超高精度大规模测序mapping算法FANSe在非模式生物中的应用

曹欣, 张弓*

暨南大学生命与健康工程研究院, 功能蛋白质广东省普通高校重点实验室, 广州 510632

* 联系人, E-mail: zhanggong@jnu.edu.cn

收稿日期: 2017-04-30; 接受日期: 2017-06-25; 网络版发表日期: 2017-07-12

国家高技术研究发展计划青年科学家专题项目(批准号: 2014AA020504)和国家自然科学基金(批准号: 81322028)资助

摘要 大规模测序技术以其高通量、低成本的优势在生物学中得到广泛的应用. 但对于基因组尚未被准确测定的非模式生物, 传统算法受制于较低的精度、稳健性和容错性, 难以有效应对. FANSe系列算法作为目前精度最高、容错性最好的大规模测序快速比对算法, 解决了非模式生物二代测序应用所面临的参考库不准确问题, 可为非模式生物的基因组、转录组、蛋白质组研究提供准确的结果. 本文总结了目前基于FANSe的多种分析方案在非模式动物、植物、微生物以及复杂的共生体系中的应用, 为非模式生物的研究提供参考.

关键词 二代测序, FANSe算法, 非模式生物, 高精度

随着大规模测序技术的蓬勃发展, 人们测定了越来越多的基因组. 到2016年3月为止, NCBI数据库中已测定基因组序列的真核生物有2759种, 原核生物有61484种, 病毒有5541种, 细胞器有7832种. 但仍然有大量的生物属于“非模式物种”, 即基因组序列未被准确测定的生物以及已初步测定基因组但是注释不完整的生物. 在大多数测序数据分析流程中, 测序所得的短的核酸片段(reads)首先需要比对到参考基因组序列上(通常称为mapping), 这一步是整个数据分析的基础, 不精确的比对会导致后续所有计算失准, 得出不可靠的结论^[1,2]. 由于缺乏准确的参考基因组及注释信息, 非模式生物的基因组、转录组、蛋白质组等研究也就比较困难.

对于像人这样的“模式生物”虽然基因组在2003年

就已被测定, 但是这个参考基因组序列只是人这一物种的“代表性个体”的基因组, 却无法作为所有人类的准确参照. 至2017年4月, dbSNP数据库150版中共收录了3.26亿个高可信reference SNPs, 相当于人基因组的10.5%, 这证实不同人种之间存在较大的差异, 而国际上被广泛使用的GRCh37, GRCh38等人类参考基因组序列均为高加索人种(欧美白人), 与中国人有着较大的差异. 对中国人基因组进行高质量的组装的“华夏一号”与目前广泛使用的人类参考基因组版本GRCh38相比, “华夏一号”组装填补了GRCh38近1/3的缺口, 并发现且验证出许多真实存在的在GENCODE中从未有过注释的转录本^[3]. 人的不同个体间差异度尚且这么大, 更何况其他物种. 细菌的同种菌的不同菌株之间的差异可高达5%~20%^[4], 而其差异部分与致病性、抗

引用格式: 曹欣, 张弓. 超高精度大规模测序mapping算法FANSe在非模式生物中的应用. 中国科学: 生命科学, 2017, 47: 702–707.
Cao X, Zhang G. Application of the hyper-accurate mapping algorithm FANSe for next-generation sequencing in non-model organisms (in Chinese). Sci Sin Vitae, 2017, 47: 702–707, doi: [10.1360/N052016-00280](https://doi.org/10.1360/N052016-00280)

生素耐药性高度相关^[5], 因此, 获得准确的参考序列就成为非模式物种测序分析的首要条件.

获得非模式生物的参考序列主要有两种策略, 在有近缘参考基因组可用时, 基于mapping算法进行修正^[6]: 将reads往近缘参考基因组上mapping, 找到差异并修正. 这种策略可完整利用近缘物种中各基因的注释. 但传统的mapping算法效果很不理想, 只能修正1.6%的基因组差异^[7-10], 传统算法往往会将大部分的序列丢失或是产生大量的错配^[11]. 在完全没有近缘参考基因组时, 则只能采用*de novo*拼接的方法拼接出参考基因组^[7]. 这种做法不但需要大量的前期调研, 而且需要多种片段长短文库的构建^[12], 测序数据常常需要高达基因组实际大小的100倍左右^[13], 使得测序成本居高不下. 处理如此海量的数据, 还需要对拼接算法非常了解和强大的计算集群. 尽管耗时耗力, 拼接出来参考基因组也并不能达到预期, 这一方面受制于测序仪的性能, 另一方面拼接算法也存在一定的弊端. 目前主流的拼接算法都基于de Bruijn图^[14-16], 而此类算法对测序中偶然的错误十分敏感, 组装错误率高, 如*Alcanivorax borkumensis* SK2基因组用最好的方法组装也会有5%的错误率, 即每20个碱基中有一个错误^[7], 如果想要对其功能进行研究, 就需要大量的一代测序校正. 非模式生物单一物种的研究尚且这么难, 对于多物种之间的相互作用或是连物种情况都不明确的共生体系就更加困难.

目前被广泛使用的mapping算法如Bowtie, BWA等的数学核心有两部分, 一部分是BWT (Burrows-Wheeler transform), 它是利用重复序列只存储一次, 轮转排序消除位置差异, 从而达到压缩空间、节省时间的目的, 但是丢失了子序列的重复位置的信息. 另外一部分是FM-index (Ferragina-Manzini index), 这是一种全文索引结构, 通常实现为双向无环数据结构, 随着搜索的逐位进行, 搜索空间迅速减小, 因此可大大缩短搜索时间, 但是一旦有错配(测序错误或SNV)就无法回到正确的分支上, 造成错误mapping或丢失. 所以这些传统算法虽然内存消耗小, 速度快, 但是其精度差、稳健性差、容错差^[17,18]. 一般只能容许1%~2%的错误率, 无法应对非模式物种动辄5%以上的差异. 而基于子序列索引的seed-based类算法虽然精度高、稳健性好, 但是运算极慢, 对大数据的处理存在严重弊端^[17]. 因此, 对非模式生物的研究急需一种准确、快速、容错

性好的mapping算法.

张弓等人^[1]开发的FANSe系列算法即是这种思想下开发出来的mapping算法. 它采用的seed-hash的数学核心以保证其精度和稳健性好, 很好地规避传统算法的缺陷; 然后通过计算技术的深度优化达到了优于Bowtie系列的运算速度^[19]. FANSe系列算法相对于传统算法的优势在于: (i) 稳健的高精度, FANSe系列算法设计有概率论保证, 是世界上第一种可以预估错误率的mapping算法, FANSe2 (FANSe算法第二代)的mapping错误率可以稳健的低至 10^{-6} 以下, 实践中基本可以忽略不计. (ii) 高容错, 容错率可高达12%, 而传统算法在变异度超过2%的时候准确度就剧烈下降. (iii) 运算快, FANSe2的速度约为Bowtie2的4倍^[19]. 综上所述FANSe系列的算法特别适合于非模式生物的研究.

目前FANSe算法对于非模式生物的应用分为以下几大策略. 近缘基因组变异度 $\leq 7\%$ 的物种可采用FANSe2高容错一步mapping^[19]; 近缘基因组变异度 $\leq 20\%$ 的物种采用FANSe2迭代修正; 近缘基因组变异度大于20%或者完全没有近缘参考基因组的单一物种采用从头拼接加FANSe2高容错mapping; 完全没有参考序列的多物种共生体系采用拼接加FANSe2多次高容错mapping(图1).

1 FANSe在近缘基因组变异度 $\leq 7\%$ 的非模式生物研究中的应用

虽然水稻(*Oryza sativa*)基因组计划2005年已经完成^[20], 但是在长期的育种过程中水稻品种之间的差异普遍在1%以上^[21], 因此对算法的容错性有较高的要求. Hussain等人^[22]首次对引发水稻幼苗在淹水胁迫下的转录表达谱进行研究. 为了解决测序样品与参考序列物种之间的变异引发的分析错误, 采用了高精度、高容错、高稳健的FANSe算法来进行比对. 使用FANSe2在7%的容错率下(许多研究表明6%~8%的错配适用于大部分研究工作^[2,23,24]), 将4组8个样品的转录组数据与参考转录组库进行比对. 最终得到的比对率都在84%以上, 差异表达的基因可很好地被qRT-PCR验证, 证实了基于FANSe2的基础分析的可靠性.

Li等人^[25]从当地肉菜市场上购买牛眼, 培养并工程化改造角膜内皮细胞, 使其发生上皮-间质转化(epi-

图1 FANSe算法解决非模式生物研究策略

thelial-mesenchymal transition, EMT), 从而提高增殖速率. 为研究该过程中细胞内各通路的变化, 确定其机制, 对其进行了RNA-seq. 牛的参考基因组虽已测定, 但育种过程使得个体差异大, 且注释不完善, 也属于典型的非模式物种个体. 使用FANSe2进行高容错mapping, 得到差异表达基因并能被实验证实, 从而指导进一步的功能验证实验并顺利获得成功. 从其准备样品测序到最终所有表型功能实验全部完成、文章写作投稿, 仅用时3个月左右, 可见FANSe2对非模式物种的基因表达定量十分准确, 符合实际状况, 精准地指导了后续实验的方向, 避免反复试误所浪费的时间、人力和财力.

2 FANSe在近缘基因组变异度 $\leq 20\%$ 的非模式生物研究中的应用

同种细菌不同菌株之间的差异常达到5%~20%, 这样会超过FANSe系列算法12%的容错极限, 因此一次mapping可能并不能达到足够好的效果. 为解决细菌基因组高变异率的问题, 吴晓蕙等人^[2]利用迭代的方法不断修正基因组序列, 成功解决了变异率20%的细菌基因组测定. 该方法采用FANSe算法将环境中分离出一株*Bacillus pumilus*细菌基因组测序结果比对至该种标准菌株*Bacillus pumilus* SAFR-032的基因组上, 允许7%的错配(含插入或缺失), 然后通过Fisher exact test统计检测SNV, 把基因组上的SNV位点进行修正, 同时更新注释位置信息. 每一轮修正完毕都更接近于真实的基因组序列. 经过7轮的修正后, 73.2%

的reads能够比对到基因组上, 这一数值已非常接近HiSeq-2000型测序仪所能达到的76.6%的理想mapping率, 而检测到的SNV数量已非常少, 可以认为基本修正完全, 得到了分离出的这株菌的真实基因组^[26]. 该团队同时还使用了另外两种方法进行比较, 一种是基于Bowtie2与Samtools结合的方法^[27], 一种是Stampy比对并用Samtools检测SNV^[28]. 其验证结果显示, 基于FANSe算法方法共修正182620碱基, 最后的mapping率达到73.2%; Bowtie2与Samtools结合的方法只能修正167个碱基; Stampy和Samtools的方法共修正10000多个SNV. 用Sanger测序对3种方法修正的结果进行验证, 结果显示基于FANSe算法的修正结果与Sanger测序结果完全一致, 所验证的近2000个位点无一假阳性、无一假阴性, 对约20%的变异区段也能做到完美的修正; 而其他两种方法都有大量的假阳性和假阴性. 这说明FANSe算法在高变异率下的准确性与容错能力远强于传统算法. 并使用FANSe修正后的蛋白参考库, 这株菌的蛋白质组质谱鉴定提高了11.5%的蛋白鉴定率和14.2%的肽段鉴定率, 覆盖突变位点的肽段比原始库提高了近5倍. 这说明基于FANSe的迭代修正方法能有效地在基因组、转录组、蛋白质组研究上应对高变异率的细菌.

3 FANSe在高变异近缘基因组或无参考基因组的单一物种研究中的应用

烟草(*Nicotiana tabacum* L.)作为重要的经济作物, 幼苗期的冷胁迫对其生长和产量产生巨大影响. 但是

长期育种导致烟草的变异度非常大, 标准株参考序列库基本不可用, 导致烟草的各种生物学机制研究非常困难. Hu等人^[29]对NC567耐寒株和Taiyan8冷敏感株两种烟草4个时间点(0, 4, 24, 48 h)的转录组比较分析揭示烟草低温应答机制的特异基因型. 由于这两种烟草与已测烟草基因组的差异度比较大, 所以对其转录组进行从头组装. 用Trinity^[14]拼接出转录组作为参考库, 由于拼接结果存在相当的错误率, 所以研究者用FANSe2在7%的容错量下进行序列比对, 8组数据比对率都在95%以上. 紧接着进行进一步的数据分析, 在分析出来基因中挑选10个与转录因子、泛素蛋白酶体系统(ubiquitin-proteasome system, UPS)和生物钟有关的基因进行qRT-PCR验证. 结果表明, 转录组分析数据与qRT-PCR的一致性非常好, 证明FANSe算法能够很好地解决由于测序仪的性能以及拼接算法弊端所导致拼接出来的参考序列具有一定错误率而引发的计算错误.

4 FANSe在复杂共生体系研究中的应用

4.1 FANSe在两个物种体系研究中的应用

尹晗琪研究组^[30]通过对18 h感染周期的宿主斜纹夜蛾细胞转录组组装, 来研究在感染过程中宿主和病毒的基因变化情况. 由于斜纹夜蛾没有参考基因组序列, 并且斜纹夜蛾核型多角体病毒基因能够整合到宿主基因上并进行表达^[31], 所以组装的错误率比单一物种的错误率更高, 为了去除组装出来不可靠的unigenes, 使用FANSe2^[19]在7%的容错率下进行mapping, 将reads小于10的unigenes去除, 保留99180个unigenes作为参考转录组库. 把0, 6, 12, 18 h感染后的测序数据用高精度FANSe2算法在6%的容错条件下进行mapping, 然后进行差异基因计算. 随机抽取2个差异宿主基因和2个病毒基因进行qRT-PCR验证, 均成功, 说明计算非常准确.

另外, 翁庆北研究组^[32]也对甜菜夜蛾核多角体病毒持续感染第8代的细胞P8-Se301-C1和甜菜夜蛾细胞Se301进行转录组测序来研究其杆状病毒潜伏感染和超感染的分子机制. 首先分别对两组测序数据进行了常规的低质量序列去除和拼接. 然后采用高精度算法软件FANSe2在允许7个错配的情况下进行mapping. 最

后找到10个只存于P8-Se301-C1而不在Se301中的核多角体病毒基因, 并且这些基因都通过qRT-PCR确证了P8-Se301-C1确实存在. 进一步说明了FANSe算法应用在拼接策略中的准确性.

4.2 FANSe在多个物种复杂体系研究中的应用

植物根际共生系统是典型的复杂共生系统, 往往包含数十至数百个物种. 以往对于此类共生系统研究有两种方法, 一种是把植物和微生物分开来研究, 但是分开来研究的所获得的基因表达谱与其共生情况不一样. 宏基因组测序能够避免这种问题, 但是由于宏基因组测序数据的高复杂性导致难以分析可能导致错误的结论^[33], 并且这种研究需要知道植物的全基因组序列, 而对于植物和共生菌全基因组都未知的情况无从下手^[34].

罗滨等人^[35]首次提出直接从头拼接宏转录组的策略, 并用这一策略对紫根水葫芦根系宏转录组与普通水葫芦进行比较分析. 这项研究需要攻克许多技术难题, 水葫芦的同科植物的基因组都没有被测定, 其根系共生真菌高达数百种, 共生情况也不明, 这就导致只能将整个共生系统看成一个大的整体, 做宏转录组的从头拼接, 这对测序样品的制备、测序质量、拼接软件、参数设置有很高的要求, 因为很容易把不同物种的序列拼接到了一起^[33]. 极为复杂的宏转录组体系的拼接势必产生大量的错配, 所以mapping算法就必须有较高的容错能力; 另一方面, 植物和真菌的mRNA原本就有一定的同源性, 那么mapping算法又必须能区分来源不同物种的基因. 高容错和高区分度本身是相互矛盾的两个要求, 但FANSe2以其高精度和高容错性可以同时满足. 最终紫根水葫芦的比对率高达86.5%, 普通水葫芦比对率86.6%. 而传统算法在非模式物种组装比对效果最好的Bowtie2^[36]也只能比对40%左右, 可见传统算法无法应对如此复杂的共生系统数据. 然后对其差异表达分析进行宏转录本的物种分布估计, 并用PCR-DGGE和18S ITS测序进行物种确证, 找出物种变化最大的共生菌是镰刀菌(*Fusarium*), 然后将reads在用高容错FANSe2 mapping到镰刀菌的参考序列上并进行真菌的差异表达分析, 最终找到重金属结合功能基因. 用这种策略, 研究者得以揭示紫根水葫芦具备强力吸附重金属的功能基因及植物-真菌的相互作用.

5 小结与展望

高精度mapping算法FANSe系列是目前实验已验证假阳性、假阴性最少的二代测序快速mapping算法。FANSe以其高精度、高稳健性以及高容错的特点能很好应对非模式生物必然出现的高变异度问题, 远远优于传统算法, 使以往无从下手的研究变得容易; 并可以使用mapping的方法取代一部分传统上只能依赖从头

拼接的方法, 从而大大提高了准确度和可靠性。由于基于FANSe的分析结果的高度实验可验证性, FANSe以其精准的计算为基因组学、转录组学、蛋白质组学研究提供可靠的结果, 很好地解决了当前大规模测序结果不能被验证、重复的危机^[37]。以上列举的不同应用场景, 展示了FANSe系列算法应对各种非模式生物研究时的基本策略和极强的适应性, 为大规模测序应用到更广泛的领域提供强有力的计算保障。

参考文献

- 1 Zhang G, Fedyunin I, Kirchner S, et al. FANSe: an accurate algorithm for quantitative mapping of large scale sequencing reads. *Nucleic Acids Res*, 2012, 40: e83–e83
- 2 Wu X, Xu L, Gu W, et al. Iterative genome correction largely improves proteomic analysis of nonmodel organisms. *J Proteome Res*, 2014, 13: 2724–2734
- 3 Shi L, Guo Y, Dong C, et al. Long-read sequencing and *de novo* assembly of a Chinese genome. *Nat Commun*, 2016, 7: 12065
- 4 Baba T, Bae T, Schneewind O, et al. Genome sequence of *Staphylococcus aureus* strain newman and comparative analysis of staphylococcal genomes: polymorphism and evolution of two major pathogenicity islands. *J Bacteriol*, 2008, 190: 300–310
- 5 Martinez J L, Baquero F. Mutation frequencies and antibiotic resistance. *Antimicrobial Agents Chemother*, 2000, 44: 1771–1777
- 6 Otto T D, Sanders M, Berriman M, et al. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics*, 2010, 26: 1704–1707
- 7 Kisand V, Lettieri T. Genome sequencing of bacteria: sequencing, *de novo* assembly and rapid analysis using open source tools. *BMC Genomics*, 2013, 14: 1
- 8 Storer C G, Pascal C E, Roberts S B, et al. Rank and order: evaluating the performance of SNPs for individual assignment in a non-model organism. *PLoS ONE*, 2012, 7: e49018
- 9 Carlsson J, Gauthier D T, Carlsson J E L, et al. Rapid, economical single-nucleotide polymorphism and microsatellite discovery based on *de novo* assembly of a reduced representation genome in a non-model organism: a case study of Atlantic cod *Gadus morhua*. *J Fish Biol*, 2013, 82: 944–958
- 10 Yang X, Chockalingam S P, Aluru S. A survey of error-correction methods for next-generation sequencing. *Briefings Bioinform*, 2013, 14: 56–66
- 11 Glenn T C. Field guide to next-generation DNA sequencers. *Mol Ecol Resour*, 2011, 11: 759–769
- 12 Li R, Li Y, Fang X, et al. SNP detection for massively parallel whole-genome resequencing. *Genome Res*, 2009, 19: 1124–1132
- 13 田洋, 曾严, 张静, 等. 辣木(*Moringa oleifera* Lam.)的高质量参考基因组. *中国科学: 生命科学*, 2015, 45: 488–497
- 14 Haas B J, Papanicolaou A, Yassour M, et al. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*, 2013, 8: 1494–1512
- 15 Luo R, Liu B, Xie Y, et al. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaSci*, 2012, 1: 18
- 16 Zerbino D R, Birney E. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res*, 2008, 18: 821–829
- 17 Trapnell C, Salzberg S L. How to map billions of short reads onto genomes. *Nat Biotechnol*, 2009, 27: 455–457
- 18 Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings Bioinform*, 2010, 11: 473–483
- 19 Xiao C L, Mai Z B, Lian X L, et al. FANSe2: a robust and cost-efficient alignment tool for quantitative next-generation sequencing applications. *PLoS ONE*, 2014, 9: e94250
- 20 Project I R G S. The map-based sequence of the rice genome. *Nature*, 2005, 436: 793–800
- 21 Huang X, Kurata N, Wei X, et al. A map of rice genome variation reveals the origin of cultivated rice. *Nature*, 2012, 490: 497–501
- 22 Hussain S, Yin H, Peng S, et al. Comparative transcriptional profiling of primed and non-primed rice seedlings under submergence stress. *Front Plant Sci*, 2016, 7: 1125
- 23 Zhong J, Cui Y, Guo J, et al. Resolving chromosome-centric human proteome with translating mRNA analysis: a strategic demonstration. *J Proteome Res*, 2013, 13: 50–59
- 24 Chang C, Li L, Zhang C, et al. Systematic analyses of the transcriptome, translome, and proteome provide a global view and potential strategy

- for the C-HPP. *J Proteome Res*, 2013, 13: 38–49
- 25 Li S, Han Y, Lei H, et al. *In vitro* biomimetic platforms featuring a perfusion system and 3D spheroid culture promote the construction of tissue-engineered corneal endothelial layers. *Sci Rep*, 2017, 7: 777
 - 26 Liu L, Li Y, Li S, et al. Comparison of next-generation sequencing systems. *Biomed Res Int*, 2012, 2012: 251364
 - 27 Wang X, Slebos R J, Wang D, et al. Protein identification using customized protein sequence databases derived from RNA-Seq data. *J Proteome Res*, 2011, 11: 1009–1017
 - 28 Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res*, 2011, 21: 936–939
 - 29 Hu R, Zhu X, Xiang S, et al. Comparative transcriptome analysis revealed the genotype specific cold response mechanism in tobacco. *Biochem Biophys Res Commun*, 2016, 469: 535–541
 - 30 Yu Q, Xiong Y, Liu J, et al. Transcriptome analysis of the SL221 cells at the early stage during *Spodoptera litura* nucleopolyhedrovirus infection. *PLoS ONE*, 2016, 11: e0147873
 - 31 Blissard G W. Baculovirus-insect cell interactions. In: *Insect Cell Culture: Fundamental and Applied Aspects*. Berlin: Springer, 1996. 73–93
 - 32 Fang Z, Shao J, Weng Q. *De novo* transcriptome analysis of *Spodoptera exigua* multiple nucleopolyhedrovirus (SeMNPV) genes in latently infected Se301 cells. *Virol Sin*, 2016, 31: 425–436
 - 33 Sharon I, Banfield J F. Genomes from metagenomics. *Science*, 2013, 342: 1057–1058
 - 34 Schatz M C, Witkowski J, McCombie W R. Current challenges in *de novo* plant genome sequencing and assembly. *Genome Biol*, 2012, 13: 1
 - 35 Luo B, Gu W, Zhong J, et al. Revealing crosstalk of plant and fungi in the symbiotic roots of sewage-cleaning *Eichhornia crassipes* using direct *de novo* metatranscriptomic analysis. *Sci Rep*, 2015, 5: 15407
 - 36 Singhal S. *De novo* transcriptomic analyses for non-model organisms: an evaluation of methods across a multi-species data set. *Mol Ecol Resour*, 2013, 13: 403–416
 - 37 Nekrutenko A, Taylor J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nat Rev Genet*, 2012, 13: 667–672

Application of the hyper-accurate mapping algorithm FANSe for next-generation sequencing in non-model organisms

CAO Xin & ZHANG Gong

Key Laboratory of Functional Protein Research of Guangdong Higher Education Institutes, Institute of Life and Health Engineering,
College of Life Science and Technology, Jinan University, Guangzhou 510632, China

Next-generation sequencing (NGS) has been widely used in biology studies for its high throughput and low cost. However, for the non-model organisms, whose genome has not been accurately sequenced, the traditional mapping algorithms cannot process them efficiently due to the low accuracy, robustness and error tolerance. FANSe series algorithms is the most accurate and error tolerant mapping algorithm for NGS. It solves the problem of the inaccurate reference database, which is the common problem in the NGS for non-model organisms. Therefore, FANSe provides accurate results for genome, transcriptome and proteome studies of non-model organisms. This review summarizes multiple analysis strategy using FANSe in non-model animals, plants, microorganisms and complex symbiosis system as examples for non-model organisms studies.

next-generation sequencing, FANSe algorithm, non-model organism, hyper-accurate

doi: [10.1360/N052016-00280](https://doi.org/10.1360/N052016-00280)