We propose Regional Hashing-based Alignment Tool (rHAT)

主要不同在于播种方式

The major difference is in the way of seeding.

我们将rHAT为真实的和模拟的SMRT数据集比对，从不同的原核生物和真核生物基因组。

bwa - sw将阅读的子串对齐到引用的基因组中，通过后缀trie遍历，找到近似的匹配，几乎没有什么位置。BLASR和bwa - mem在读取和引用基因组之间使用了精确匹配的短标记，该概率模型发现所有的短种子匹配的时间都超过预先定义的阈值。

rHAT认为一个主要的瓶颈时播种的过程

其他比对工具都是从种子的所有偏移量开始作为种子，然而，对于BLASR和bwa - mem，由于基因组的重复，在整个read中，密集的短种子会产生大量的点击量，这对于合并和优先级是很昂贵的。对于bwa - sw来说，成本也可能是巨大的，以使许多长时间的嘈杂的子串对齐到参考基因组中

此外，由于所有三个对准器都使用了Burrows-Wheeler Transformations（BWT）（Ferragina和Manzini，2000）用于索引参考基因组，执行LF映射的另外的成本（Langmead等，2009）来定位精确的基因组位置 的命中。 因此，对于BWA-MEM和BLASR，检索所有短种子的命中的开销也是不可忽视的。

本方法操作通过一个基于hash表的索引，区域哈希表(RHT)来实现。在扩展阶段，基于sdp的启发式设计是为了有效地将读取信息对齐到候选站点上。

对于rHAT，在设计基因组索引和read比对模块以提高性能时，充分考虑了SMRT读取的特征。

它也是指BLASR（Chaisson和Tesler，2012）的研究，表明对于具有中等错误率的长查询字符串（例如> 1000bp），它在实际位置发生许多短标记匹配的可能性很高

L is selected as 2048 bp for guaranteeing high number of k-mer matches.

The parameter k is typically configured as 11–15 bp, considering the error rate of SMRT reads.

rHAT subsequently performs a heapsort to prioritize the windows, and selects the top *M* windows as candidates.

The L/2 bp overlapping between neighboring windows is necessary, since it guarantees that a L/2 bp long substring of the read can be fully accommodated by a specific window for seeding

rHAT指提取reads中中间的一半长度L/2，那随着L增加丢失的长度也增加。

对于大多数类似于P5 / C3或SMRT测序的以上版本来说，它们的长度足够大，可以在实践中能跨越重复，这样种子比对不到正确位置的错误相对少见。

此外，由于端到端对齐的失败也可能是由假阳性候选站点造成的，因此该策略也有助于纠正此类错误，也就是说，对片段的重新播种可能会将片段重新对齐到它们的实际站点上。

默认配置

Firstly, rHAT can provide outstanding throughput (Fig. 2a, b). With the default setting (−l 11, −w 1000, −m 5, −k 13),

rHAT比对了大量碱基，这是碱基水平敏感，我们进一步检查了这些比对的细节，发现更多的reads有很大一部分的碱基被其他三个对准器剪掉了。

 rHAT aligned most bases, indicating that it also achieved the best base-level sensitivity.

However, rHAT aligned fewer reads with large clippings, which is the main reason of the higher base-level sensitivity of rHAT.

if too many SV-free reads were split aligned, the downstream analysis could be also affected in practice.

数据来源

Two SMRT P5/C3 release datasets respectively from H. sapiens and D. melanogaster genomes were used for evaluating rHAT on real SMRT data.