cDNA的大规模平行测序已经能够对转录组进行深入而有效的探测。目前的从这样的数据重建转录本的方法通常依赖于将reads对齐到参考基因组，因此不适用于具有部分或者参考基因组缺失的样品。这里我们介绍从头组装全长转录本的Trinity方法，并在尚未提供参考基因组的裂殖酵母、小鼠和粉虱的样品上对该方法进行评估。通过有效地构建和分析de Bruijn 图，Trinity可以重建大部分的转录本，包括来自最近复制的基因的可变剪切异构体和转录本。与其他从头组装的转录组组转软件相比较，Trinity能够在大范围的表达水平上恢复更多全长转录本，其敏感度与依靠基因组对比的方法相似。我们的方法为任何样本的转录组组装提供了统一的解决办法，尤其是针对缺失参考基因组的样本。

大规模的并行cDNA测序（RNA-seq）的最新成果为从器官和组织中得到大量的转录组数据提供了一种经济有效的方法。原则上，从转录本起始位点到转录本结束位点之间的完整的和连续的mRNA序列，这可以使我们识别多个可变剪切变体的同种型的所有表达的转录本。然而，使用具有相当大的测序错误率的短reads重建全长转录本拥有实质性的计算挑战：（1）一些转录本的覆盖股很低，而另外一些转录本则高度表达。

（2）由于测序偏差，整个转录本的read的覆盖度是不均匀的。

（3）经由一个高度表达的转录本测出来的序列错误的read，可能比经由一个表达水平不那么高的转录本测出来的正确的read的丰度要大。

（4）由相邻基因座编码的转录物可能会重叠并且会因此错误地融合形成嵌合转录本。

（5）由于可变剪切，数据结构需要适应每个基因座上上的多个转录本。

（6）在不同基因中重复的序列带来了模糊性。

一个成功的方法应该要攻克研究过程中遇到的每一个挑战，我们的程序要能适用于复杂的哺乳动物的基因组和基因密集的微生物基因组，并且能够重建不同大小、不同表达水平和不同蛋白质编码能力的转录本。

对于转录组重建有两种可选择的计算方法。映射优先法，如Scripture和Cufflinks。映射优先法首先将所有的reads与参考基因组（未注释）匹配，然后合并具有重叠联配的序列，连接reads和配对末端的剪切位点。组装优先法（de novo），例如AbySS,SOAPdenovo，Oases，该方法使用reads直接组装转录本，随后如果有参考基因组的话可将其映射到参考基因组。原则上，映射优先法承诺最大灵敏度，但是需要依赖于正确的read到参考基因组的联配，可变剪切、测序错误和参考基因组的缺失或者不完整使得整个任务变得复杂。相反，组装优先法不需要任何的read和参考基因组的联配，这在基因组序列不可用、基因组有缺口、基因组高度片段化或者基因发生改变（如在癌细胞中）的情况下是具有很大优势的。

在过去几年的时间里，映射优先法发展地很成功，但迄今为止，在制定有效的组装优先法方面进展很是缓慢。随着reads数量的增加，确定哪些reads应该被连接成连续的contigs变得越来越难。de Bruijn图提供了一个优雅的解决办法，该图可以说是几个全基因组装程序的基础。在该图中，节点定义为有k个核苷酸长度的序列（即为k-mer，其中k比read的长度要短得多），并且如果两个顶点有k-1个核苷酸重叠并且序列数据支持这个连接的话，就用边连接这两个顶点。通过k-1个核苷酸的重叠，可以重建线性序列并且枚举所有可能的解决方案。对于转录组组装，图中的每一条路径便代表着一条转录本。应用于图结构的评分方案可以依靠最初的reads序列和配对信息来丢弃没有意义的转录本并且计算出所有的转录本。

在de novo组装RNA-seq数据中应用de Bruijn图面临着三个关键挑战：

1. 从大量的原始数据（数百万个剪辑对）中高效构建该图。
2. 定义合适的评分和枚举算法以恢复所有合理的剪接形式和旁系同源转录本。
3. 对有数据中的测序错误和其他人为因素引起的噪声提供稳健性。

特别应注意的是，测序错误会造成很大数量的节点是错误的，从而产生一个可能具有数百万条路径的巨大的图。

这里我们介绍Trinity，一个强大而有效的基于de novo重建转录组的方法。Trinity由三个软件模块构成：虫子、蝶蛹和蝴蝶，按顺序使用以处理大量的RNA-seq reads我们用两个已经注释好的物种的数据来评估Trinity，这两个物种一个是微生物（裂殖酵母），一个是哺乳动物（老鼠），再用基因组尚未被测序的一只昆虫（粉虱）对Trinity进行评估。在每种情况下，Trinity可以恢复大部分的已经表达的参考转录本，表现优于其他其他可用的基于de novo的转录组组装软件，并且其结果类似于依赖基因组对比的方法。

结果：Trinity是基于de novo进行转录组组装的一个方法。

与对基因组进行de novo（从头）组装相比较，de novo组装很少有大的连接序列图可以代表整个染色体上reads的连续性，在组装转录本数据时我们希望看到大量的独立的断开的图，每一个图都代表了非重叠基因座上的复杂性。因此，Trinity将这些序列数据分成许多独立的图，然后单独处理每一个图来提取全长异构体并梳理来自旁系同源基因的转录本。

Trinity的第一步，虫子，将reads组装为转录本的特定序列。这一步使用贪婪的基于k-mer的方法进行快速高效的转录本组装，仅恢复共享k-mer（可变剪切、基因重复或等位基因变异）的一组可选变体中的一个（即选择最好的一个）。第二步，蝶蛹。将相应的对应于可变剪切转录本的一部分或者对应于旁系同源基因的其他独特部分的contig聚集在一起。然后对相应的contigs的每个聚类创建一个de Bruijn图，每个图都反映了变体之间重叠的复杂性。最后一步，蝴蝶。分析由reads和read对应的上下文在对应的de Bruijn图中产生的路径，并报告所有的合理的转录本序列，分解可变剪切的同种型和由旁系同源基因产生的转录本。下面我们详细描述Trinity的每个模块。

虫子，贪婪高效地组装contigs。

这一步共有六个小步骤重建线性转录本contig。

（1）利用所有的reads序列构建k-mer字典（实际上，k=25）。

（2）从k-mer字典中移除可能有错的k-mers。

（3）选择k-mer字典中最频繁的k-mer（不包括复杂度很低，只出现一次的k-mer）作为种子进行contig组装。

（4）寻找出现频度最高的与种子k-mer有k-1个碱基重叠的k-mer，将该k-mer的碱基末端连接至不断增长的contig序列，以达到对种子进行各个方向的延伸的目的（一旦k-mer被使用，则在k-mer字典中移除该k-mer）。

（5）在各个方向上延伸序列直至不能再进行延伸为止，然后得到了线性contig。

（6）选择下一个最大丰度的k-mer，重复（3）到（5）的步骤，直到k-mer字典中的k-mer都被使用过。

仅由第一步得出来的contigs并不能完整地捕捉转录本的复杂性。例如，每个全长基因座只能产生一个可变剪切变体，同时可以产生可变剪切转录本的特定区域的部分序列。然而，其contigs确实保留了后续Trinity组件所需要的信息，以重建和搜索包含所有可能序列的整个图。事实上，除了从种子k-mer中排除的或者在比规定的最小长度还要小的contigs中丢弃的低复杂度和单一k-mer，第一步的contigs提供了基于重叠的de bruijn图的序列的完整表示，其中，每个k-mer在集合中都是独一无二的，并且其k-1的子序列隐含地定义了图的边。 该方法比一次性从所有的reads中计算一个完全的图要高效的多，并且它可以快速提供由reads中的k-mer强支持的contigs的有意义的中间输出。在contig延伸的过程中，通过消除单个k-mer作为初始种子，第一步可能因为序列错误进一步减少了k-mer聚体的包含。

蝶蛹，建立de Bruijn 图

这一步将第一步中最小重叠的contigs聚集为连接组件的集合，并且为每一个组件创建一个完整的de Bruijn图。每一个组件定义了可能来源于可变剪切形式或者密切相关的庞爰同系物的contigs的集合。第二步分为三个阶段：

1. 它递归地将第一步中的contigs分组为连接组件。如果contigs之间有完美的k-1个重叠的碱基并且如果跨越这些contigs的连接处的reads有最小数目，其中contigs在（k-1）-mer的两端都有（k-1）/2个碱基匹配，那么这些contigs将被分为一组。
2. 为每一个组件构建一个de Bruijn图，并使用k-1大小的序列作为节点，k定义为连接这些节点的边。定义de Bruijn图上边的权重为原始read集合中所支持的k-mer的数量。
3. 将每个read分配给共享最大数量的k-mers的组件，并确定每个read为组件贡献k-mer的区域。

蝴蝶，解决可变剪切和旁系同源转录本。

第三步通过协调第二步使用reads和配对末端生成的单个de Bruijn图来重建合理的、全长的、线性的转录本。它对剪接异构体和旁系同源基因重构了不同的转录本，并解决了由错误或者由转录本之间共享的序列大于k个碱基长度引起的模糊性。

第三步有两部分组成。第一部分被称之为图简化，它重复以下步骤：（1）合并de Bruijn图中的线性路径上的连续的节点，形成能够代表更长的序列的节点。（2）修剪较小偏差的边（由相对较少的reads支持），因为这样的边可能对应着测序错误。预计二倍体多态性要比测序错误更加频繁，并且可能会被保持下去。第二部分被称之为合理的路径评分，其使用贯穿图中潜在路径并且维持支持这些路径的reads的动态编程过程来由实际reads和reads对支撑的路径。因为reads和序列片段的长度普遍要比k大，因此可以解决歧义性并将路径的组合数量减少到实际转录本（被列为线性序列）的数量。

裂殖酵母的RNA-seq。

我们首先从裂殖酵母中产生了RNA-seq数据。裂殖酵母转录组具有相对较多的真核微生物的剪切，这些剪切包含较短的内含子（平均内含子长度为80.6bp）和致密转录本（平均基因间区域=仅基于编码基因的938bp）。为了最大限度地提高转录本的覆盖率，我们合并了来自四个生物情况（中期-对数增长、所有葡萄糖都被消耗之后的增长、晚期静止期、热休克）下的约1.54亿特异性链，有76个碱基的Illumina read序列。

重建全长转录本的敏感度限制。

接下来，对于给定特定的序列集合的数据，如果能从中重建出完美的注释的转录本，那么我们对这些转录本规定一个敏感度上限。基于特定的长度为k的低聚物的任何组装方法仅限被由那些RNA-seq数据集合中确切的k-mer组成表示的序列使用。为了确定这种经验灵敏度上线，我们根据所有的reads数据创建了k-mer字典并且鉴定了所有已知的参考的蛋白质编码序列。因为可以通过它的邻接和全长序列上的k-mer重叠进行填充，所以这些序列可以根据已给的reads数据重建全长转录。我们将这些序列集合称之为“Oracle 集”。因为这个集合包含了由k-mers覆盖的转录本序列，但不包含完整的reads，因此有些转录本虽然显示的是可以重建的但是实际上并没有被重建。相反，Oracle集仅反映了已被注释的已知基因和已知异构体，但这很有可能被低估，尤其是在哺乳动物中。不过，Oracle集提供了一个有用的敏感度基准。

在裂殖酵母的数据集中，几乎所有的（91%，4600/5064）参考蛋白质编码序列都出现在了Oracle集中（25-mer 字典，154M配对reads），几乎所有的编码转录本（98%）都在测量的条件下（>=0.5FPKM）被表达，与之前对酵母的研究结果是一致的。当我们通过随机次采样减少覆盖率时，Oracle集的大小在配对reads增至50M时达到饱和，我们选择这个读数作为后续研究的基准测试集。

Trinity恢复了大部分的裂殖酵母的转录本。

根据50M的reads对，Trinity完全重构了已注释的全长转录本的86%，其中包括94%的严格定义的oracle转录本。在未被完全重建的276条oracle转录本中，有90条（33%）至少超过他们长度的90%的转录本被重建，有177条（64）至少超过他们长度的50%的转录本被重建。

总的来说，Trinity产生了27841个超过100bp的线性contigs，这些contigs又被分为23232个组件。在这27841个Trinity产生的contigs中，只有2454个contigs没有被GMAP联配到基因组上。其中，30%的contigs与Uniref90蛋白质匹配上，其几乎90%为粟酒裂殖酵母蛋白质。

Trinity重建跨越广泛的表达水平和序列深度的全长转录本。例如，它精确地捕获了第二五分位（5-10%）中71%的基因的全长转录本，并在剩下的五分位中全长覆盖了81%-95%的已注释的转录本。同时考虑全长或者部分重建，Trinity重建了每个转录本上大部分的碱基片段。

在很多情况下，Trinity准确地解决了密切相关的旁系同源转录本的序列问题。在包含185个旁系同源基因的77个基因家族中，Trinity全长恢复了33个家族中的所有成员（68个基因），其他33个家族中至少有一个成员被恢复（发现了46个基因，缺失了45个基因），并且错过了剩下11个家族的所有26个基因，这些家族通常涉及到不高度表达的基因。裂殖酵母中大部分高度表达的转录本都是由具有相似序列的旁系同源基因得到的（例如，编码核糖体蛋白的基因），这个问题Trinity也已解决。

裂殖酵母中延伸的UTRs和长反义转录本。

与现有注释相比，Trinity延伸了312个转录本的5‘非翻译区，并延伸了543个转录本的3’非翻译区。同时，它还发现了3726个以前没有注释的5‘非翻译区和3415个3’非翻译区。

Trinity鉴定了1235个基因间位点上的2319条转录本，并将这些转录本作为新的转录序列，也鉴定了612条长反义转录本，这些转录本覆盖了相对应转录本长度的75以上，并且不可能来自邻近基因的扩展转录。130个基因间转录物和612个长反义转录物是多克隆的。尽管比已注释的蛋白质编码基因的表达水平要低，但是49条长反义转录本要比相应的感知编码的转录本的表达水平高出5倍。这表明了粟酒裂殖酵母在减数分裂中反义转录的调控作用，并且与以前的发现一致。

Trinity恢复了老鼠的大多数的表达的且带注释的转录本。

与酵母相比，哺乳动物转录组显著地表现出了更为复杂的选择性剪切模式，为了测试Trinity

测试不同亚型的能力，我们对来自C567BL/6小鼠原代免疫树突细胞的大约5260万个76bp的配对reads进行了测序。不同于粟酒裂殖酵母，小鼠已知基因（10724）中只有54的基因被鉴定为表达，其中，Oracle集确定了8358个基因是可以全长重构的。

Trinity报告指出，有48497个contigs要长于350bp，对应于7749个基因座来说，可以捕获到8185个转录本是全长转录本。在广泛的表达水平范围内，转录本可以全长恢复的比例和长度捕获的分数很高。

Trinity以与老鼠Oracle集一致的方式解决了剪接同种型和基因旁系同源的问题。Trinity从385个位点中发现了872个全长、可变剪切的同种型，并且在Oracle集中匹配了752个旁系同源转录本中的463个全长转录本。Trinity在5265个转录本中延长注释的5‘非翻译区，并且在305个例子中包含了一个或多个额外的5’非翻译区外显子。在2918个转录本中延长3‘非翻译区，在62个例子中增加了3个3’非翻译去外显子。UTR长度的差异通常是由于UTR的可变剪切的限制。

重建转录本的高序保真度。

我们通过将全长转录本联配到相应的参考基因上去来计算组装转录本的碱基错误率，并从最高联配得分中捕捉错配、插入、缺失的次数。在裂殖酵母中，错配率、插入率和缺失率均小于1/10000。在老鼠中，这样的概率大约是裂殖酵母的两倍，这样的结果反映了较低的转录本折叠覆盖率。由于原始的reads的错误率大约为1%，因此，Trinity解决了约99%的测序错误。

比较Trinity和其他方法。

我们通过几种方法比较了Trinity和其他几个组装方法的表现。首先，我们检查了每种方法重建为全长参考转录本的数量（灵敏度）。在粟酒裂殖酵母中，Trinity的表现要好于其他的从头组装程序，如AbySS、Trans-AbySS和SOAPdenovo，并且也要比映射优先程序Scripture和Cufflinks要好。Trinity在整个10M到150M的输入reads序列数据中表现良好，而其他的方法在大约50M或更小的输入时趋于峰值。在老鼠中，Trinity的表现要好于其他的从头组装程序，如AbySS、Trans-AbySS和SOAPdenovo，并且也要比映射优先程序Scripture和Cufflinks表现出更好的灵敏度。此外，Trinity和Cufflinks在广泛的表达水平中，他们的灵敏度得到了最佳调整。与Trinity不同的是，其他的一些从头组装算法在表达度最高的五分位中，其重建全长转录本的表现不是很好。

其次，我们评估了拼接模式检测的准确性。我们将所有重建的转录本映射回参考基因组，并考虑每个单独的内含子或者由映射定义的内含子组合。我们将由每种方法捕捉的注释的参考内含子的数量进行比较，并比较由每种方法的转录本定义的之前未被注释的内含子的数量有多种方法捕捉的未被注释的内含子都不太可能是假阳性的。在粟酒裂殖酵母中，Trinity鉴定了最大数目的参考内含子和1582个未注释的内含子，其中大部分推测是未注释的UTRs。其中，1174个内含子也通过至少一种方法被确定，因此它们更可能是真实的。Trinity在粟酒裂殖酵母中鉴定了最大数量的注释的剪切摸式。其它的方法也报道了了大量的错误融合的粟酒裂殖酵母转录本，这些不同的转录本是由一些被称之为单一合并的转录本这样的邻近基因所产生的产物。这导致了其他的方法灵敏度有所降低。

在老鼠中，大多数方法对检测单个注释的内含子有较高的相似的灵敏度，但在检测完整的剪切模式时又有所不同Scripture可以鉴定出最多的注释的剪切模式，Trinity紧随其后。然而，Scripture产生了110000多个不同的剪切模式，比Trinity和其他的方法大约多出10倍，这也表明了Scripture中有许多错误的数据，而在Trinity中则有较好的精度。总体而言，每个至少被其他一种方法支撑的方法预测出相对较少的非注释的剪切模式。

最后，我们检查了映射到每个参考基因组上的不同的contigs的数量，以及每个基因座上重建转录本的覆盖率。这说明了多个转录本代表了基因座的同一个区域。比如，由于选择性剪切导致的等位基因的变异或者在未检测到序列错误的情况下枚举了转录本。在粟酒裂殖酵母中，Trinity将7057个转录本映射到4874个基因上，平均覆盖度为每个基因1.37层，和除了Scripture、trans-AbySS之外的其他方法类似。在小鼠中Trinity的表现和除了trans-AbySS之外的其他方法类似。大量的trans-AbySS转录本覆盖在基因座的相似区域并不能反应不同的剪切模式的数量，只能表明多个相似的转录本序列由各个基因座产生，而不是代表着许多不同的剪切模式。如果仅是AbySS，相比较于trans-AbySS而言，其缺乏高灵敏度，产生的contigs也更少。

粉虱转录组的重新组装。

在没有测序基因组的情况下，基于RNA-seq的从头组装是研究至今大多数生物体转录组组装的唯一可行的选择。比如，尽管高度多样化的昆虫纲包含了集中关键的模型生物，但是其并没有被高质量的基因组序列草案密集地覆盖。另外昆虫的转录组也表现出了复杂的选择性剪切模式。白粉虱B.tabaci就是这样的一个例子，它的基因组并没有被测序，并且由于它们来自于远交群体的个体的混合物，其RNA-seq样品是遗传多态的。

我们将Trinity应用于粉虱的公开的RNA-seq数据集，该数据集由2190万对Illumina的76bp reads组成，并使用常规的非链特异性方法进行测序。Trinity产生了96000个转录本捕获了等位基因变体。其中，至少有80%的相应的同源蛋白质序列中有4323个具有最高BLASTX匹配至2880个独特的Unire90蛋白质序列。由Trinity产生的几近全长转录本的数量要远远大于其他从头组装转录本。

为了评估由Trinity捕获的可变剪切模式的程度，我们对所有从单个图组件中得到的contigs进行对比，并寻找至少有一个可选内部外显子的最小长度为21并且长度是3的倍数的证据。通过此定义，325个组件包含至少两个不同的亚型。