的原因可能是特异性方法常被设计于某种特定类型突变的预测，所考虑的特征范围及所依赖的训练数据相对有限。这或许是未来开发致病框内突变特异性预测新方法时需要注意的问题。

考虑到数据平衡性对评估结果的影响，从数据集 IS 中随机选取 88 条正样本，和其中的负样本组成平衡数据集，记为数据集 IB（balance）。基于这 2 个基准数据集，从数据平衡方面评估了这些方法的预测性能。图 1 中绘制了不同预测方法在数据集 IS 和 IB 上的 ROC 和 PR 曲线。由于没有致病分数，MT2021 未能在图中呈现相应曲线。从图 1a 和图 1b 中可以看出，VEST-I 在数据集 IS 上的表现出色，AUC 和 AUPR 分别为 0.930 和 0.961，是所有方法中最高的。其次是 CAPICE，AUC 和 AUPR 均接近 0.9，这可能得益于其采用了大量高置信度且平衡的训练数据。图 1c 和图 1d 分别展示了这些方法在数据集 IB 上的预测表现。其

中，VEST-I 的这两项指标略微增加，均达到了 0.963。CAPICE 的性能表现稳定，AUC 和 AUPR 的整体变化幅度不大。从指标数值来看，特异性方法 MetaRNN-I 在这 2 个数据集上 AUC 均大于 0.730，性能表现可观。相反地，SIFT-I 的 AUC 和 AUPR 数值均处于较低水平，这一定程度上受其使用的训练数据数量相对有限影响。除了 SIFT-I，其他方法在数据集 IB 和 IS 上的指标差异不大。综合分析这些结果，本文得出结论，VEST-I 和 CAPICE 在癌症基因组框内突变的致病性预测上表现最为突出，在不同规模的数据集上均展现出了优异的性能。此外，本研究还表明，对于大多数预测方法而言，数据集的平衡性对其预测性能的影响并不显著。



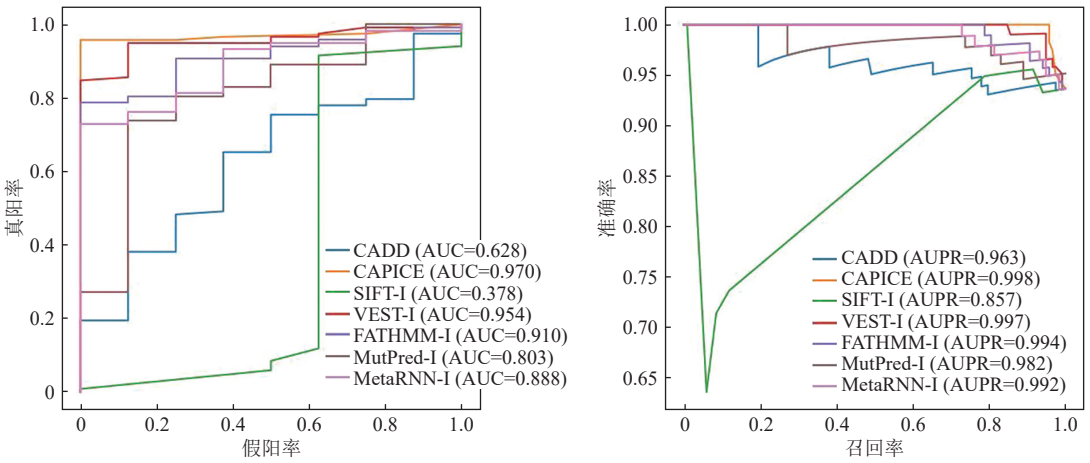
a. ROC 曲线（数据集 IS）　b. PR 曲线（数据集 IS）

c. ROC 曲线（数据集 IB）　d. PR 曲线（数据集 IB）

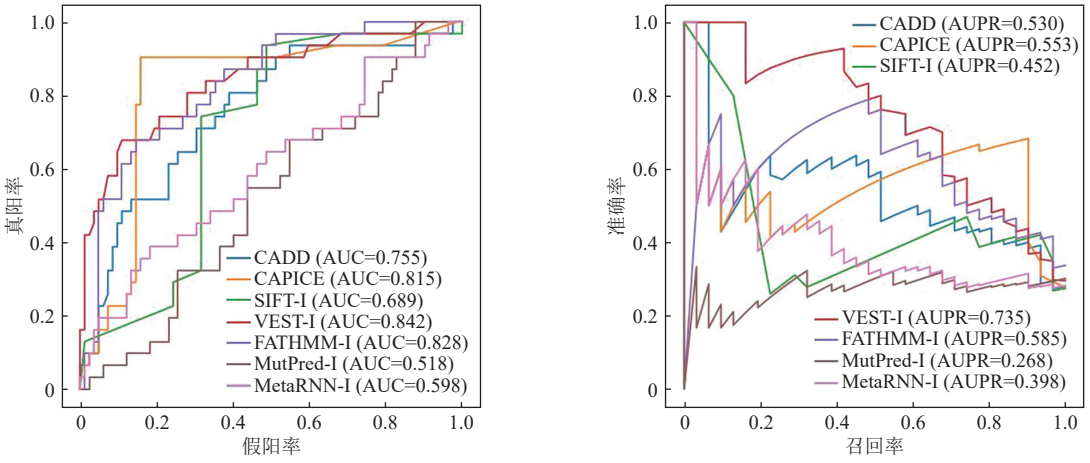
图 1　数据集 IS 和 IB 的 ROC 和 PR 曲线

框内突变和基因的分布密切相关，对癌症基因组研究具有重要意义。为从基因层面深入评估不同预测方法的性能，根据框内突变是否位于已知癌症基因[25]上，将数据集 IS 划分为分布在癌症基因上的数据集 IS-D（driver）和没分布在癌症基因上的数据集 IS-P（passenger）。如图 2a 和图 2b 所示，多数预测方法展现出了良好的预测性能，在癌症基因上的突变数据集 IS-D 上的 AUC 和 AUPR 数值处于较高水平。尤其是 VEST-I、CAPICE 和 FATHMM-I，这些方法的 2 个指标均在 0.9 以上，显示出极高的预测准确性。特异性方法 MetaRNN-I 也在该数据集上维持着稳定的预测性能。

然而，对于非癌症基因上的突变数据集 IS-P，情况则有所不同。如图 2c 所示，除了 CADD 和 SIFT-I，其他方法的 AUC 指标均有所减小。更加值得注意的是，当预测非癌症基因上的突变时，所有方法的 AUPR 数值均有着不同程度的降低，如图 2d 所示。这表明，基因的分布对现有预测方法的性能有着明显影响。尽管如此，VEST-I 和 CAPICE 这 2 个方法仍然能够在不同基因类型的基准数据集上保持着出色性能，这进一步体现了二者广泛的适应性。



a. ROC 曲线（数据集 IS-D）　　b. PR 曲线（数据集 IS-D）



c. ROC 曲线（数据集 IS-D）　　d. PR 曲线（数据集 IS-D）

图 2　数据集 IS-D 和 IS-P 的 ROC 和 PR 曲线

框内突变可能涉及一个或多个氨基酸的插入或缺失，其碱基变化长度（简称长度）是不均匀的。为从长度层面评估现有预测方法，根据不同长度对基准数据集 IS 进行子集划分，如表 2 所示。对于长度为 3 bp 的数据集 IS-L3 而言，VEST-I 和 CAPICE 的性能表现突出，AUC 和 AUPR 均达到了 0.910。相比之下，SIFT-I 在该数据集上的表现不尽人意，其各项指标均低于 0.650，数值相对较低。在长度为 6 bp 的数据集 IS-L6 上，不同方法的性能与数据集 IS-L3 接近，各指标的变化幅度不大。其中，VEST-I 仍呈现出最高水平的性能评估指标，CAPICE 和 FATHMM-I 紧随其后。当长度追加到为 9 bp 时，CADD 以及特异性方法 MutPred-I 和 MetaRNN-I 在数据集 IS-L9 上的指标均有小幅度的提升。特别是对 CADD 而言，在面对长度大于等于 12 bp 的数据集 IS-L12，其性能提升的趋势依然持续。综上，不同长度的框内突变对预测方法的性能确实存在影响。具体来说，CAPICE、VEST-I 和 FATHMM-I 面对不同长度的框内突变数据时，均能维持着稳定的性能。而 CADD、MutPred-I 和 MetaRNN-I 在处理长度中等及偏上的框内突变时，其预测性能有所提高。

本研究从突变数据平衡性、突变基因分布以及突变长度这 3 个关键维度，对现有的 8 种致病突变预测方法在癌症框内突变上的性能进行了系统性评估。评估结果显示，广谱性方法 VEST-I 和 CAPICE 在多个基准数据集上均保持着高水平的 AUC 和 AUPR 指标，彰显了在癌症中框内突变预测方面的突出性能。FATHMM-I 在基准数据集上的 AUC 指标保持在 0.800 以上，表现出了较高的预测准确性。特异性方法中，MetaRNN-I 的 AUC 指标处于中等水平，相较于其他方法，性能表现良好。综合评估结果，本文得出结论：VEST-I 和 CAPICE 在不同维度上展现出稳定出色的预测性能，适用于识别癌症中的驱动框内突变。

表 2 不同长度数据集上的 AUC 和 AUPR 指标

方法	IS-L3		IS-L6		IS-L9		IS-L12	
	AUC	AUPR	AUC	AUPR	AUC	AUPR	AUC	AUPR
CADD	0.664	0.821	0.596	0.707	0.829	0.887	0.887	0.896
CAPICE	0.916	0.935	0.831	0.877	0.786	0.814	0.908	0.854
SIFT-I	0.610	0.636	0.519	0.642	0.607	0.632	0.702	0.500
VEST-I	0.927	0.967	0.846	0.923	0.966	0.981	0.981	0.979
FATHMM-I	0.816	0.863	0.792	0.889	0.855	0.917	0.872	0.871
MutPred-I	0.691	0.811	0.696	0.763	0.863	0.898	0.708	0.735
MetaRNN-I	0.735	0.841	0.729	0.853	0.761	0.879	0.719	0.783

2.1.2 一致性分析

由于 MT2021 没有提供具体的致病分数，因此 2.1.1 小节指标分析中未涉及该方法。为了解决这一问题，同时明确不同预测方法，尤其是 MT2021 预测结果的可靠性，本文借鉴了文献 [26] 的工作思路。具体来说，根据预测方法的输出，采用二元标记的方式，将预测为驱动突变的结果标记为 0，预测为乘客突变的结果标记为 1，并基于这些二元分类结果进行了各方法间的一致性计算。

图 3 展示了在具备所有预测方法致病分值的数据集 IS 上不同方法间的一致性。方法与自身间一致性系数默认为 1，在图主对角线上不作显示。结果显示，VEST-I 与 CAPICE、VEST-I 与 FATHMM-I、CAPICE 与 FATHMM-I 这 3 组方法间的一致性系数分别为 0.843、0.812 和 0.786，这些数值表明了它们之间具有较高水平的一致性。MT2021、VEST-I、CAPICE 间一致性系数分别为 0.528、0.616，整体上处于中等水平，这说明了 MT2021 对癌症驱动框内突变预测结果的可靠性需增强。值得注意的是，MT2021 与 CADD、MutPred-I、MetaRNN-I 间的一致性水平较高，相关系数分别为 0.790、0.856 和 0.852，这表明在与其他方法结合使用时，MT2021 能够提供一定的互补信息。因此，为了获得更加准确的研究结果，推荐将 MT2021 与其他方法的预测结果结合起来考虑。此外，本文发现 VEST-I、CAPICE 和 FATHMM-I 能对数据集 IS 中的大部分框内突变作出相同预测，这一结果表明这些方法在预测癌症驱动框内突变方面的可靠性较高，对突变的致病性评估具有较好的共识。

为了进一步探究不同预测方法在特定条件下的一致性，分别对分布在癌症基因上的数据集 IS-D 和长度等于 3bp 的数据集 IS-L3 上方法间的一致性系数进行分析。与数据集 IS 相比，CAPICE、VEST-I 和 FATHMM-I 在数据集 IS-D 和数据集 IS-L3 上的一致性有所提高，这表明这些方法在预测癌症基因上或短长度框内突变时，其预测结果的可靠性得到进一步提升。然而，MT2021 与前述方法间的一致性却有所下降，这表明该方法在相同条件下预测结果的可靠性可能需要进一步增强。此外，特异性方法在这些数据集上与 CAPICE 和 VEST-I 的一致性系数也有小幅降低，这反映了这类方法在特定类型的突变预测上可能需要更多优化。基于这些分析结果，本文建议在应用突变预测方法时，应考虑突变的基因分布和长度特征等特定因素，以获得更加可靠的预测结果。

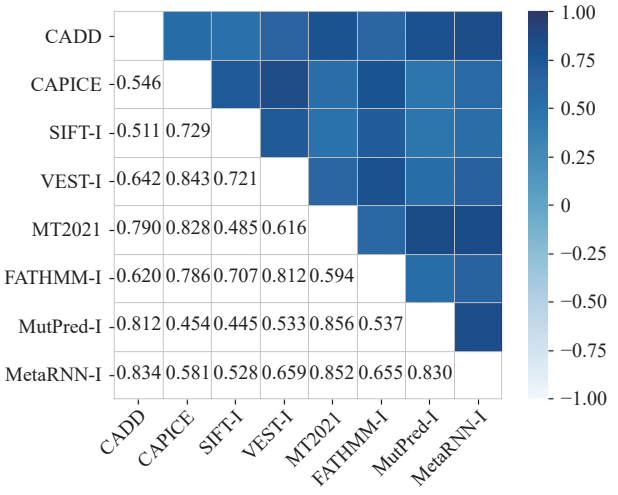


图 3 数据集 IS 上预测方法间的一致性

2.1.3 致病阈值分析

癌症基因组相关研究中，预测阈值是判断突变是否为驱动突变的一个重要依据。为深入了解致病阈值下预测方法的样本分类能力，对数据集 IB 上不同方法预测分数的分布直方图进行了分析，如图 4 所示。直方图的横坐标左右两端分别为分值的低分和高分区域。可以看出，CAPICE 中负样本（乘客突变）的分数主要集中在 0 附近，这反映了其在区分负样本和正样本（驱动突变）方面的准确性。VEST-I 的分数清晰的分布在 2 个区域：代表乘客突变的低分区域和代表驱

动突变的高分区域，显示其良好的分类能力。FATHMM-I 的致病分数也呈现出类似的分布趋势。然而，特异性方法 MetaRNN-I 致病分数的分布需要进一步优化，多数样本的分数在 0.2 附近，且存在分数混合的现象。特别地，CADD 和 MutPred-I 的预测分数分布较为混乱，多个区间重合现象明显，这可能是这 2 个方法在性能上表现一般的原因之一。

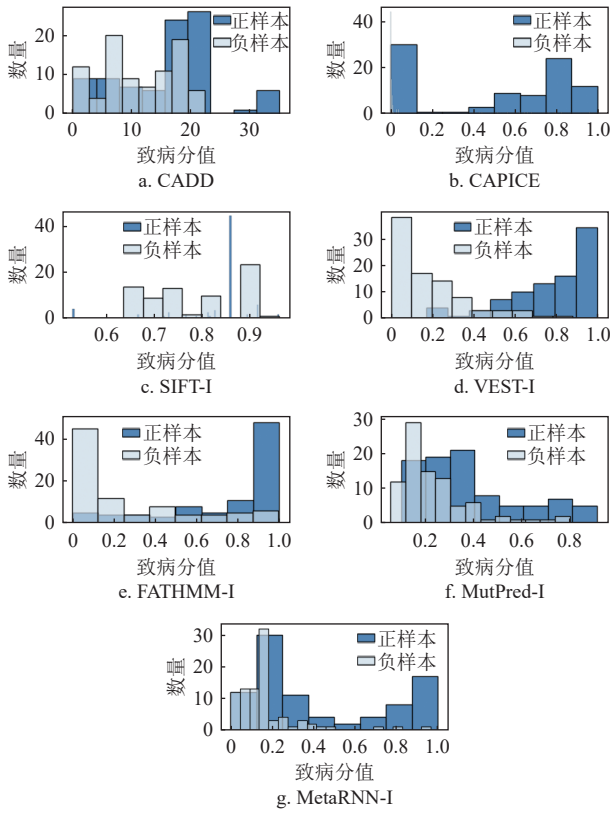


图 4　不同预测方法致病分数分布直方图

进一步地，参考文献 [27] 的研究思路，基于数据集 IB 为不同预测方法计算推荐的致病阈值，并探究不同致病阈值下方法样本分类能力的变化。结果显示，SIFT-I 的推荐阈值为 0.842。在所推荐的阈值下，SIFT-I 的敏感性（sensitivity, SEN）指标有所下降，这表明该阈值一定程度上限制了 SIFT-I 对正样本的识别能力。显然，将这一阈值作为 SIFT-I 在数据集 IB 中预测癌症框内突变致病性的判断依据是不合适的。此外，所有方法在推荐阈值下的 SEN 均有不同程度的变化，这意味着推荐阈值对预测方法样本分类能力存在一定影响。值得注意的是，推荐阈值的计算依赖方法的预测结果，具体数值在不同的数据集下会有所变动。因此，在处理多个数据集时，采用统一的推荐阈值作为癌症驱动突变（正样本）判断的唯一依据可能不是最佳选择。

综上所述，本文对致病突变预测方法的性能进行了评估。首先，从数据平衡性、基因分布和突变长度分布方面评估了预测方法的性能；其次基于多个基准数据集计算分析各方法间的预测一致性；最后探究了这些方法在预测阈值下的样本分类能力。这些综合性的评估工作证实了几种现有预测方法，特别是 VEST-I 和 CAPICE 在识别癌症驱动框内突变方面的适用性和可靠性。这些方法不仅在预测准确性上表现出色，而且在不同数据集和条件下保持了较高的一致性和样本分类能力。

2.2　癌症驱动框内突变预测及分析

随着癌症基因组研究工作的深入，涌现了以 COSMIC 为代表的大量未被解析的框内突变。本文挖掘了这些框内突变中潜在的癌症驱动突变，并探索其和癌症发生发展的联系。具体流程是，首先基于前述 8 种方法对数据集 M 中的 6 142 条突变数据进行致病分值量化。然后，基于 CAPICE、VEST-I、FATHMM-I 和 MetaRNN-I 的结果筛选出一致被预测为驱动框内突变的数据，共计 548 条，记作数据集 MP（Pathogenic）。选择这些方法的原因是：在 2.1.1 的性能评估中，CAPICE 和 VEST-I 在最大的基准数据集 IS 上的 AUC 均接近 0.900，表明了在识别癌症驱动框内突变方面的高准确性和可靠性。FATHMM-I 在数据集 IS 上的 AUC 值为 0.840，略低于前两者，性能表现依然可观。MetaRNN-I 的 AUC 指标大于 0.730，表现出良好的预测性能。将其纳入分析，可以提供特异性方法的视角，与广谱性方法形成互补，从而获得更全面的实验结果。最后，从突变长度、涉及基因、参与生物通路方面分析了数据集 MP 与基准数据集 I 中的 203 条正样本（记为数据集 IP）两者间的相似性。

图 5 详细展示了数据集 MP 和数据集 IP 中框内突变的长度分布。图中横坐标大于 0 的值代表了插入类型的框内突变长度，而小于 0 的值的绝对值则表示缺失类型的框内突变长度。可以看出，对于数据集 IP 而言，突变长度集中在 3bp 的区域，这和文献 [28] 在人类基因组研究工作中观察到的结果是一致的。同样地，在数据集 MP 中，超过 60% 的突变长度也集中在 3 bp，这意味小长度的框内突变在癌症的发生发展可能起着关键作用。此

外，数据集 MP 和 IP 在单一长度的框内突变数量上都显示出随长度增加而减少的趋势，这种现象一定程度上可归因于 DNA 修复机制的影响[29]。通过对比这 2 个数据集，本文发现数据集 MP 中的潜在驱动突变与已知的真实驱动框内突变在长度分布上整体是相似的。

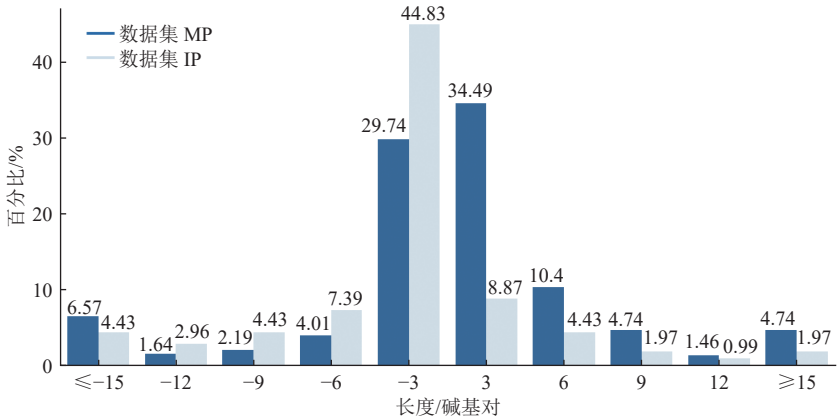


图 5 数据集 MP 和 IP 中突变的长度分布

涉及基因的种类及数目是分析一组驱动框内突变数据的一个关键依据。因此，本文统计了数据集 IP 和 MP 的基因分布情况。对数据集 IP 而言，其突变数据不均匀地分布在 91 个基因上，其中包含 58 个癌症基因。对数据集 MP 而言，与之相关的基因中也有 70 个癌症基因。单条突变上可能涉及多个基因，同时为避免 2 个数据集突变数目不平衡带来的影响，特备对比了其中仅分布在癌症基因上的突变以及基因本身的数量。具体来看，数据集 IP 中有 168 条框内突变仅分布在癌症基因，这些突变共关联了 58 个癌症基因。尽管数据集 MP 中的突变总数偏多，仍有 134 条突变被鉴定为仅位于癌症基因上，并且这些突变涉及的癌症基因数目与数据集 IP 一致。综合这些结果，数据集 MP 和数据集 IP 在癌症基因的涉及方面具有显著的相似性，这意味着数据集 MP 中的潜在驱动突变在基因层面与已知的真实驱动突变共享相似的特征。

为进一步探索数据集 MP 和数据集 IP 中框内突变生物学意义，基于这 2 个数据集进行了 KEGG 通路富集分析[30]，部分通路和经 Boferroni 校正法校正后的 *P* 值如表 3 所示。可以观察到，数据集 MP 中突变所在基因最显著富集的通路是 hsa05200: Pathways in cancer。hsa05200 是一个与涉及多种癌症通路的基因集，这说明这些潜在的驱动框内突变涉及的基因与癌症的发生发展密切相关。除 hsa05200 外，数据集 MP 上其余生物通路均和特定类型的癌症相关。此外，数据集 MP 所在基因富集的生物通路中有 25 个和数据集 IP 上的相似，这意味着这些潜在驱动突变基因可能通过参与相同的生物通路来推动癌症发展。值得注意是，潜在驱动框内突变相关基因在甲状腺激素信号通路上显著富集，文献 [31] 也证明了此通路在癌症进程中的影响。从参与生物通路角度出发，本文发现数据集 MP 中潜在驱动框内突变与真实致病框内突变类似，与癌症的发生联系紧密，这为后续开展其分子机制的功能实验并应用于癌症治疗提供了依据。

表 3 数据集 IM 和 IP 的 KEGG 通路富集分析

数据集MP		数据集IP	
KEGG通路	*P*值	KEGG通路	*P*值
hsa05200:Pathways in cancer	6.56×10^{-6}	hsa03440:Homologous recombination	2.80×10^{-14}
hsa05210:Colorectal cancer	6.56×10^{-6}	hsa03460:Fanconi anemia pathway	8.21×10^{-10}
hsa05226:Gastric cancer	6.56×10^{-6}	hsa05200:Pathways in cancer	6.57×10^{-6}
hsa05213:Endometrial cancer	6.56×10^{-6}	hsa05224:Breast cancer	9.14×10^{-5}
hsa05230:Central carbon metabolism in cancer	6.56×10^{-6}	hsa04218:Cellular senescence	1.20×10^{-4}
hsa05215:Prostate cancer	7.57×10^{-6}	hsa05210:Colorectal cancer	1.32×10^{-4}

综上所述，本文首先利用了 VEST-I、CAPICE、FATHMM-I 和 MetaRNN-I 这 4 个预测方法的结果，整理了 548 个潜在的癌症驱动框内突变。然后，从多方面将这些突变和已知的癌症驱动框内突变进行对比。结果显示，无论是在突变长度分布、涉及的癌症基因，还是在参与通路方面，2 组突变数据均表现出显著的相似性。基于这些结果，本文证实了通过预测方法挖掘的癌症基因组中潜在驱动框内突变作为癌症驱动突变样本是合理的。

2.3　功能注释数据库

众多突变数据库和基因组分析工具过于分散，致使研究人员获取全面的突变信息面临挑战。部分致病突变预测方法的本地复现，对个人用户的计算能力有一定要求。为给科研人员的研究工作创造便利，本文将前述 8 种方法对数据集 M 中 6 142 条癌症基因组框内突变量化的预测分数进行整理，并构建了一个用户访问友好、集成多种注释信息的线上数据库 dbCCID（http://dbccid.xialab.info/）。

需要说明的是，dbCCID 支持 2 个主要的人类参考基因组版本，确保了突变数据的兼容性，满足研究人员根据研究需求选择合适的版本。除了框内突变，dbCCID 同时支持移码突变注释信息的查询，以协助科研人员专注于癌症基因组插入缺失突变的研究。为用户可以从多角度进行数据挖掘，dbCCID 提供了包括按突变、按转录本、按基因、按癌组织在内的 8 种检索方式。通常情况下，ddCCID 会在每次查询后返回包括 8 种预测方法致病分数在内的 37 种不同信息，这些信息有助于研究人员深入理解突变的功能影响。

此外，为避免过长的查询等待时间，本文对可能涉及大量数据的搜索方式单次查询允许输入的信息数量进行了合理限制。同时，考虑到不同检索方式的反馈时间存在差异，推荐优先选用按突变（search by InDel）的方式进行检索。目前为止，dbCCID 共记录了 2 个主要参考基因组版本下的近 12 万条突变数据的相关信息，其中框内突变约占 10%。dbCCID 的建立，旨在提供一个用户友好的平台，为研究人员的研究工作创造了便利。

3　结束语

本文围绕癌症基因组框内突变功能注释及预测方法性能比较分析展开了一系列工作。首先，基于基准数据集对现有致病突变预测方法进行了系统地性能评估，验证了这些方法在识别癌症中致病框内突变的适用性和可靠性。这些方法中，VEST-I 和 CAPICE 的预测性能表现最为出色。其次，基于性能表现良好的 4 种方法的预测结果挖掘了癌症基因组中潜在的驱动框内突变数据，并与已知的驱动突变数据进行相似性对比，证实了其作为驱动突变的合理性。最后，构建了一个集成多种预测方法注释信息的线上数据库 dbCCID，该数据库收录了大量突变信息、支持多种检索方式，旨在为研究人员提供便利。本文为面向癌症框内突变预测方法的选择及开发提供了理论支撑。

参 考 文 献

[1] SUNG H, FERLAY J, SIEGEL R L, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. CA: A Cancer Journal for Clinicians, 2021, 71(3): 209-249.

[2] BRAY F, LAVERSANNE M, WEIDERPASS E, et al. The ever-increasing importance of cancer as a leading cause of premature death worldwide[J]. Cancer, 2021, 127(16): 3029-3030.

[3] BALL E V, STENSON P D, ABEYSINGHE S S, et al. Microdeletions and microinsertions causing human genetic disease: Common mechanisms of mutagenesis and the role of local DNA sequence complexity[J]. Human Mutation, 2005, 26(3): 205-213.

[4] STENSON P D, MORT M, BALL E V, et al. The human gene mutation database (HGMD®): Optimizing its use in a clinical diagnostic or research setting[J]. Human Genetics, 2020, 139: 1197-1207.

[5] PON J R, MARRA M A. Driver and passenger mutations in cancer[J]. Annual Review of Pathology: Mechanisms of Disease, 2015, 10: 25-50.

[6] YUE Z, CHU X, XIA J. PredCID: Prediction of driver frameshift indels in human cancer[J]. Briefings in Bioinformatics, 2021, 22(3): 1-9.

[7] NIAVARANI A, SHAHRABI F A, SHARAFKHAH M, et al. Pancancer analysis identifies prognostic high-APOBEC1 expression level implicated in cancer in-frame insertions and deletions[J]. Carcinogenesis, 2018, 39(3): 327-335.

[8] YAZAR M, OZBEK P. Assessment of 13 in silico pathogenicity methods on cancer-related variants[J]. Computers in Biology and Medicine, 2022, 145: 105434.

[9] 邵向阳, 徐伟文. 下一代测序 (NGS) 技术的发展及在肿瘤研究的应用[J]. 分子诊断与治疗杂志, 2016(5): 289-296.
SHAO X Y, XU W W. Development of next generation sequencing (NGS) technology and its application in cancer research[J]. Journal of Molecular Diagnostics and Therapy, 2016(5): 289-296.

[10] SATAM H, JOSHI K, MANGROLIA U, et al. Next-generation sequencing technology: Current trends and advancements[J]. Biology, 2023, 12(7): 997.

[11] TATE J G, BAMFORD S, JUBB H C, et al. COSMIC: the

catalogue of somatic indel in cancer[J]. Nucleic Acids Research, 2019, 47(D1): 941-947.
[12] MCDONALD E F, OLIVER K E, SCHLEBACH J P, et al. Benchmarking alphamissense pathogenicity predictions against cystic fibrosis variants[J]. PloS One, 2024, 19(1): e0297560.
[13] DOUVILLE C, MASICA D L, STENSON P D, et al. Assessing the pathogenicity of insertion and deletion variants with the variant effect scoring tool (VEST-Indel)[J]. Human Mutation, 2016, 37(1): 28-35.
[14] LI S, VAN D V K J, DE R D, et al. CAPICE: A computational method for consequence-agnostic pathogenicity interpretation of clinical exome variations[J]. Genome Medicine, 2020, 12(1): 1-11.
[15] STENSON P D, MORT M, BALL E V, et al. The human gene mutation database: Towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies[J]. Human Genetics, 2017, 136: 665-677.
[16] KARCZEWSKI K J, FRANCIOLI L C, TIAO G, et al. The mutational constraint spectrum quantified from variation in 141, 456 humans[J]. Nature, 2020, 581(7809): 434-443.
[17] KENT W J, SUGNET C W, FUREY T S, et al. The human genome browser at UCSC[J]. Genome Research, 2002, 12(6): 996-1006.
[18] RENTZSCH P, WITTEN D, COOPER G M, et al. CADD: Predicting the deleteriousness of variants throughout the human genome[J]. Nucleic Acids Research, 2019, 47(D1): 886-894.
[19] HU J, NG P C. SIFT indel: Predictions for the functional effects of amino acid insertions/deletions in proteins[J]. PloS One, 2013, 8(10): e77940.
[20] STEINHAUS R, PROFT S, SCHUELKE M, et al. Mutation taster 2021[J]. Nucleic Acids Research, 2021, 49(W1): 446-451.
[21] FERLAINO M, ROGERS M F, SHIHAB H A, et al. An integrative approach to predicting the functional effects of small indels in non-coding regions of the human genome[J]. BMC Bioinformatics, 2017, 18(1): 1-8.
[22] PAGEL K A, ANTAKI D, LIAN A J, et al. Pathogenicity and functional impact of non-frameshifting insertion/deletion variation in the human genome[J]. PLoS Computational Biology, 2019, 15(6): e1007112.
[23] GUNNING A C, FRYER V, FASHAM J, et al. Assessing performance of pathogenicity predictors using clinically relevant variant datasets[J]. Journal of Medical Genetics, 2021, 58(8): 547-555.
[24] CANNON S, WILLIAMS M, GUNNING A C, et al. Evaluation of in silico pathogenicity prediction tools for the classification of small in-frame indels[J]. BMC Medical Genomics, 2023, 16(1): 36.
[25] SONDKA Z, BAMFORD S, COLE C G, et al. The COSMIC cancer gene census: Describing genetic dysfunction across all human cancers[J]. Nature Reviews Cancer, 2018, 18(11): 696-705.
[26] PAYER B, LEE J T. X chromosome dosage compensation: How mammals keep the balance[J]. Annual Review of Genetics, 2008, 42: 733-772.
[27] WANG D, LI J, WANG Y, et al. A comparison on predicting functional impact of genomic variants[J]. NAR Genomics and Bioinformatics, 2022, 4(1): lqab122.
[28] LIN M, WHITMIRE S, CHEN J, et al. Effects of short indels on protein structure and function in human genomes[J]. Scientific Reports, 2017, 7(1): 9313.
[29] 刘博雅, 杨鑫, 任梦梦, 等. DNA 损伤修复机制——解读 2015 年诺贝尔化学奖[J]. 中国生物化学与分子生物学报, 2015, 31(12): 1322-1329.
LIU B Y, YANG X, RENG M M, et al. DNA damage repair mechanisms—interpreting the 2015 Nobel Prize in Chemistry[J]. Chinese Journal of Biochemistry and Molecular Biology, 2015, 31(12): 1322-1329.
[30] HUANG D W, SHERMAN B T, LEMPICKI R A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources[J]. Nature Protocols, 2009, 4(1): 44-57.
[31] LIU Y C, YEH C T, LIN K H. Molecular functions of thyroid hormone signaling in regulation of cancer progression and anti-apoptosis[J]. International Journal of Molecular Sciences, 2019, 20(20): 4986.

编 辑 刘飞阳