遗传 2006 28 (3): 329-333 ISSN: 0379-4172 CN: 11-5450/R

技术与方法

基于RefSeq数据库的人类标准转录数据集的构建

李稚锋[1,2], 李玉鉴[3], 赵东升[4], 杭兴宜[1], 王正志[2], 骆志刚[5], 张成岗[1]

1. 军事医学科学院放射与辐射医学研究所，北京100850；2. 国防科技大学机电工程与自动化学院，长沙410073；3. 北京工业大学计算机学院，北京 100822；4. 军事医学科学院卫生勤务与医学情报研究所，北京 100850；5. 国防科技大学并行与分布处理国防科技重点实验室，长沙 410073

收稿日期 2005-3-16 修回日期 2005-7-12 网络版发布日期 2006-3-7 接受日期

摘要

美国国家生物信息技术中心(NCBI)提供了具有生物意义上的非冗余的基因和蛋白质序列的RefSeq参考序列数据库。然而，由于基因普遍存在的多态性以及不同实验室对于序列测定的质量控制存在差异等原因，已发现RefSeq数据库可能存在部分质量问题。文章基于“中心法则”提出“标准转录数据集”的概念，以人类基因和基因组序列为例，利用BLAT、Sim4和自行设计的EIparser等基因结构解析程序分析了RefSeq人类基因转录数据(2005-4-18)与目前所公布的人类标准基因组(2005-4-20)的对应关系。对于有实验证据支持的标记为NM_和NR_的记录，多种程序分析结果表明其与标准基因组完全相对应的记录为9 771个；符合多个程序修订标准的记录10 943个；而与标准基因组有较大差异的记录为203个，多种程序分析结果不一致的记录为2 676个，提示研究人员在使用此非标准转录组数据时，必须考虑到其存在非标准转录的原因甚至存在错误的可能性。本文为基于标准、高质量转录数据集的生物信息学数据分析、分子生物学实验设计、基因多样性和遗传变异分析等提供了重要的参考标准。相关结果可通过http://biocompute.bmi.ac.cn/transcriptome/index.htm访问。

关键词 RefSeq数据库；转录组；质量控制；人类标准转录数据集

分类号 Q754

Construction of Standard Human Transcript Dataset Based on RefSeq and Human Genome Sequence Database

LI Zhi-Feng[1,2], LI Yu-Jian[3], ZHAO Dong-Sheng[4], HANG Xing-Yi[1], WANG Zheng-Zhi[2], LUO Zhi-Gang[5], ZHANG Cheng-Gang[1]

1. *Beijing Institute of Radiation Medicine, Beijing* 100850, *China*;
2. *College of Mechatronics Engineering and Automation, National Univ. of Defense Technology, Changsha* 410073, *China*; 3. *Beijing University of Technology, Beijing, China*; 4. *Beijing Institute of Health Administration and Medical Information*, 100850, *China*; 5. *National Lab of Parallel and Distributed Processing, National University of Defense Technology, Changsha* 410073, *China*

Abstract

The NCBI Reference Sequence (RefSeq) database aimed to provide a biologically non-redundant collection of DNA, RNA, and protein sequences and to promote the research on genes and proteins of human beings and other species. However, because of widely distributed polymorphisms and different quality control of experiments in individual laboratories, there are potential problems need to be identified in the RefSeq database. Regarding which, we herein define the concept, standard transcript, based on the Central Dogmas of Biology that each standard transcript should be perfectly mapped to the standard genomic DNA sequence at the exon level. A large scale analysis for mapping all of the RefSeq records of human being (2005-4-18) to the officially released human genome sequence database (2005-4-20) was further performed using BLAT, Sim4 and a homemade program, EIparser, which was especially designed for this purpose. The standard transcripts based on the RefSeq database were obtained according to the alignment with standard human genome database. There are 9,771 RefSeq records of human being labeled with “NM_” and “NR_” could be perfectly mapped to human genome sequences, while other 10,943 records could be considered as standard transcripts after reasonable revision by comparing with the genome sequences according to all of the three methods. Moreover, the left 203 unrevisable records and 2,676 inconsistent records reported by the above programs could not be considered as standard transcripts and should be checked critically before using because of potential errors in them. Our study has thus provided a reference standard dataset of human beings with high quality for further bioinformatic and experimental analysis such as polymorphism and mutation of human genes. The reference standard dataset based on above criteria could be retrieved from <A href="http://biocompute.bmi.ac.cn/transcriptome/index.htm">http://biocompute.bmi.ac.cn/transcriptome/index.htm</A>.

Key words RefSeq database transcriptome quality control database of standard transcript sequences of human

DOI:

通讯作者 张成岗 zhangcg@bmi.ac.cn