**哈尔滨工业大学计算学部**

**读书/论文笔记**

课程名称：生物信息学

课程类型：选修

项目名称：基因组变异检测 Prism

班级：2103601

学号：2021112845

姓名：张智雄

|  |  |  |
| --- | --- | --- |
| **设计成绩** | **报告成绩** | **指导老师** |
|  |  | **刘博** |

1. **论文的主要研究问题描述**

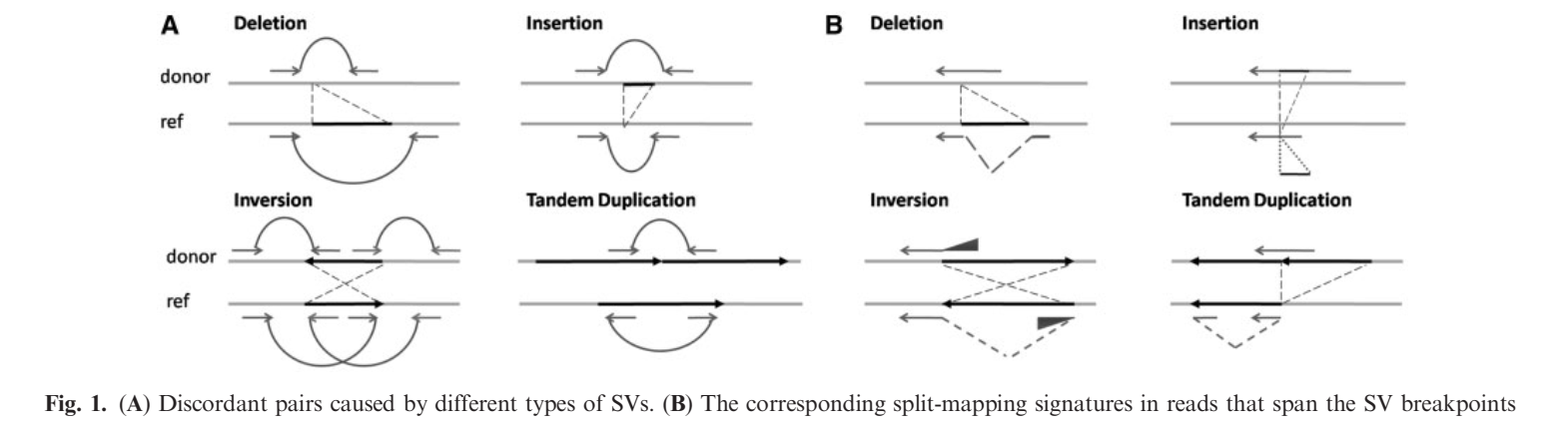
该论文的主要研究问题是在高通量测序技术的背景下，针对检测结构变异（SVs）存在的问题提出了新的解决方案。传统方法通常只能报告SV的大致位置，而不是准确的断点位置，这限制了对SV的准确识别和分析。论文指出，现有的方法主要基于覆盖深度或成对末端映射聚类，例如Pindel、Splitread和SVseq等，虽然能够识别SV的断点，但受到了识别大规模结构变异的限制。这些方法对于大型变异的识别受到了严重的限制，而且程序的运行时间和准确性与所设定的参数密切相关，这对于大规模的基因组数据处理来说是一个挑战。

因此，该论文提出了一种新的方法，即Pair-read Informed Split Mapping（PRISM），来解决这些问题。PRISM利用成对末端的不一致聚类来指导分割读取的映射，从而显著减少了分割映射的搜索空间，提高了准确性和运行速度。该方法还利用了Needleman-Wunsch（NW）算法的修改版，对分割读取进行基础级别的对齐，从而在存在SNPs、小插入/缺失和测序错误时实现高准确度。

PRISM的优势在于能够识别各种类型的SV，包括任意大小的倒置、缺失和串联重复，而且具有高灵敏性和准确性。论文通过对比先前的数据集和模拟实验，以及对PRISM结果的PCR验证，包括先前未表征的变异，证明了PRISM的优越性，整体精度达到了90%。

论文的研究问题在于如何提高对结构变异的准确识别和分析，以及如何克服现有方法的局限性，为基因组研究提供更可靠的工具和技术。通过提出并验证PRISM方法，论文为解决这一问题提供了新的思路和解决方案，具有重要的理论和应用意义

1. **论文的主要方法**

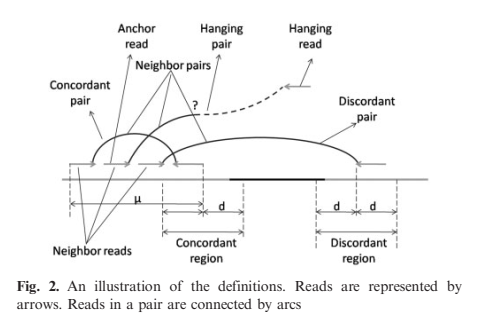


PRISM 方法依赖于对悬挂读对的彻底分析，它们具有异常的映射距离、方向或顺序，作为大型缺失、倒置和串联重复的信号。如果未映射读对跨越 SV 的断点，则应将其分割映射到断点的两侧。PRISM 通过两种不同的策略选择数据库和查询序列来处理具有和不具有不一致聚类的 SV。这些策略用于识别小的插入/缺失和悬挂对周围的不一致区域，以及不一致对的聚类来识别不一致区域。

**2.1 问题背景**

以下定义了 PRISM 方法中使用的术语，包括映射对、不一致对、一致对、锚定读取、悬挂读取、邻居、一致区域和不一致区域等，所有这些术语都在figure 2中进行了说明。

* 和分别是插入大小的**均值和标准差**。
* **映射对：**给定一个由读取和组成的读取对，如果和都被映射，则是一个映射对。
* **不一致对：**给定一个由读取和组成的映射对，如果满足下述条件，则读取对是一个不一致对。
* 和之间的映射距离大于
* 或的映射方向与测序方向不同
* **一致对：**如果映射对不是一个不一致对，则是一个一致对。
* **锚定读取和悬挂读取：**给定一个由读取和组成的读取对，如果被映射而没有被映射，带有一个或多个插入/缺失，或者具有未对齐部分（软裁剪），则是一个锚定读取，而是一个悬挂读取。
* **悬挂对：**如果读取对包含一个锚定读取和一个悬挂读取，则是一个悬挂对。
* **邻居：**给定读取对包含和，而包含和。假设不失一般性，被映射到，而被映射到。如果，和是邻居读取，而和是邻居对。
* **一致区域：**给定一个悬挂对，包含读取和，其中是锚定读取，被映射到，假设是加上插入大小。给定一个间隔（该值取决于SV断点预期位于的区域，在通常情况下约为），区间是对于的一致区域。当没有歧义时，称区间为一致区域。
* **不一致区域：**给定一个不一致对，包含读取r1和r2，其中r1被映射到pos1，而r2被映射到pos2。给定另一个读取对，包含读取s1和s2，其中s1是锚定读取，被映射到pos3，而s2是一个悬挂读取。如果r1是s1的邻居读取，则给定一个间隔d（该值取决于SV断点预期位于的区域，在通常情况下小于），称区间为来自P的的不一致区域。当没有歧义时，简单地将区间称为不一致区域。实践中，使用不一致对的聚类来识别不一致区域：每个聚类有两个脚，对应于一个对的两端。



**2.2 PRISM 工作流程**

运行 PRISM 包括五个阶段：映射读取、预处理映射结果文件、聚类不一致对、分割映射和调用结构变异（SVs）。

1. **映射读取。**使用 BWA以默认设置映射读取。生成一系列 SAM文件，并在后续阶段中进行处理。
2. **预处理。**从 SAM 文件中识别不一致对和悬挂对。不一致对在(3)中进行聚类。悬挂对按照锚定读取的位置排序，并在(4)中进一步使用。
3. **聚类。**识别所有不一致对，并使用 CNVer 中使用的成对读取聚类工具对其进行聚类。该程序使用一种贪婪算法，将具有相似映射距离和方向的对进行聚类。生成的不一致聚类与悬挂对一起在下一阶段中使用。
4. **分割映射。**这个阶段是 PRISM 的核心。PRISM扫描悬挂对，尝试分割映射。它首先尝试在一致区域对齐每个悬挂读取，允许插入或删除一个固定惩罚。如果存在不一致聚类，悬挂读取将分部对齐到一致和不一致区域。PRISM使用修改的NW算法进行分割映射，并以SAM格式呈现读取对齐。对于倒置或复制变异，PRISM修改原始读取序列，使其能够线性映射到基因组，并在SAM字段中存储修改。
5. **过滤和调用 SVs。**在对齐后，PRISM 从 SAM 文件中调用 SV 位点。PRISM 根据支持读取的数量和对齐分数对初始变异列表进行过滤。用户可以设置这些阈值以在灵敏度和特异性之间进行权衡。

**2.3 分割映射算法**

**2.3.1 修改的NW算法**

为了对分割映射的读取进行与参考基因组的对齐，使用修改的 NW 算法。对于删除，查询是读取序列，而数据库是参考基因组的两个片段，期望起始和结束都能映射。这两个区域可能是相同的。

为（矩阵1）和（矩阵2）构建两个动态规划矩阵如下。

* 矩阵1中的每个单元格的计算与传统的 NW 算法相同。
* 对于矩阵2，计算了一个额外的递归，该递归使用矩阵1中上一行的所有单元格的最大分数（在figure 3C 中说明）。

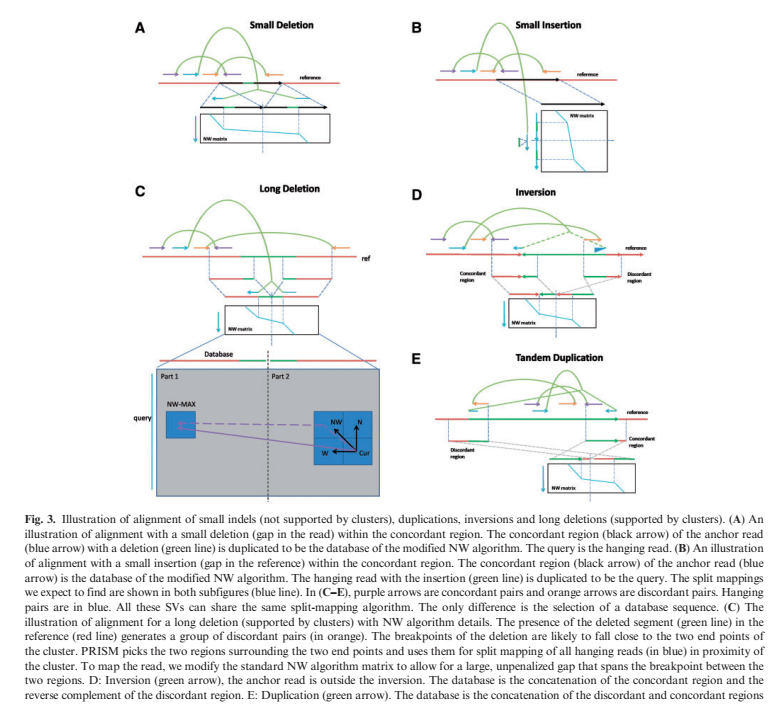
删除的修改后的 NW 矩阵使用以下递归方程构建。

使用此分数允许对齐从矩阵2“跳跃”到矩阵1，引入一个大的间隙，对应于读取中的分割。这个间隙的惩罚是一个与长度无关的常数。对于插入，算法类似，只是将基因组的单个区域与读取的两个副本进行对齐。对齐的算法相同，只是“跳跃”是在两个读取副本之间。

对于其中的满足以下条件：

而可以通过如下公式进行计算：

需要注意的是，跳跃分数不会计算在 和 中，因为删除后不能直接跟随插入，而删除后直接跟随跳跃可以被包括在跳跃中。与删除相同的算法用于倒置和串联重复的分割映射。插入的算法类似，只是读取被复制，而不是参考片段，最后，为了优化动态规划步骤的性能，使用了锚定对齐方法。



**2.3.2 选择修改 NW 矩阵的查询和数据库**

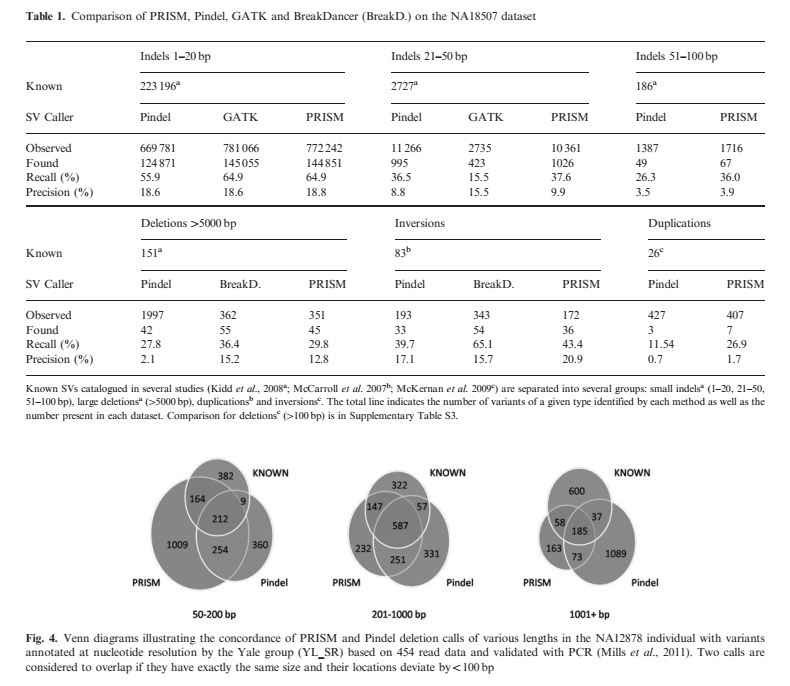
PRISM根据SV的类型和大小选择查询和数据库序列的策略。数据库可以是一致区域、不一致区域或它们的反向互补序列。查询可以是读取序列或其双倍（用于插入）。在一致区域内，针对每个锚定读取，PRISM尝试将悬挂读取对齐，允许插入或删除，以处理小的插入/删除。在一致区域和不一致区域之间，如果悬挂对有相邻读取属于不一致聚类，则PRISM尝试将悬挂读取分部对齐到一致和不一致区域，根据SV类型选择不一致区域。

**2.4 模拟数据集**

为了评估PRISM在已知真实数据集上的性能，研究者将已知的人类插入/删除植入到人类基因组染色体1中，并模拟了具有高斯分布插入的100 bp成对读取（平均500 bp，标准偏差30 bp）。他们采用了真实Illumina数据的测序错误模型，其中总体错误率为1%。

1. **论文的主要实验结果**

本研究旨在评估基于全基因组配对末端数据的PRISM方法在真实数据上的表现，并通过PCR实验验证其假阳性率和识别的新变异。实验分为两个部分：第一部分是对PRISM方法在真实数据上的性能评估，第二部分是针对PCR实验的验证。

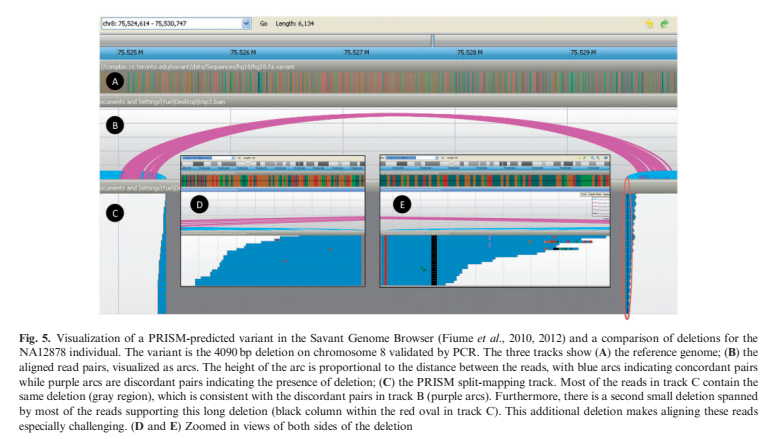


为了评估PRISM方法在真实数据上的性能，研究使用了两个来自不同人种的HapMap个体的全基因组配对末端数据集。NA18507个体（来自Yoruban人种）的数据集具有47倍的测序覆盖率，插入片段大小为500 bp，PCR重复读取率为7.4%；而NA12878个体（来自CEU人种）的数据集具有15倍的测序覆盖率，插入片段大小为300 bp，PCR重复读取率为3.6%。研究使用了hg18参考基因组，并参考了之前对这些基因组进行的多项研究。

针对NA18507基因组数据，PRISM方法成功检测到了大量插入缺失（784,319个1-100bp的插入缺失），其中约有19%的变异在之前的研究中已经被确认。与Pindel方法相比，PRISM方法能够识别出更多已知的变异，并且在检测更大变异时表现出更好的灵敏性。而对于CEU NA12878基因组数据，PRISM方法同样表现出优势，在识别较大变异方面具有显著的优势。此外，PRISM方法在识别倒位和重复方面也取得了令人满意的结果，尤其是在检测重复方面表现出优于Pindel的性能。

通过PCR实验对58个变异进行验证，结果显示，大部分实验成功验证了PRISM的预测。在实验过程中，仅有少数情况（2/58）出现引物无法正常工作或两个等位基因大小不匹配的情况。最终，在47/52（90%）的实验中，实验证实了PRISM的预测，表明其在识别新变异方面具有较高的准确性。值得注意的是，还发现了少量的变异与PRISM的预测不符，可能是由于引物错位或参考基因组的错位所致。

通过对PRISM方法在真实数据上的性能评估和PCR实验的验证，本研究得出了以下主要结论：PRISM方法在识别插入缺失、倒位和重复等结构变异方面具有良好的性能，尤其是对于较大变异的识别。此外，通过PCR实验验证，确认了PRISM方法在识别新变异方面的高准确性，这为进一步的基因组研究提供了可靠的数据支持。



1. **论文方法的优缺点分析**

PRISM相对于以往的方法在从HTS数据中检测结构变异方面有几个优势，包括结合配对读分析和分割映射来检测任意大小的变异的确切断点。PRISM使用的敏感对齐算法能够准确识别断点，即使在其他变异或测序错误附近也能如此。图5显示了一个例子，其中PRISM预测的一个大的缺失（经PCR验证）紧接着一个小的缺失，这使得跨越缺失的读对齐尤其具有挑战性。**PCR验证不仅确认了PRISM结果的总体高准确性，还证实了其能够识别以前方法无法访问的新变异。**