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间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