近似匹配，比BWA-MEM快290X，召回率为96%

mapping noisy PacBio reads (each ≥5 kbp in length) to the complete NCBI RefSeq database containing 838 Gbp of sequence and 60,000 genomes.

然而map原始序列仍然是一个瓶颈

However, mapping raw sequences continues to be a bottleneck for many applications.

然而重复种子会转换错误的映射，当有很高错误率时

However, repetitive seeds that do not translate to correct mappings combined with high sequencing error rates limit their scalability.

很多分析不需要详细的比对，这些应用包括

Such applications include depth-of-coverage analysis, metagenomic read assignment, structural variant detection, and selective sequencing

Broder[4]证明了在两个集合之间的Jaccard相似系数的不偏估计可以通过一个hash后的元素子集stretch来有效地计算。

Broder proved that an unbiased estimate of the Jaccard similarity coefficient between two sets can be computed efficiently using a subset of hashed elements called a sketch.

Schleimer等[25]提出了winnowing算法，它从每一个连续的文本窗口中挑选一个最小的hash元素(也称为minimizer[23])，以更快地估计web文档之间的本地相似性。

Schleimer et al. [[25](https://link.springer.com/chapter/10.1007/978-3-319-56970-3_5#CR25)] proposed the winnowing algorithm, which picks a minimum hashed item (also known as a minimizer [[23](https://link.springer.com/chapter/10.1007/978-3-319-56970-3_5#CR23)]) from each consecutive window of text as a means to more quickly estimate local similarity between web documents.

这些方法只是在经验上得到证实，MinHash Alignment Process [[3](https://link.springer.com/chapter/10.1007/978-3-319-56970-3_5#CR3)], minimap [[14](https://link.springer.com/chapter/10.1007/978-3-319-56970-3_5#CR14)], and BALAUR [[21](https://link.springer.com/chapter/10.1007/978-3-319-56970-3_5#CR21)].

These ideas have been used to develop new mapping and assembly algorithms for long reads such as the MinHash Alignment Process [[3](https://link.springer.com/chapter/10.1007/978-3-319-56970-3_5#CR3)], minimap [[14](https://link.springer.com/chapter/10.1007/978-3-319-56970-3_5#CR14)], and BALAUR [[21](https://link.springer.com/chapter/10.1007/978-3-319-56970-3_5#CR21)]. To date, the effectiveness of these approaches has only been demonstrated empirically.

该理论通过PacBio和MinION数据集进行验证，我们通过将PacBio metagenomic读到整个RefSeq数据库演示了我们的方法的可伸缩性。我们的算法的速度和空效率可以实现实时映射，与minimap相比，我们的方法对大的，重复的基因组保持较高的灵敏度。

The theory is validated using PacBio and MinION datasets, and we demonstrate the scalability of our approach by mapping PacBio metagenomic reads to the entire RefSeq database. The speed and space efficiency of our algorithm enables real-time mapping, and compared to minimap, our method maintains high sensitivity with better precision for large, repetitive genomes.

https://gss2.bdstatic.com/-fo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D204/sign=6f6d394f364e251fe6f7e3f89387c9c2/aa18972bd40735fac894d5cc98510fb30f24085a.jpg

The expected number of errors in a k-mer is k⋅ϵ，the probability of no errors within each k-mer, assumed independent, is



让c和n分别代表A中无错的和总共的k-mer，那么c/n的生存概率的期望值是

 Let A be a read derived from Bi, where Bi denotes the length |A| substring of reference B starting at position i. If c and n denote the number of error-free and total k-mers in A, respectively



S(A)是A中选择的元素个数，stretch越大，估计精度越大

where S(A) (called the *sketch* of A) is the set of the smallest *s* hashed items in A, i.e., S(A)=mins{Ω(x):x∈A}S(A)=mins{Ω(x):x∈A}

This estimate is unbiased provided S(A) is a simple random sample of A. Increasing the sketch size improves the accuracy of the estimate.

As an example, Fig. [1](https://link.springer.com/chapter/10.1007/978-3-319-56970-3_5#Fig1) illustrates this distribution for a read with known Jaccard similarity  j=G (ϵ=0.15,k=16) and sketch size s varying from 200 to 500

降低阈值用det，是在90%的置信区间边缘错误

we lower this threshold by δ to account for variation in the estimate. The parameter δ is defined as the margin of error in Jaccard estimation using a 90% confidence interval.

提出结合两种算法来计算J相似度

We present an algorithm to estimate J(A,Bi) efficiently using a combination of MinHash and winnowing techniques

我们的方法依赖于我们开发的索引和搜索策略来有效地修剪错误的映射位置。

Our method relies on an indexing and search strategy we developed to prune the incorrect mapping positions efficiently.

先用winnowing方法得到W(A)个minimizers，然后再用minhash

Let *W*(*A*) be the set of minimizers computed for read *A* using the winnowing method with window-size *w*. We sketch *W*(*A*) instead of sketching *A* itself.

先使用winnowing是为了minhash减少采样，通过经验发现这样评估相似度一样好

In contrast to the MinHash approximation (Eq. [2](https://link.springer.com/chapter/10.1007/978-3-319-56970-3_5#Equ2)), our estimator J(A,Bi) uses winnowing to reduce the sampling frame before picking the minimum hash values.

we empirically show in Sect. [8.1](https://link.springer.com/chapter/10.1007/978-3-319-56970-3_5#Sec12) that the quality of the Jaccard estimation using*J(A,Bi*)is as good as the MinHash estimation.



这个作为过滤条件，因为从前面推导后面，但不一定从后面推导前面，只能作为第一次过滤条件

下面出现排序

We begin by computing the minimizer hashed values Wh(A)Wh(A) by winnowing the read A, and compute the positions of their occurrence in the reference (line 4). Accordingly, L={pos:h∈Wh(A)∧⟨h,pos⟩∈W(B)}. Next, we sort the array L to process all the positions in ascending order.

如果满足过滤条件，说明至少m个入口（相同的k-mer）在 [i,i+|A|)中，那么说明L中从第一个到第m个入口的位置pos之间距离小于|A|，能够评估Bi用线性时间

If Bi satisfies the filtering criterion, there should be at least mentries in L with values between [i,i+|A|). It also implies that m consecutive entries should exist in L with positional difference between the first and mthmth entry being <|A|. This criterion is efficiently evaluated for all BiBi using a linear scan on L (lines 6–9).

这里弯L是保存并集hash元素，为了实现L用了一种C++数据结构，让hash的值作为key，如果在参考组和read都出现就置1，负责置0

We use L to contain the minimizer hashed values {h∈Wh(A)∪Wh(Bi)}. To implement L, we make use of the C++ ordered map data structure that supports logarithmic time insertion, deletion and linear time iteration over unique ordered keys. We keep the hashed value as the map’s key, and map it to 1 if it appears in both the reference and the read, and 0 otherwise.

扫描M而不是扫描H，是因为M中是按pos递增顺序排列的

The function getMinimizers gathers the reference minimizer hashes Wh(Bi) by sequentially iterating over M in the required position range and populating the minimizers associated with each Bi into the map L (lines 4, 8-9

选择窗口w和个数s大小

Jaccard相似性估计的stretch大小与窗口大小w成反比

The sketch size for Jaccard similarity estimation is inversely proportional to the window size w

在搜索过程中，较大的窗口大小可以改善运行时和空间需求，但也会对统计意义和估计精度产生负面影响。

假设两个文档相似度为p，那么相应位数相似的概率也是p，那么一个桶里全然同样的概率是p^r，不同样的概率是1-p^r，那么m个桶都不同样的概率是(1-p^r)^m。所以至少有一个桶同样的概率是1-(1-p^r)^m，我们能够依据我们想要的概率p去分配m和r。

窗口大小随着P或l0增加而增加，但是和emax成反比

The window size w increases with increasing pmax or l0, but has an inverse relationship with ϵmax

我们能够进一步降低采样比例，对超过l0的且满足p值限制的。

 From Fig. [3](https://link.springer.com/chapter/10.1007/978-3-319-56970-3_5#Fig3)(c), notice that we can further reduce the sampling rate (i.e. use a larger window size) for reads longer than l0l0 while still satisfying the p-value constraint.

我们提出了多级winnowing的概念，通过为每个输入选择自定义窗口大小来进一步优化算法的运行时间。

 Suppose Ww(B) denotes the set of winnowed fingerprints in the reference computed using window size w, then W2w(B)⊆Ww(B) [[25](https://link.springer.com/chapter/10.1007/978-3-319-56970-3_5#CR25)]

我们构造一个长度为5 kbp的随机序列，每个字符具有相等的概率为a、C、G或t，我们在每个位置引入替换错误时生成read，概率为0.15

We construct a random sequence of length 5 kbp with each character having equal probability of being either A,C,G or T. We generate reads while introducing substitution errors at each position with probability 0.15.

估计错误随着sketch尺寸增大而减小，用本文方法，对相似性评估与true jaccard相似性之间差异微不足道

 Similar to MinHash approximation, we note that the magnitude of estimation error reduces with increasing sketch size.

像mashmap,minimap使用winnowing来索引基因组，但不使用MinHash近似来估计Jaccard相似或核苷酸标识。

 Like mashmap, minimap uses winnowing to index the reference, but does not use the MinHash approximation to estimate Jaccard similarity or nucleotide identity.

Mashmap是设计找全的匹配

在运行时计算采样比例

The ability to compute the sampling rate at runtime gives mashmap its edge in terms of memory usage.

高精度，能避免重复产生的假阳性

Mashmap also achieves high precision, avoiding false positives on the repetitive human genome. Minimap’s low precision on human is largely driven by false-positive mappings to repetitive sequence, which could potentially be resolved with alternative clustering parameters.

避免重复的原因

It may be possible to avoid such mappings by considering the positional distribution of shared sketch elements during the second stage filter, or by adopting a local alignment reporting strategy like minimap.

未来研究，将分开read匹配

Future work aims to extend this method to split-read mapping, compressed reference databases, and additional error models.