遗传学报 2004 31 (4): 326-334 ISSN: 0379-4172 CN: 11-5450/R

用电子克隆新基因C17orf32和ZNF362对NCBI人类基因数据库模式参考序列5种错误类型的分析与纠正

张德礼1 季梁1 马大龙2 李衍达1

1清华大学生物信息学研究所生物信息学教育部重点实验室、清华大学信息科学技术学院自动化系智能技术与系统国家重点实验室;北京100084;2北京大学人类疾病基因研究中心;北京100083

收稿日期 修回日期 网络版发布日期 接受日期

摘要

采用生物信息学分析与实验确认相结合的技术路线，通过所识别的基因在非冗余数据库比对发现了网上公布的计算机注释人类基因组编码序列存在各种类型的多处错误。该策略既有助于发现更多的人类新基因， 又有助于纠正美国国家生物技术信息中心（NCBI）基因组注释项目公布的参考序列（REFSEQs）中所存在的错误。比如他们采用基因预测方法通过自动计算分析从NCBI contig NT_010808预测到两个模式参考序列LOC124919 和LOC147007，本该都是C17orf32，但却都是C17orf32的不同错误形式，分别为第1和2类型错误； 再如，他们采用基因预测方法通过自动计算分析从NCBI contig NT_004511 预测到三个模式参考序列LOC14907， LOC200084和LOC91126，实际上都是ZNF362一种基因，却提交了ZNF362的3种不同错误形式，分别为第4，5和7类型错误。我们利用计算机识别并结合实验验证能够纠正或避免现有的人类基因组编码序列错误，以前公开发表的文献没有明确指出NCBI人类基因模式参考序列存在错误。因此，应当慎重看待计算机注释的可能存在各种类型错误的人类基因组编码序列。人类新基因的正确识别和注释仍是一项长期而繁重的任务。

关键词 人类基因组 表达序列标签 计算机克隆 模式参考序列 生物信息学

分类号

(Department of medical genetics;Second Military Medical University;Shanghai 200433 China)

Abstract

Key words nuclear receptor transcription factor co-regulator regulation of gene expression

DOI:

通讯作者