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。

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.

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 gigabytes of memory.