首先找到簇的方法：

First, find clusters of short exact matches between the read and the genome using either a suffix array or BWT-FM index[[7](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-238#CR7)].

决定两个度量：

1. the number of matches of minimal length expected to exist between a read and the genome at a given sequencing accuracy and read length；
2. the number of false positive clusters the read is expected to have elsewhere in the genome.

为什么这样考虑？

If the chances of finding a match between the read and the genome are low, or if there are many regions a read may map to incorrectly with high identity, our proposed approach would not be feasible.

人类基因组重复性带来匹配困难

We next examine the repeat structure of the human genome to determine how difficult it is to map to due to the repetitive nature of the genome.

怎么度量两个序列之间相似度

we measure a different similarity metric on the human genome, the anchor similarity, where sequence similarity is measured as the number of shared anchors between the two sequences from the genome.

方法的可行性平衡点：

The feasibility of the method depends on the balance of having enough anchors to detect the correct interval to align a read to, vs. having so many anchors that clustering takes a prohibitive amount of time.

结果分成两部分，第一部分展示锚点比对到基因组上的理论，第二部分是实际PacBioRS数据比对的比较

 in the first, we examine characteristics of PacBioRS reads, and present theory on how these sequences contain matches that may be used to anchor alignments to the genome. In the next, we present a practical comparison of alignment methods on PacBioRS sequences.

限制锚点数量的方法，就是限制低多样性的锚点，可以用长锚点

One approach to limiting the number of anchors is to limit to a set of anchors of low multiplicity in the genome;  this is commonly done by using longer anchors.

长度为5的序列在基因组出现3百万次

Every word of length 5 occurs on average over 3 million times in the human genome

先不考虑错误分布，先考虑无错分布error-free，对于它是一个几何分布，且在PacBioRS测序大肠杆菌比对发现，有95%的为几何分布

Rather than focusing on the average case, it is more informative to consider the distribution of runs of error-free sequences;  for a uniform distribution of errors across a read, this is a geometric distribution.

先做假设，根据几何概率分布，在误差0.05范围内各种测序精度p时，想要获得一个无错的stretch延伸，文中给出需要测序的碱基数的评估

**Waiting length to sequence a word of length** ≥ k **at** ε = 0.05 **.** The waiting lengths to sequence a word of length ≥ k at ε = 0.05 at varrying accuracy. This gives an estimate of the number of bases required to sequence before having an error free stretch that may serve as an alignment anchor.

当测序错误已知时，能够测定一定数量anchors的概率

 Instead of using waiting lengths, it is possible to directly compute the probability of sequencing a certain number of anchors when the error rate is known.

定义一个函数作为分布M个错误的方法数，要求得到至少N个最大的子串中长度大于等于K

define **NumConfigurations*(M,N,K,L)*** as the number ways to distribute the positions of M errors when reading from the template such that there are at least N maximal substrings of length ≥ K not interrupted by error.

假定测序错误是均等随机的，则得到N个锚点的概率为：

 NumConfigurations(M,N,K,L)/(LM)

我们计算了L = 1000, and K = 15, 20, and 25时，这个概率值，研究比对anchors的数量，结果表明M = 200, 150, 100, and 50, corresponding to read accuracies of 80%, 85%, 90%, and 95%，且在测序精度为85%时，几乎所有的结果显示有至少10个长度为15的anchors

This indicates that with minimum anchor size K = 15, one would would expect to find at least 10 anchors at the correctly mapped interval in the genome.

当从基因组中重复抽样，会有一些密集的锚点anchors簇使reads匹配到基因组中，找到锚点后，在基因组中间隔使用详细的动态规划

由于人类中像ALU，LINE这种重复序列存在，计算需要太高以至于不能将所有的read去比对到重复区实例，在另一方面，如果只有一个限制性数量的详细匹配区域，那么找到具体位置机会也是很小的

基因组中重复的相似度通常是由两个序列的成对比对的百分数确定的

The similarity of repeats in a genome is typically defined by percent identity from a pairwise alignment of the two sequences

然而，具有高相似度的序列可能不会共享许多长stretch的精确匹配，这就是在使用基于锚的mapping。

我们引入了另一个度量指标:两个序列的锚相似度是在两个序列之间共享，在锚定距上有一定的约束固定长度、非重叠、有序锚点的最大数目。

we introduce an alternative metric: the **anchor similarity** of two sequences is the maximum number of fixed-length, non-overlapping, ordered anchors, shared between two sequences, with certain constraints on anchor spacing

锚点相似性需要两个参数：

Anchor similarity requires two parameters: K, the minimum anchor size; and δ, the indel rate, which may change the spacing between anchors.

随机抽取1million基因组中长度L=1kb的随机间隔，计算所有间隔之间相似度，间隔长度L1=（1+a）L=（1+0.15）\*1000=1150bp，假定插入删除比率为0.15，

在锚点15,20,25的的这些长度中，找到间隔中>=s-similar的数量，绘制直方图，

从这个图中，可以解释在使用锚点映射read时必须搜索的间隔数。

从四个图中，能得到间隔长度，所选的锚点长度和匹配上的锚点数量s-similar，了解到97.5%的样本需要至少20个锚点相似，才能将候选位置限定在100及以上很小范围。

我们比较人类基因组中锚点相似性值的分布和匹配锚点期望分布

We compared the distribution of values of anchor similarity from the human genome with values ofNumConfigurations(M,N,K,L)/(LM) to see how the mapability of sequences compares to the expected distributions of matching anchors.

有些基因组间隔中的reads 与基因组其他间隔具有很低锚相似度，因此假的匹配簇就少，得到特异匹配，然而有的间隔reads与其他间隔之间具有高相似度，那就会有很多匹配簇。

表四表明对于任何reads至少有8个长度为20或更大的锚

The green points in C show the number of matching intervals when using a similar set of parameters: at least 10 anchors of length 20. Importantly, 95% of the samples match uniquely in the genome.

然后模拟多个物种基因组，发现95%的长为1000碱基的reads能够正确匹配到基因组中

For example, 95% of 1000-base reads from the human genome simulated with a 15% error rate map to the correct location in the genome.

在图四中，K=15时，选择>=1-similar的相似度时，只有22%的reads能将候选间隔位置限定在100左右，但如果相似度选择为10时，即两个间隔之间锚点至少十个相似时，那么在K=15时，就能让超过90%的reads确定特定的匹配

As shown in Figure[4](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-238#Fig4)B, a read with a 15% error rate has a 97% chance of having 10 anchors of length 15 or more. The anchor similarity corresponding to these reads uses parameters δ = 0. 15,L = 1000, and k = 15, and is shown by the red curve in Figure[5](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-238#Fig5)A. Over 90% of the sampled intervals only have one location with at least 10 anchors of length 15, indicating they map uniquely under this repeat under this repeat metric. The other two genomes, E. coli, and A. thaliana, are shown for guidance