The new strategy chooses the mapped genomic location for the read directly from the seeds.

 When the read length is <160 bp, overlapping subreads are used.

通过各种方式扩展种子的流行对齐方式包括Bowtie（3），Bowtie2（4），BWA（5），Novoalign（http://www.novocraft.com，2013），Maq（6）和MrsFast（7）。 扩展步骤通常包括回溯（3,5）， Smith–Waterman 动态规划（4-6,8）或Needleman-Wunsch动态规划（9）（Novoalign）。

一般情况下，动态规划的运行时间以读取长度(11、12)进行四次递增。

已经提出了许多巧妙的算法，使扩展步骤更有效，包括有界回溯(5)和动态规划的带状和位向量版本(13,14)。

Many clever algorithms have been proposed to make the extension step more efficient, including bounded backtracking (5) and banded and bit-vector versions of dynamic programming (13,14).

新策略种子更短，无错配，

The new strategy differs from previous procedures in a number of ways: the subreads are shorter and more numerous than conventional seeds; they are mapped without mismatches; and the local alignment is determined directly by counting subreads without further intermediate steps.

数据集

We used a 1000 Genomes data set, a SEQC data set and simulation data sets to compare alternative methods for read mapping and exon–exon junction detection.

Real quality scores, extracted from a 101 bp SEQC Sample A read data set, were used for simulation reads. Sequencing errors were generated according to the quality score each read base has.

通过修改的人类参考基因组GRCh37(hg19)生成了100个碱基对模拟数据，其中80个bp或更长的重复序列被删除，以使每个模拟读取都有一个独特的已知的映射位置

产生模拟数据集

In addition to the simulation data sets generated from the filtered human genome, we generated a 101 bp simulation data set from the unfiltered human genome, in which repetitive regions were kept. This data set contained indels in it. We also used Mason (32) and Art (33) to generate two extra simulation data sets.

建立参考组索引，每隔三个取一个k-mer，那么read就需要扫描三次

To build an index, 16 bp sequences were extracted from the reference genome in every three bases

Correspondingly, each read has to be scanned three times for the mapping,

实际的峰值存储器使用量将略高于这些值，因为当执行比对时，整个基因组的序列也被加载到存储器中。 索引构建功能提供将索引分解成多个部分的选项，以减少内存占用（任何时候只有一部分存在于内存中）

测试平台所有的程序都在HP刀片超级计算机上进行了测试，其中包括16个Xeon 2.93 GHz CPU核心和128 GB的内存。

我们不允许不匹配，而是保持子读相对较短，在灵敏度和准确性之间达到一个很好的平衡。

最佳种子长度

Tests show that a range of subread lengths, from 10–25 bp, work well from this point of view (data not shown).

使用16-mer

Subread uses subreads of length 16 because that is in the optimal range for sensitivity and accuracy and because sequences of this length will fit exactly into a machine word on a 32-bit computer system or into half a word on a 64-bit computer system.

为了使种子策略得到有效的工作，每一个种子（k-mer）都有必要具有合理的特异性，所以对应于高度重复或过度共同序列的种子路由集中删除。

我们定义为无信息量的任何子读，其序列在参考基因组中发生> 24次。这些subreads信息subreads因此发生在参考基因组≤24次。

仿真结果表明，较高的截断阈值能导致较高的映射灵敏度，但精度较低，在本研究中，除另有说明外，还使用了24次重复的截止时间。Subread在索引构建程序中提供了一个选项(' -f ')，这样用户就可以在适当的时候调整这个阈值。

每个读取的最大共识集确定其映射位置。 当没有唯一的最大共识集合时，由于两个或多个映射到不同位置的共识集具有相同的投票数，所以选择了覆盖基因组更多基础的一致性集。 如果还是一个领带，那么它是以MQS为基础的，或者通过读取和每个候选区域之间的汉明距离来破坏。

The largest consensus set for each read determines its mapping location. When there is no unique largest consensus set, because two or more consensus sets mapping to different locations have the same number of votes, the one covering more bases in the genome is chosen. If there is still a tie, it is broken on the basis of either MQSs or by the Hamming distance between the read and each candidate region.

下面说明subread匹配到参考组

Encoded value for each subread is used as its key in the hash table. The key’s value gives the chromosomal location/s in the genome to which the corresponding subread is perfectly matched (no mismatches allowed).

*consensus set*.是指那些投到基因组中同一个位置的subread的集合

Any set of informative subreads that vote for the same mapping location for the read is called a *consensus set*. In general, a read will have more than one consensus set.

一致性阈值是指每个区域所需要的最小的投票数

The consensus threshold is the minimum number of subreads (votes) required for reporting a mapping location.

在模拟中设置每个subread数量，以及一致性阈值

Numbers of subreads ranging from 7 to 28, and consensus thresholds ranging from 10 to 70% of the number of subreads, were examined for the mapping of 10 million 101 bp reads.

一致性阈值增加，敏感性降低，精度增加

 Not surprisingly, sensitivity decreased and accuracy improved with the consensus threshold increase for any fixed subread number

将一致性阈值设置为~30%达到最好效果

However, setting the consensus threshold at ∼30% of the subread number gave good performance with respect to both accuracy and sensitivity across a wide range of subread numbers and cut-offs for removing uninformative subreads

这个数据集设置的测序错误率在0-10%

在双端测序，把成对的reads中mapping到基因组中最好那一条read匹配上的位置作为锚点，然后另一条read的距离信息作为参考，且在一致性阈值中设置也相对宽松，可以低至为1，如果出现多个候选位置，那么就将距离信息作为度量。

如果两个端点的映射位置都不满足距离标准，Subread和Subjunc仍然报告这两个读取的映射位置，只要它们满足最小数量的一致性序列数(默认为3)。这些reads可能有一个片段长度大于或小于指定的片段长度，或者它们可能起源于嵌合（奇异）序列

建立参考组索引需要1个小时

It takes about 1 hour to build an index for human or mouse genome on a Linux computer with a few gi

当投票多个位置具有相同的票时，将选择覆盖碱基最多的那个

When there is no unique largest consensus set, because two or more consensus sets mapping to different locations have the same number of votes, the one covering more bases in the genome is chosen.

要绘制paired - end读取，Subread(或Subjunc)首先映射两个相同的读取一对个体，就好像它们是单端末端。一个读取的映射位置与来自同一对(如更大的)的另一个读取相比，该映射具有更高的可信度吗获得的选票数量，被确定为“碎片的位置”。然后用paired - end距离信息帮助地图准确地映射。由于对于距离信息的可用性，我们使用一个轻松的共识阈值来调用未映射对的映射位置，低到只有一个投票

gabytes of memory.