DOI: 10.13376/j.cbls/2017032
文章编号：1004-0374(2017)03-0230-07

李川昀，博士，北京大学分子医学研究所研究员，博士生导师。2009 年获北京大学理学博士学位，曾在美国 NIH 从事复杂疾病研究，2011 年在北京大学建立实验室，运用猴作为人类近缘模式动物的优势，在全基因组尺度探索人类演化和复杂疾病的分子机制。主持国家优秀青年科学基金、“万人计划”青年拔尖人才等科研项目，曾获“贝时璋青年生物物理学家奖”等奖项，所指导的博士生获 Ray Wu Prize、北京大学优秀博士学位论文等。

在恒河猴基因组学框架下研究人类演化与调控

钟晓明，申　晴，彭继光，李玉梅，李川昀*
(北京大学分子医学研究所，北京 100871)

摘　要：恒河猴作为人类近缘的模式生物，在基础与转化医学研究中具有独特优势，但其应用受到功能基因组学数据匮乏、基因结构混乱、研究平台缺乏等限制。近年来，深度测序技术的发展为突破这些技术瓶颈提供了机遇。现综述在深度测序技术支撑下，以恒河猴为背景开展的基因组学与分子演化研究，以期抛砖引玉，推动非人灵长类领域的研究进程。

关键词：恒河猴；深度测序；人类演化；猴基因组学
中图分类号：Q343.1；Q789；Q959.848　　**文献标志码：**A

Understanding human biology in the genomic context of rhesus macaque

ZHONG Xiao-Ming, SHEN Qing, PENG Ji-Guang, LI Yu-Mei, LI Chuan-Yun*
(Institute of Molecular Medicine, Peking University, Beijing 100871, China)

Abstract: With human-comparable genome sequence and the advantages as model animal, rhesus macaque poses a unique model in molecular and translational study of human diseases. Despite these unique advantages, several unresolved issues have limited the current use of the model--inadequate functional genomics annotations, error-prone gene models, and lack of a platform for visualizing and assessing high-throughput data. With the development of the deep sequencing technology, some of these key issues have been addressed these years. Here we summarized the monkey genomics and molecular evolution studies we performed in the deep sequencing era.

Key words: rhesus macaque; deep sequencing; human evolution; monkey genomics

收稿日期：2016-04-30
基金项目：国家自然科学基金项目(31522032, 31471240, 31221002, 31171269)
*通信作者：E-mail: chuanyunli@pku.edu.cn

恒河猴 (rhesus macaque) 作为非人灵长类模式动物，兼具模式动物环境因素可控、取材检测方便以及基因组和生理病理接近人类的优势，为研究人类演化与复杂疾病提供了独特视角，极大地推动了分子演化、神经科学、行为学、病理学等多个学科的发展 [1-4]。然而，恒河猴资源稀缺，且存在功能基因组学数据匮乏、基因结构注释混乱、研究平台不成熟等技术瓶颈，限制了其应用 [5]。近年来，深度测序技术的发展为解决这些问题提供了契机。本文将综述在这一时代背景下，以恒河猴为研究视角，探究人类演化与调控的部分工作，以期抛砖引玉，推动非人灵长类领域的研究。

1　恒河猴研究的优势与瓶颈

作为人类近缘的模式生物，恒河猴与人类的分歧时间约为 2 500 万年，其基因组与人的相似性约为 93%[6]，在研究人类演化与复杂疾病方面具有独特优势。一方面，相比于以人类为研究对象的研究工作，恒河猴作为模式生物取材方便，且环境因素可控，机制研究容易深入。例如，可以获取同一恒河猴个体的多个组织，避免个体差异对研究的影响。同时，其可控的饲养条件有利于减少环境因素对研究的干扰。另一方面，相较于小鼠等模式生物，恒河猴与人类的亲缘关系更近，相应的研究成果更容易转化到人类。此外，恒河猴主要分布在我国南部，使得在我国开展恒河猴研究具有独特的资源优势。

基于这些优势，恒河猴已被广泛应用于药物研发、行为学等细胞、个体层面的研究中。例如 Ohkawa 等 [7] 利用恒河猴为实验动物，系统研究了 SIV (simian immunodeficiency virus) 疫苗引起的免疫反应，辅助了疫苗研发。然而，相较于细胞、个体层面的应用，恒河猴在分子层面的研究则要少很多。造成这种现象的原因，主要包括恒河猴匮乏的功能基因组学数据、混乱的基因结构注释，以及不成熟的研究平台。早期分子生物学数据主要依赖于费时费力的第一代测序方法，极大限制了数据产出，如恒河猴的表达序列标签数据 (EST) 还不足人类的 1%[5]。数据的稀缺则进一步增加了对恒河猴基因组进行拼装和基因结构注释的难度。一方面，由于用于拼接恒河猴基因组的数据覆盖度不高 (仅为 5× 左右 [6])，造成恒河猴基因组序列不准确、存在漏洞、错误拼接等问题，进而产生了诸如移码突变、外显子缺失等基因注释的错误 [8]。另一方面，由于针对恒河猴的转录组数据较少，猴 90% 以上的基因结构源于预测，质量差，极大地阻碍了对恒河猴基因的功能研究。此外，这些注释信息主要分散在文献和多个二次数据库中，缺乏对数据的有效整合与利用。长期以来，MGI [9]、FlyBase [10]、WormBase [11] 对小鼠、果蝇、线虫等领域的研究起到巨大的推动作用，如果能在非人灵长类研究领域填补这一空白，将为开展特色的猴基因组医学研究铺平道路。

2　解决恒河猴基因组学研究存在的三个技术瓶颈

如前所述，恒河猴作为非人灵长类模式动物，为研究人类演化与复杂疾病提供了独特视角。但恒河猴研究存在功能基因组学数据匮乏、基因结构注释混乱和研究平台不成熟等技术瓶颈，限制了其应用。近年来兴起的深度测序技术为解决这些技术瓶颈提供了契机。

2.1　解决猴功能基因组学数据匮乏问题

深度测序技术以其高准确性、高灵敏度、高通量、低成本的优势，为解决上述技术瓶颈问题提供了基础 [12-13]。笔者首先建立了包含 120 个个体、56 种组织的恒河猴组织样本库，并开发了完善的生物信息学分析与评估流程，对 24 只恒河猴进行了全基因组测序，对恒河猴多个体、多组织开展了系统的全转录组研究，并进行了多个调控层次的组学研究，总测序片段数达 1 000 亿条，对恒河猴基因组和转录组的覆盖率分别达到 99% 和 98%。与 2011 年笔者进入这个领域时相比，目前对恒河猴的功能注释数量有了大幅的提升，极大地丰富了对恒河猴基因组的认识，也为解决恒河猴基因结构注释混乱等瓶颈问题提供了基础。2011 年以来，领域内进一步对更多调控层次的数据进行了研究，如转录起始位点和蛋白质 - 核酸相互作用等 [14-17]，进一步完善了对恒河猴基因组、转录组复杂性的认识，为研究人类演化与复杂疾病机制提供了全新的切入点。

深度测序技术产生的大量数据有效地解决了恒河猴数据匮乏的问题。然而，这些数据在原始样本质量、处理条件、测序平台上存在的差异会对后续的分析造成影响。因而，开发一套标准化的处理流程，对这些数据进行系统的分析、评估和整合是非常必要的。笔者根据各类深度测序研究的性质，制定了相应的评估体系 [18]。以转录组测序数据为例，一套好的测序数据需要具备以下特性：(1) 较高的 RNA 样本完整性；(2) 较高的碱基测序质量和较长的测序片段，以保证测序片段回贴的准确性；(3)

测序片段保留转录本链特异性；(4) 转录本外显子区的测序片段覆盖密度要远高于内含子区和基因间区；(5) 测序片段在转录本上的覆盖度是均一的；(6) 绝大部分的转录区测序片段覆盖度较高。根据上述条件，笔者从 9 个方面对每一套 RNA-Seq 数据进行评估，最终得到代表每套数据质量的标准化分数。当研究对数据质量要求很高时，可根据上述得到的质量分数，过滤掉低质量数据。此外，打分系统中每一分项的分数也有重要的参考意义，如针对双向转录的研究，应选择分链项得分较高的数据进行研究。类似地，笔者开发了针对全基因组重测序、外显子组测序、CLIP-Seq、ChIA-PET、Poly (A)-Seq、Small RNA-Seq、ChIP-Seq 等测序数据的评估体系，对 1 600 多套深度测序数据进行了系统的分析、评估和标准化 [18]，为系统开展恒河猴基因组学研究建立了数据标准。

2.2 实现对恒河猴基因结构的精确修正

由于恒河猴转录组研究较少，猴 90% 以上的基因结构源于预测，质量差，是该领域研究的主要技术瓶颈之一。由于 RNA-Seq 的测序片段来源于转录组，一方面，测序片段在基因组上的位置标识了转录活性区域，其密度代表了该区域的转录水平，可用于对外显子位置的粗略定义；另一方面，跨越多个外显子的测序片段，可用于对内含子 - 外显子边界进行精确定义。利用这一特征，笔者开发了相应的基因结构修正算法，运用 60 亿条 RNA-Seq 自产测序数据，精确定义了猴全基因组两万多个基因的精细结构 [5]，发现之前该领域对高达 28.7% 的基因结构注释存在错误，包括错误的内含子 - 外显子边界、错误的非翻译区 (UTR) 边界，以及丢失的外显子和转录本。进一步，通过比较修正前后内含子区域与外显子区域在测序片段密度分布、跨物种保守性分值以及转录本特定区域的序列特征分布 (如剪切位点信号、Poly(A) 信号、转录起始位点信号等)，确认修正后的基因结构是准确的 [5,18]。

由于上述用于基因结构修正的数据来源于同一个恒河猴个体，为了确保上述发现的错误并非反映了个体间的差异，笔者进一步对 97 套恒河猴 RNA-Seq 数据进行了整合与重分析，涵盖了来自 38 个个体、18 种组织类型的恒河猴样本，总数据量是第一次修正的 9.7 倍。在此基础上，进一步对恒河猴基因结构进行了精细修正和评估，验证了之前绝大部分的修正结果 [18]，仅有少量新错误被发现 (被修正的比例从 28.7% 提高到 30.2%)。总之，经过上述两次修正，恒河猴基因结构注释混乱的问题已得到妥善解决。

2.3 恒河猴“一站式”基因组知识库RhesusBase

长期以来，FlyBase (http://flybase.org)、WormBase (http://www.wormbase.org)、MGI (http://www.informatics.jax.org) 等著名的模式动物数据库通过对相应物种注释信息的系统整合，对果蝇、线虫、小鼠等领域的研究起到巨大的推动作用。而非人灵长类研究领域的数据信息则主要分散在文献和多个二次数据库中，缺乏对数据的有效整合与利用。随着深度测序技术的产生与发展，这一问题显得尤为突出：一方面，不同的功能基因组学数据源自不同的实验平台和分析流程，缺乏标准化，严重影响着对这些数据的整合与解读；另一方面，对这些组学数据的分析和解读需要大量的计算资源，以及专业的生物信息学技术支持，导致生物学家很难使用这些数据用于假设驱动的机制研究。

笔者采用上述纠错修正后的精细的恒河猴基因框架图，在自产数据基础上，进一步整合并重新分析了 1 667 套组学数据和 65 个在线数据库，构建了一个集基因结构、表达、调控、遗传变异、疾病、功能及药物开发等信息于一体的恒河猴“一站式”基因组知识库 RhesusBase (http://www.rhesusbase.org)，总功能注释多达 76 亿条。近期，又陆续开发了分子演化界面 (molecular evolution gateway)、转化医学界面 (translational medicine gateway) 和群体遗传学界面 (population genetics gateway)，建立了基因页面、基因组浏览器、iPhone 手机 APP 等多个访问界面，以方便用户对 RhesusBase 基因组大数据的检索与分析。

仅以基因表达数据为例，针对用户感兴趣的基因，可以直接从 RhesusBase 基因页面中获得该基因在 97 个恒河猴样本、105 个人类样本和 58 个小鼠样本中的表达情况，这些表达数据均基于对 RNA-Seq 原始数据在同一分析平台下细致的分析、评估与标准化，可直接用于物种、组织、个体间的比较分析。此外，用户还可以获取由表达芯片技术和原位杂交技术 (*in situ* hybridization) 鉴定得到的多组织、多发育时期的表达谱数据，为进一步的功能研究提供线索。尤其重要的是，在 RhesusBase 基因组浏览器中，可直接展示原始的测序数据，辅助用户确证表达、转录本结构等信息，并识别新的可变剪切形式。而 RhesusBase 整合的其他信息，如跨物种保守性 [5]、转录因子结合位点 [5]、miRNA 调

控位点[5]、表观遗传学调控信息[18]、群体遗传学水平的自然选择信号[19]等，也为理解该基因的功能提供了新的切入点。

3　在恒河猴基因组学框架下研究RNA编辑调控

如前所述，通过解决恒河猴基因组学研究存在的3个技术瓶颈，尤其是通过建立恒河猴基因组知识库RhesusBase，笔者实现了在配置有丰富注释信息的基因组框架下深入理解基因功能。事实上，这一整合有转录因子结合位点、miRNA调控位点、染色体相互作用、可变剪切等多个调控层次信息的综合基因组框架，促进了对基因如何发挥功能的研究，也为研究新调控层次的功能提供了广阔的视角。接下来将以RNA编辑(RNA editing)调控为例，简要介绍这一应用。

RNA编辑可引起基因组与其编码的转录组在特定位点的序列差异，在个体发育、复杂疾病调控中发挥重要作用[20]。早在1986年，Benne等[21]就报道了RNA编辑的现象，此后由于技术限制，相关研究进展缓慢。深度测序技术的发展推动了该领域的发展，使得能够在全基因组的尺度上发现RNA编辑事件。然而，如何准确鉴定RNA编辑位点，仍然是该领域面临的一个技术挑战。已有多篇评论，如2011年发表在*Science*杂志的人类编辑组集合[22]，表明90%以上是由技术误差导致的假阳性[23-25]。在此背景下，人们只能把不同研究工作呈现出的编辑组异质性笼统地归因于调控“复杂性”。此外，在灵长类动物中，腺嘌呤(A)脱氨基形成次黄嘌呤(I)(A-to-I RNA editing)被认为是这类调控的主要形式。造成这种现象的原因是腺嘌呤脱氨基的催化过程依赖于RNA腺苷脱氨酶(adenosine deaminases acting on RNA, ADARs)特异性地识别初始转录本折叠形成的双链二级结构(dsRNA)[26]，而在灵长类中dsRNA的形成是由广泛分布于基因组内的*Alu*重复序列所介导(人类基因组中约有10%的区域是由*Alu*元件组成[27])。由于小鼠等非灵长类模式生物缺少*Alu*重复序列，近年的研究主要集中在人类细胞系，故机制研究难以深入，在RNA编辑生物学意义的探讨方面存在较大争议。

笔者运用恒河猴作为模式动物和人类近缘物种的双重优势，对来自同一个恒河猴个体的多个组织样本进行了基因组和转录组深度测序、质谱和低通量验证，基于RhesusBase对恒河猴基因结构的精确修正和丰富的功能注释信息，解决了RNA编辑鉴定中的多个技术难点，成功获得了包含三万多个RNA编辑位点的恒河猴全编辑组，涵盖猴多组织、多个体、多发育时期，实验验证率达到97%，呈现出高度的准确性、完整性和定量精度[28]。同时，运用恒河猴多组织、多个体数据的优势，通过比较不同组织的RNA编辑水平与ADARs表达的关系，以及同一组织内不同编辑位点RNA编辑水平与ADARs结合序列的相关性，清晰地阐述了RNA编辑其实在很大程度上受控于*ADAR*基因介导的时空调控，澄清了该领域当前对RNA编辑调控“复杂性”的含糊不当的认识[28]。

RNA编辑研究的另一个难点是其整体功能性。目前虽有个案研究揭示，RNA编辑可以通过改变蛋白质的氨基酸组成从而在复杂疾病中发挥作用，但这些调控多数(>99%)位于不编码蛋白质的基因组区域。这些由高通量测序得到的数以万计的编辑事件是否在整体上具有生物功能，是该领域悬而未解的难题。运用恒河猴作为人类近缘物种的优势，通过跨物种比较，笔者发现自然选择在维持RNA编辑组过程中发挥重要作用，为阐明RNA编辑的整体功能性提供了更具说服力的分子演化证据[28]。

为了探究它们具体通过怎样的调控途径来实现其功能，笔者进一步针对猴多组织、多个体、多调控层次开展了系统的组学研究。运用RhesusBase的平台优势，通过比较RNA编辑与其他调控层次在时间和空间上的相关性，笔者首次发现一类由发生RNA编辑的长转录本前体经剪切而形成的piRNA分子，并将其命名为epiRNA (editing-bearing PIWI-interacting RNA)[29]。研究进一步发现，RNA编辑对piRNA生成具有显著的促进作用，而这些发生在epiRNA上的RNA编辑事件正经历着显著的负向自然选择作用，提示在人类、恒河猴等灵长类动物中，与piRNA调控的互作可能是RNA编辑发挥生物学功能的主要途径之一[29]。该项研究在灵长类中首次建立了RNA编辑与piRNA调控的功能联系，为针对这两类调控的机制研究提供了新视角[29]。

4　在恒河猴基因组学框架下研究人类演化

人类与其他动物在生理病理方面存在巨大差别，人之所以为人的分子演化基础是什么，这个问题是人类演化研究领域备受关注的问题之一，也是*Science*杂志总结的重要科学问题之一[30-31]。恒河猴作为非人灵长类模式动物，是研究这一问题的完美模型。一方面，作为人类近缘物种，恒河猴基因组

与人类差别较小，而相较于黑猩猩等与人亲缘关系更近的物种，人、猴基因组间较大的差异有利于研究者在高度相似的基因组背景下进行高精度的比较基因组学研究，是研究人类特异性基因和调控的完美对照；另一方面，作为模式动物，恒河猴还具有环境因素可控、取材方便的优势，机制研究可以深入。在成功解决以猴为模型开展基因组学研究的三个主要技术瓶颈后，笔者开始在恒河猴功能基因组学注释框架下，从新基因和新调控起源的角度，来探究人之所以为人的分子演化基础。

首先，人类特有基因的产生被认为是导致人类特异性状的一个重要原因[32-34]，其起源模式、演化特征与功能研究越来越受到人们关注。在此类研究中，利用近缘物种作为外类群对基因的产生时间进行精确定义，是最为关键的一环。而这对人类近缘物种的功能基因组学注释精度要求很高。利用RhesusBase对恒河猴完善的基因组注释，结合多物种的比较基因组学研究，笔者准确鉴定得到了43例以从头模式起源的人类特有的蛋白质编码基因；进一步研究发现，这些新蛋白质在黑猩猩和恒河猴的同源区域多数(85%)以长非编码RNA (lncRNA)形式存在。有趣的是，它们已具有与人类同源基因相似的转录结构和基因表达模式，提示这些人类特有的蛋白质可能起源自具有精细表达和剪切特征的lncRNA。那么，这些人类特异的新蛋白究竟是功能性的，还是仅仅代表一些翻译层面的噪音？从群体遗传学的角度，通过对目标区域进行捕获测序，笔者获得了这些区域在82个恒河猴个体中的遗传多态性图谱。在此基础上，通过整合67个人的公共全基因数据，并辅以分子生物学实验验证，发现这些以从头模式起源的人类特有新基因，在人类群体中其外显子区的群体突变率(θ_w)和核苷酸多样性(π)均小于内含子区，其蛋白质编码区非同义突变水平明显低于同义突变，且倾向于低频分布[35]。而在对应的猴群中，这些分布均没有显著性差别。这一自然选择信号仅在人群中发现，提示这些新基因编码的蛋白质可能已经发挥了人类特有的生物功能[35]。据此，笔者提出“蛋白质的lncRNA起源假说”，认为lncRNA是新蛋白诞生的温床，并明确了由lncRNA逐步产生功能蛋白的详细过程。这一假说不仅为研究人类特异性性状提供了新基础，也为认识lncRNA的功能提供了新角度，得到了国际同行的广泛认可[36-39]。

全新基因的引入可以触发新性状，而发生在已有基因上的新调控也有可能是新性状的源泉[40]。例如，通过对人、黑猩猩和恒河猴不同性别肝脏组织的RNA-Seq数据进行分析，Blekhman等[41]发现人类特有的剪切事件可能与形态发生、组织解剖特征有关，提示可变剪切在人类演化历程中发挥了重要作用。人类特异的表观遗传学修饰[42]、增强子和启动子[43]同样可能导致新性状的产生。基于自产数据，笔者开发了系统的生物信息学分析流程，实现了在全基因组水平对猴多个调控层次的精确测定，包括基因突变、基因表达、基因可变剪切等。这些恒河猴调控数据的积累为鉴定人类特有调控事件提供了有效对照。例如，笔者对猴编辑组的总覆盖度已经超过人类编辑组覆盖度的10倍以上，这样，结合黑猩猩等其他外群数据，就可以准确地鉴定人类特有调控事件。运用比较基因组学方法，笔者共发现了9 295例人类特有的调控事件，为认识人之所以为人的分子演化基础提供了新的切入点[18]。

除了新基因和新调控的产生，保守基因和调控在人类特异的丢失等因素也可能是人类特异性状的分子基础。恒河猴功能基因组学数据的完善也为深入研究这些分子演化过程提供了重要资源。

5 结语

恒河猴作为非人灵长类模式动物，被广泛用于病理学、行为学等研究中。然而，恒河猴存在功能基因组学数据匮乏、基因结构注释混乱、研究平台不成熟等技术瓶颈，限制了这一特色模型在基础与转化医学研究中的应用。在过去的研究中，笔者将前沿的深度测序技术引入到非人灵长类研究领域，深入研究了恒河猴的基因组和转录组，基本解决了以猴为模型开展研究的三个主要技术瓶颈。然而，恒河猴基因组学领域仍有一些技术问题亟待解决，如虽然目前恒河猴基因组的测序覆盖度已达到了47.4×，但由于测序读长较短，基因组质量仍然不高；另一方面，虽然第二代测序技术所产生的测序读段能够实现对转录本局部结构的精细定义，但过短的序列读长限制了视野，导致无法获知转录本的整体结构，如多个可变剪切事件在转录本层次是如何搭配的。近年来，以PacBio技术为代表的第三代测序技术得以快速发展，这些新测序技术的单分子、长读长特性有望为解决上述技术问题提供契机。

在解决上述技术瓶颈的基础上，笔者进一步以猴为视角，从新基因和新调控角度探究了人类特异性状的分子基础，并对RNA编辑等新调控进行了

整体组学层面的研究。这些工作运用了恒河猴作为人类近缘模式动物的优势，为探究人类演化与调控提供了大量的新线索。在今后的工作中，这些线索有待进一步深入研究，以促进对人之所以为人等基础科学问题的认识。此外，恒河猴基础调控数据的积累也有望促进以恒河猴为模型的转化医学研究，发现新的复杂疾病标记物与治疗靶点。

[参 考 文 献]

[1] Bliss-Moreau E, Moadab G, Bauman MD, et al. The impact of early amygdala damage on juvenile rhesus macaque social behavior. J Cogn Neurosci, 2013, 25: 2124-40

[2] Melnick DJ. The genetic-structure of a primate species - rhesus macaques and other cercopithecine monkeys. Int J Primatol, 1988, 9: 195-231

[3] Richard AF, Goldstein SJ, Dewar RE. Weed macaques - the evolutionary implications of macaque feeding ecology. Int J Primatol, 1989, 10: 569-97

[4] Hayreh SS, Jonas JB. Appearance of the optic disk and retinal nerve fiber layer in atherosclerosis and arterial hypertension: an experimental study in rhesus monkeys. Am J Ophthalmol, 2000, 130: 91-6

[5] Zhang SJ, Liu CJ, Shi M, et al. RhesusBase: a knowledgebase for the monkey research community. Nucleic Acids Res, 2013, 41: D892-905

[6] Gibbs RA, Rogers J, Katze MG, et al. Evolutionary and biomedical insights from the rhesus macaque genome. Science, 2007, 316: 222-34

[7] Ohkawa S, Wilson LA, Larosa G, et al. Immune responses induced by prototype vaccines for AIDS in rhesus monkeys. AIDS Res Hum Retrov, 1994, 10: 27-38

[8] Zhang X, Goodsell J, Norgren RB Jr. Limitations of the rhesus macaque draft genome assembly and annotation. BMC Genomics, 2012, 13: 206

[9] Eppig JT, Blake JA, Bult CJ, et al. The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. Nucleic Acids Res, 2015, 43: D726-36

[10] Attrill H, Falls K, Goodman JL, et al. FlyBase: establishing a gene group resource for *Drosophila melanogaster*. Nucleic Acids Res, 2016, 44: D786-92

[11] Howe KL, Bolt BJ, Cain S, et al. WormBase 2016: expanding to enable helminth genomic research. Nucleic Acids Res, 2016, 44: D774-80

[12] Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. Nat Rev Genet, 2011, 12: 87-98

[13] Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet, 2010, 11: 31-46

[14] Yan G, Zhang G, Fang X, et al. Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. Nat Biotechnol, 2011, 29: 1019-23

[15] Fang X, Zhang Y, Zhang R, et al. Genome sequence and global sequence variation map with 5.5 million SNPs in Chinese rhesus macaque. Genome Biol, 2011, 12: R63

[16] Cain CE, Blekhman R, Marioni JC, et al. Gene expression differences among primates are associated with changes in a histone epigenetic modification. Genetics, 2011, 187: 1225-34

[17] Liu Y, Han D, Han Y, et al. *Ab initio* identification of transcription start sites in the rhesus macaque genome by histone modification and RNA-Seq. Nucleic Acids Res, 2011, 39: 1408-18

[18] Zhang SJ, Liu CJ, Yu P, et al. Evolutionary interrogation of human biology in well-annotated genomic framework of rhesus macaque. Mol Biol Evol, 2014, 31: 1309-24

[19] Zhong X, Peng J, Shen QS, et al. RhesusBase PopGateway: genome-wide population genetics atlas in rhesus macaque. Mol Biol Evol, 2016, 33:1370-5

[20] Maas S. Gene regulation through RNA editing. Discov Med, 2010, 10: 379-86

[21] Benne R, Van den Burg J, Brakenhoff JP, et al. Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. Cell, 1986, 46: 819-26

[22] Li M, Wang IX, Li Y, et al. Widespread RNA and DNA sequence differences in the human transcriptome. Science, 2011, 333: 53-8

[23] Kleinman CL, Majewski J. Comment on “Widespread RNA and DNA sequence differences in the human transcriptome”. Science, 2012, 335: 1302; author reply

[24] Lin W, Piskol R, Tan MH, et al. Comment on “Widespread RNA and DNA sequence differences in the human transcriptome”. Science, 2012, 335: 1302; author reply

[25] Pickrell JK, Gilad Y, Pritchard JK. Comment on “Widespread RNA and DNA sequence differences in the human transcriptome”. Science, 2012, 335: 1302; author reply

[26] Higuchi M, Single FN, Kohler M, et al. RNA editing of AMPA receptor subunit GLuR-B: a base-paired intron-exon structure determines position and efficiency. Cell, 1993, 75: 1361-70

[27] Batzer MA, Deininger PL. Alu repeats and human genomic diversity. Nat Rev Genet, 2002, 3: 370-9

[28] Chen JY, Peng Z, Zhang R, et al. RNA editome in rhesus macaque shaped by purifying selection. PLoS Genet, 2014, 10: e1004274

[29] Yang XZ, Chen JY, Liu CJ, et al. Selectively constrained RNA editing regulation crosstalks with piRNA biogenesis in primates. Mol Biol Evol, 2015, 32: 3143-57

[30] Culotta E. What genetic changes made us uniquely human? Science, 2005, 309: 91

[31] Kennedy D, Norman C. What don’t we know? Science, 2005, 309: 75

[32] Chen S, Krinsky BH, Long M. New genes as drivers of phenotypic evolution. Nat Rev Genet, 2013, 14: 645-60

[33] Zhang YE, Long M. New genes contribute to genetic and phenotypic novelties in human evolution. Curr Opin Genet Dev, 2014, 29: 90-6

[34] Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Res, 2010, 20: 1313-26
[35] Chen JY, Shen QS, Zhou WZ, et al. Emergence, retention and selection: a trilogy of origination for functional *de novo* proteins from ancestral lncRNAs in primates. PLoS Genet, 2015, 11: e1005391
[36] Reinhardt JA, Wanjiru BM, Brant AT, et al. *De novo* ORFs in *Drosophila* are important to organismal fitness and evolved rapidly from previously non-coding sequences. PLoS Genet, 2013, 9: e1003860
[37] Ulitsky I, Bartel DP. LincRNAs: genomics, evolution, and mechanisms. Cell, 2013, 154: 26-46
[38] Wu X, Sharp PA. Divergent transcription: a driving force for new gene origination? Cell, 2013, 155: 990-6
[39] Xie C, Zhang YE, Chen JY, et al. Hominoid-specific *de novo* protein-coding genes originating from long non-coding RNAs. PLoS Genet, 2012, 8: e1002942
[40] Gagneux P, Varki A. Genetic differences between humans and great apes. Mol Phylogenet Evol, 2001, 18: 2-13
[41] Blekhman R, Marioni JC, Zumbo P, et al. Sex-specific and lineage-specific alternative splicing in primates. Genome Res, 2010, 20: 180-9
[42] Zhou X, Cain CE, Myrthil M, et al. Epigenetic modifications are associated with inter-species gene expression variation in primates. Genome Biol, 2014, 15: 547
[43] Reilly SK, Yin J, Ayoub AE, et al. Evolutionary genomics. evolutionary changes in promoter and enhancer activity during human corticogenesis. Science, 2015, 347: 1155-9