• 医学信息研究 •

参考序列数据库构建与数据管理探讨*

郑 思 孙良龙 李 姣

（中国医学科学院/北京协和医学院医学信息研究所 北京 100020）

〔摘要〕 从参考序列数据特点以及数据产生、获取与维护、应用几方面阐述参考序列数据库构建及管理情况，探讨我国大型参考序列数据库建设问题，包括构建精细化的参考序列，促进我国精准医学发展；规范数据处理流程，确保数据质量等。

〔关键词〕 参考序列；数据库；数据管理；非冗余；基因注释；精准医学

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2020.08.004

Discussion on the Building of Reference Sequence Databases and Data Management *ZHENG Si, SUN Lianglong, LI Jiao, Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, China*

〔**Abstract**〕 The paper expounds the building and management of the reference sequence database (RefSeq) from the aspects of the characteristics of reference sequence data, data generation, acquisition, maintenance and application, discusses the building of large reference sequence databases in China, including the building of refined reference sequence to promote the development of precision medicine in China; standardize data processing flow, ensure data quality, etc.

〔**Keywords**〕 reference sequence; database; data management; non-redundant; gene annotation; precision medicine

1 引言

自2003 年人类基因组计划全部完成以来，相继启动国际单倍体型计划[1]、国际千人基因组计划[2]、肿瘤基因组图谱计划[3]和环境基因组计划[4]等一系列人类生命健康相关的重大科学研究计划，对基因组学研究、疾病医疗和药物研发等领域产生巨大影响，能够帮助人们从分子水平探索人类起源和疾病发生发展历程，极大地促进疾病预防、诊断和治疗。测序技术发展大大降低测序成本，序列数据呈现爆炸式增长，其以标准格式存储在计算生物学平台或数据库中。然而公共数据库中序列的多样性给研究人员带来挑战，不同实验室或不同研究项目提交的数据存在冗余。例如国际核酸序列联盟（由美国核酸序列数据库 GenBank[5]、欧洲核酸数据库 ENA[6]和日本 DNA 序列数据银行 DDBJ[7]组成）中序列数据存在很多重复[8]。此外不同个体尤其是不同种族间序列也存在一定差异。因此需要建立一套完整、非冗余、注释信息丰富的核酸和蛋白质参考序列。美国国家生物信息中心（National Center for Biotechnology Information, NCBI）从2000

〔修回日期〕 2020-01-03

〔作者简介〕 郑思，硕士，助理研究员，发表论文10 余篇；通讯作者：李姣，研究员。

〔基金项目〕 国家重点研发计划精准医学专项“中国人群多组学参比数据库与分析系统建设”（项目编号：2017YFC0907503）；“精准医学本体和语义网络构建”（项目编号：2016YFC0901901）。

年开始建立参考序列数据库 RefSeq，为多种生物提供序列相关的数据信息及资料，10 余年来一直是生物学研究领域最具有权威性的序列数据库[9-10]。RefSeq 提供一组注释完整、非冗余、可作为参考标准的序列数据，涵盖基因组、转录本和蛋白质，能实现分子类型、版本管理、基因名称等多维度索引，为生物医学、功能基因组和种群多样性研究奠定基础。本文对 RefSeq 数据特点、产生、服务和应用等方面进行调研，为建立大型参考序列数据库提供参考。

2 参考序列数据库构建

2.1 参考序列数据特点

2.1.1 经过校正的非冗余参考序列集合 自 2000 年首次发布 3 446 条人类转录物和蛋白质记录数据以来，RefSeq 已发展成为涵盖 97 407 种生物（涵盖病毒、细胞器、原核生物、真核生物等），28 730 283 条核酸序列，157 639 958 条蛋白质序列记录的数据库（RefSeq Release 97）。该数据库每天更新，可通过 NCBI 资源中的 Gene 库、Nucleotide 库、BioProject 库、Blast 或者 NCBI 的图形显示访问。

2.1.2 规范化数据编码方式 每条参考序列都有稳定数据编号、版本号和整数识别码，涵盖丰富的数据属性，数据模型与其他国际权威核酸序列数据库相兼容。参考序列编码方式由前缀、下划线和注释 3 部分组成，其中前缀标识序列分子类型，注释部分标识序列审编状态和原始序列来源等。每条参考序列都有完整数据属性，以固定格式存储在数据库中。例如每条序列都准确标注来源物种、物种分类、基因符号和编码蛋白名称、序列组成及特征等。此外 Refseq 为用户提供参考序列相关的生物数据库的交叉引用，确保可以随时追踪到最新研究进展。

2.2 参考序列数据产生

2.2.1 RefSeq 工作流程 RefSeq 是由 NCBI 工作人员及其合作者在对提交到国际核酸序列数据库协作体（International Nucleotide Sequence Database Collaboration，INSDC）的大量冗余序列数据进行收集、审编和注释的基础上而产生的，是对原始序列数据的持续审阅、标注和重新组织。因此 RefSeq 中的参考序列包含经过补充和更新的序列，也包含一部分经过验证但未经修改的原始序列。NCBI 与领域内权威组织开展合作，采用几种不同方法来产生参考序列，具体包括审编通道（Curation Pipeline），注释通道（Annotation Pipeline）和数据提取通道（GenBank Extraction Pipeline），见图 1。

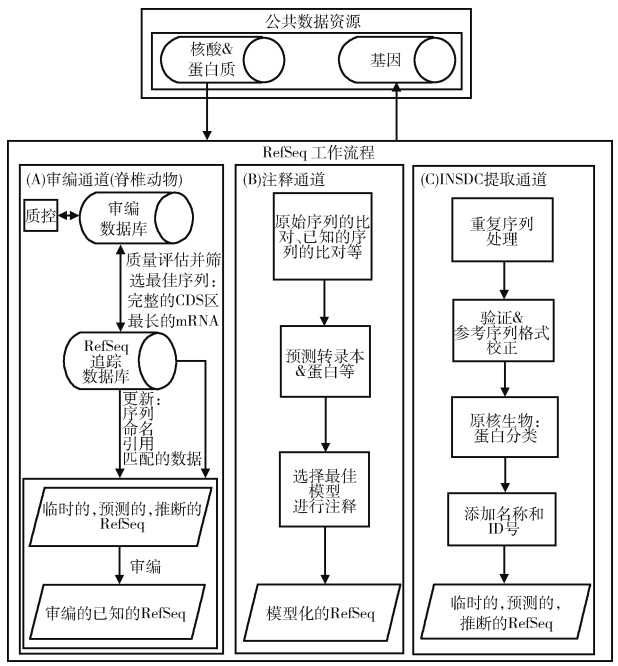

图 1 RefSeq 工作流程[11]

2.2.2 开放领域合作 RefSeq 积极与领域权威组织开展合作，获取序列、命名法、注释等相关生物学资源。例如数据库中人类参考序列的命名规则是由 HUGO 基因命名委员会（HUGO Gene Nomenclature Committee，HGNC）提供的。对于外部合作者构建并提交的参考序列，NCBI 工作人员会对这些序列进行格式调整或检测明显错误（如注释的 CDS 区不能编码相应的蛋白），但不会对其中的注释信息进行额外审编或修改。如果后续验证或实际使用过程中发现这些参考序列存在问题，NCBI 会将这些错误信息告知提交者，在数据库的后续版本中进行更新。RefSeq 数据库提供网站反馈窗口，可用于启动或修改合作协议。

2.2.3 审编通道 通过审核序列比对状态、文

献、质量评估检测以及外部合作者提交的数据资源等来产生核酸和蛋白质参考序列。由审编通道产生的序列称为已知参考序列（Known RefSeq），用 NM_，NR_ 或 NP_ 作为序列前缀标识符。使用规范化的数据审编能极大提高数据利用率[12]。RefSeq 中来源于病毒、线粒体、脊椎动物和部分无脊椎动物的参考序列是经过 NCBI 工作人员审编；而大部分来自细菌、植物和真菌的参考序列数据是由外部合作者审编并提交；还有些序列未进入审编状态。RefSeq 在每条参考序列编号的注释部分标明序列审编状态。序列审编流程包括以下几个步骤：首先，结合自动化序列比对和外部合作者提供的信息，初步定义基因和相关序列。其次，评估数据质量并筛选最佳序列，评估过程包括分析命名方法、序列相似性、基因组定位和潜在的克隆错误等。对于通过质量评估的序列，自动分配 RefSeq 序列编号并标识序列审编状态。最后，开展进一步审编，增加序列相关文献、名称、别名、基因 ID 以及与其他数据库的交叉链接等关于序列特征的注释信息来产生完整的参考序列，同时进一步更新审编状态。对于没有通过质量评估的序列数据，由 NCBI 工作人员和外部合作者共同审核来解决数据冲突。因为审编过程中的歧义必须在参考序列数据生成之前解决。该审编过程将提供更详细的序列信息（如去除污染物、扩展 UTR 区、参考最新文献信息修正序列错误、确定可变剪切位点）和注释信息（增加参考文献、丰富基因和蛋白功能描述、增加成熟蛋白产物等注释特征）。

2.2.4 注释通道 采用 NCBI 的自动化序列注释流程产生参考序列，该过程涵盖将序列比对到基因组，基于序列相似性产生转录本或蛋白产物名称，筛选最优注释模型等。由注释通道产生的序列称为模型化的参考序列（Model RefSeq），用 XM_，XR_ 或 XP_ 作为序列前缀标识符。对于真核生物基因组注释，NCBI 采用一系列计算工具，见表 1，建立包含基因组序列输入、基因组序列遮盖处理、审编过的参考基因组序列比对、蛋白和转录本序列比对、基因预测、小 RNA 注释和筛选最优模型 7 个模块的分析框架[13]，见图 2。最终产生的注释信息包括编码区、保守区、小 RNA、变异、基因和蛋白质产物名称等。

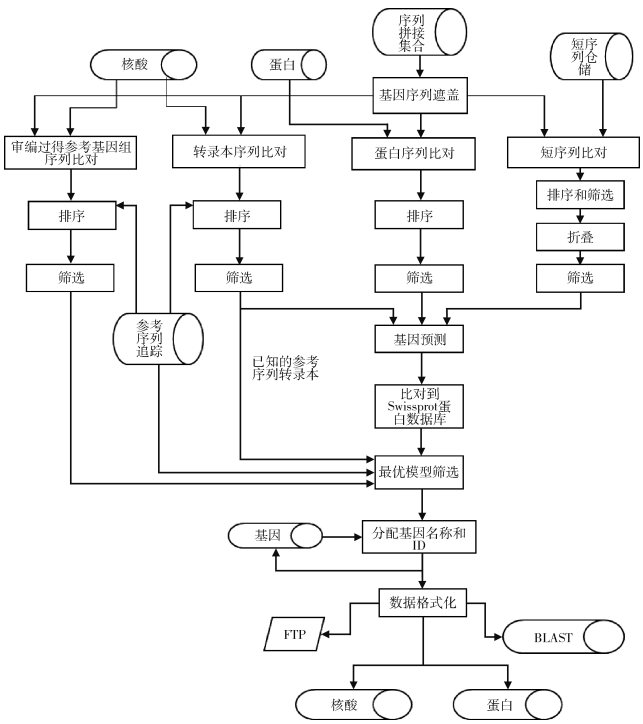

图 2 真核生物基因组注释流程[13]

表 1 真核生物基因组注释相关的预测工具及资源

工具/资源名称	描述
BLAST	序列比对工具 URL：https：//blast. ncbi. nlm. nih. gov/Blast. cgi
Splign[14]	一种用于计算 cDNA 到基因组序列的比对工具，可以准确地定位剪接位点、容忍测序错误并支持跨物种比对 URL：www. ncbi. nlm. nih. gov/sutils/splign/splign. cgi
ProSplign	NCBI 开发的用于将蛋白质序列比对到基因组的工具 URL： www. ncbi. nlm. nih. gov/sutils/static/prosplign/prosplign. html
Gnomon	NCBI 开发的用于真核生物基因预测的工具 URL： https：//www. ncbi. nlm. nih. gov/genome/annotation_ euk/gnomon/
tRNAScan－S[15]	一种用于转运 RNA 预测的工具 URL：http：//trna. ucsc. edu/tRNAscan－SE/

2.2.5 数据提取通道 直接从 INSDC 提取完整、注释过的原始序列，经过去重、格式修改、验证、增加交叉引用等构建参考序列。提取的序列数据类型可分为4 大类：染色体、微生物基因组、小的完整基因组和靶标基因位点。直接来源于 INSDC 的 cDNA 或者 EST 序列也称为已知的参考序列(Known RefSeq)，用 NM_ ，NR_ 和 NP_ 作为前缀标识符。

2.3 参考序列数据获取与维护

2.3.1 数据获取 RefSeq 数据库支持多样化数据检索、查询与获取方式，数据可开放获取，用户使用时不需要认证。(1) 基于 Entrez 检索系统的序列数据查询。支持多种关键词序列数据检索，包括记录名称、相关文献 ID、记录 ID、带注释的染色体和碱基位置和属性等；支持通过检索结果页的分类导航辅助筛选检索结果；支持构建精细化的检索式并提供可视化的检索式构建工具。(2) 基于 BLAST 的检索查询。支持基于记录号、序列片段的相似性检索。(3) 图形化界面检索。检索系统支持用户友好的图形交互界面，支持通过基因组数据浏览器、序列浏览器和基因记录中的图形图像来查看参考序列不同的功能元素注释[16]。(4) FTP 下载。

2.3.2 数据维护 RefSeq 数据库有完善的管理和维护体系，便于数据汇交、存储与共享。首先，参考序列数据库为用户和项目合作者开放数据构建和使用的反馈窗口。其次，RefSeq 数据库处于不断更新状态，更新过程中会保留原有数据条目，便于后续查询和使用。最后，RefSeq 数据库具有完善的数据共享机制，项目组参考序列数据与国际上其他权威序列数据库保持同步和更新，不同数据库之间建立相互连接。

2.4 数据应用

RefSeq 数据库构建为生物医学、功能基因组学和生物多样性研究奠定基础，但序列注释信息在某些方面与其他数据库还存在差异，需要进行持续更新和完善[17-18]。对于基因组注释、基因识别、特征描述、突变和多态性分析、表达研究和比较分析等，RefSeq 中的序列可作为稳定的参考。例如 RefSeq 中的转录本参考序列对于突变位点的功能预测具有重要作用[18]。此外在实验生物学研究中可以使用 RefSeq 中的参考序列进行引物设计等[19-20]。

3 大型参考序列数据库建设思考

3.1 概述

近几年我国医疗行业、科研机构及产业界开始开展不同规模的队列研究。随着我国启动基因组数据资源体系与开放共享平台建设，我国人群序列数据汇聚与有效整合，有助于参考基因序列数据库的构建，支撑我国生命科学的发展[21]。RefSeq 数据库构建和管理可以为国家大型参考序列数据库构建提供参考。

3.2 构建精细化参考序列，促进我国精准医学发展

通过分析基因组学和蛋白质组学等来测定疾病患者遗传学信息，将其用于指导疾病的预防、诊断和治疗，是精准医学在临床上最直接的应用[22]。这些组学分析技术得以开展的重要基础是一个精细化参考序列的构建。因为不同人种遗传背景存在差异，例如单核苷酸多态性位点及频率差异。RefSeq 通过整理、审编和注释核酸序列数据联盟中的原始序列数据，综合考虑不同层面数据，建立信息全面、稳定、非冗余的参考序列。我国是个地域辽阔、人口众多的多民族国家，不同地区、民族之间的基因表型和频率分布往往不同[23]。因此需要通过多中心合作、增加样本人群数量、扩大少数民族在样本人群中的比重等方式来优化采样方法，结合可靠、准确的测序工具及平台，构建符合我国人群遗传特征的精细化的参考序列数据库，促进精准医学发展。

3.3 规范数据处理流程，确保数据质量

我国生物组学数据产量约占全球 40%，是数据产出大国。但是不同机构在组学数据采集、生成和分析过程中采用的方法存在差异，导致获取的数据质量参差不齐，甚至包含许多错误数据，极大影响

数据解读和有效利用[24-25]。可以参考 RefSeq 构建一套规范化、涵盖数据审编和注释通道的序列数据处理流程。对于提交的原始序列数据，需要通过质量控制、计算学分析和人工审编才能进入正式使用。处于不同审编阶段的序列数据都加上明确的审编状态标识并附有详细数据来源信息。此外应建立一套完整的数据质量评估标准来发现审编过程中的问题数据，确保数据冲突在参考序列数据生成之前得以解决。例如对于高 GC 含量、复杂度低和重复序列较多的区域，不同样本之间差异较大，需要建立尽可能涵盖样本差异化的参考序列。

3.4 实现数据统一管理，完善数据服务

RefSeq 数据库通过各个模块（如数据提取、存储和浏览平台）的相互协作来实现数据统一管理，通过 Entrez 检索、FTP、BLAST 对比等方式提供数据开放接口，便于科研工作者使用数据。考虑到组学序列数据的复杂多样性，在建立大型参考序列数据资源时需要完善的数据服务平台。首先，对于接收的序列数据，综合不同的测序平台、实验环境等，将数据以统一格式收录到数据仓储中。其次，构建一个数据索引库，为用户提供检索查询、FTP 下载、API 下载等数据获取方式，确保可以随时追踪到参考序列信息及序列 fasta 格式的下载等。

4 结语

构建参考序列数据库 RefSeq 包括具有稳定注释、非冗余基因组、转录本和蛋白质参考序列数据，通过规范的数据处理流程和管理方式为数据质量提供保障。RefSeq 为人类基因组功能注解提供基础，为突变分析、基因表达和多态性发现等方向的研究提供参考，对加快推进生物医学和疾病生物学研究具有重要意义。通过分析参考序列数据库构建和管理方式可以为大型参考序列数据库组织和运作提供参考。

参考文献

1 Gibbs RA, Belmont JW, Hardenbol P, et al. The International HapMap Project [J]. Nature, 2003, 426 (6968): 789 - 796.

2 Genomes Project C, Abecasis GR, Altshuler D, et al. A Map of Human Genome Variation from Population - scale Sequencing [J]. Nature, 2010, 467 (7319): 1061 - 1073.

3 Cancer Genome Atlas Research N. Comprehensive Genomic Characterization Defines Human Glioblastoma Genes and Core Pathways [J]. Nature, 2008, 455 (7216): 1061 - 1068.

4 H WS, Kenneth O. The Environmental Genome Project: phase I and beyond [J]. Molecular Interventions, 2004, 4 (3): 147 - 156.

5 Benson DA, Cavanaugh M, Clark K, et al. GenBank [J]. Nucleic Acids Res, 2018, 46 (D1): D41 - D47.

6 Silvester N, Alako B, Amid C, et al. The European Nucleotide Archive in 2017 [J]. Nucleic Acids Res, 2018, 46 (D1): D36 - D40.

7 Mashima J, Kodama Y, Fujisawa T, et al. DNA Data Bank of Japan [J]. Nucleic Acids Res, 2017, 45 (D1): D25 - D31.

8 O'Sullivan C, Busby B, Mizrachi IK. Managing Sequence Data [J]. Methods Mol Biol, 2017 (1525): 79 - 106.

9 O'Leary NA, Wright MW, Brister JR, et al. Reference Sequence (RefSeq) Database at NCBI: current status, taxonomic expansion, and functional annotation [J]. Nucleic Acids Research, 2016, 44 (D1): D733 - D745.

10 Pruitt KD, Katz KS, Sicotte H, et al. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI [J]. Trends Genet, 2000, 16 (1): 44 - 47.

11 Bethesda (MD). The NCBI Handbook (Internet) [M]. USA: National Center for Biotechnology Information, 2002: 307 - 328.

12 Cole C, Krampis K, Karagiannis K, et al. Non - synonymous Variations in Cancer and Their Effects on the Human Proteome: workflow for NGS data biocuration and proteome - wide analysis of TCGA data [J]. BMC Bioinformatics, 2014, 15 (1): 28.

13 Bethesda (MD). The NCBI Handbook (Internet) [M]. USA: National Center for Biotechnology Information, 2013: 111 - 130.

14 Kapustin Y, Souvorov A, Tatusova T, et al. Splign: algorithms for computing spliced alignments with identification of

paralogs [J]. Biol Direct, 2008, 3 (1): 20.
15 Chan PP, Lowe TM. tRNAscan - SE: searching for tRNA genes in genomic sequences [J]. Methods Mol Biol, 2019 (1962): 1 - 14.
16 Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high - performance genomics data visualization and exploration [J]. Brief Bioinform, 2013, 14 (2): 178 - 192.
17 Zhao S, Zhang B. A Comprehensive Evaluation of Ensembl, RefSeq, and UCSC Annotations in the Context of RNA - seq Read Mapping and Gene Quantification [J]. BMC Genomics, 2015, 16 (1): 97.
18 Frankish A, Uszczynska B, Ritchie GR, et al. Comparison of GENCODE and RefSeq Gene Annotation and the Impact of Reference Geneset on Variant Effect Prediction [J]. BMC Genomics, 2015, 16 (Suppl 8): S2.
19 Kumar A, Chordia N. In Silico PCR Primer Designing and Validation [J]. Methods Mol Biol, 2015 (1275): 143 - 151.
20 Konermann S, Brigham MD, Trevino AE, et al. Genome - scale Transcriptional Activation by An Engineered CRISPR - Cas9 Complex [J]. Nature, 2015, 517 (7536): 583 - 588.
21 National Genomics Data Center M. Database Resources of the National Genomics Data Center in 2020 [J]. Nucleic Acids Res, 2020, 48 (D1): D24 - D33.
22 武奥申, 刘小娜, 刘昀赫, 等. 二代基因测序数据管理和大数据平台在精准医学中的应用 [J]. 中国生物工程杂志, 2019, 39 (2): 101 - 111.
23 杨昭庆, 褚嘉祐. 中国人类遗传多样性研究进展 [J]. 遗传, 2012, 34 (11): 1351 - 1364.
24 Yohe S, Thyagarajan B. Review of Clinical Next - generation Sequencing [J]. Arch Pathol Lab Med, 2017, 141 (11): 1544 - 1557.
25 Hardwick SA, Deveson IW, Mercer TR. Reference Standards for Next - generation Sequencing [J]. Nat Rev Genet, 2017, 18 (8): 473 - 484.

(上接第 11 页)

卫生事件，信息化在抗疫工作中发挥了重大作用，但也暴露出一些问题与不足。随着国家对医疗卫生信息化的重视和 5G、云计算、大数据、物联网等新技术的发展，医疗领域信息化将会有新一轮的升级与发展，传染病预警预测、智慧医疗、互联网医院、惠民服务等将迎来新的变革。与此同时医卫数据采集的及时性、准确性，医卫信息化人才培育，新技术与传统技术融合，数据安全等问题也是医疗卫生信息化面临的挑战，需要进一步深入研究与探讨。

参考文献

1 中国疾病预防控制中心. 新型冠状病毒肺炎疫情分布 [EB/OL]. [2020 - 04 - 08]. http://2019ncov.chinacdc.cn/2019 - nCoV/.
2 封诚. 告别“表格抗疫”，整合基层数据是数字化社会治理刚需 [EB/OL]. [2020 - 02 - 20]. https://www.hit180.com/42577.html.
3 吴天智, 龙虎, 陈文, 等. 我国区域人口健康信息平台建设现状研究 [J]. 中国数字医学, 2017, 12 (3): 7 - 8.
4 孙卫. 新一代智能区域人口健康信息平台建设探索 [J]. 中国卫生信息管理杂志, 2017, 14 (5): 681 - 685.
5 李伟, 于慧杰, 宋秀军, 等. 区域（市级）人口健康信息平台现状分析与解决方案对比 [J]. 医疗卫生装备, 2017, 38 (10): 55 - 61.
6 黎茂林, 王一峰. 基于云架构的区域人口健康信息平台 [J]. 电脑与信息技术, 2019, 27 (5): 28 - 32.
7 钛媒体 APP. 健康码是个被忽视的奇点事件 [EB/OL]. [2020 - 02 - 27]. https://baijiahao.baidu.com/s?id=1659660386527920498&wfr=spider&for=pc.
8 中国金融信息网. 新冠肺炎疫情将加速我国人工智能企业两极裂变 [EB/OL]. [2020 - 03 - 15]. http://stock.xinhua08.com/a/20200315/1922365.shtml.
9 Salathé M. Digital Pharmacovigilance and Disease Surveillance: combining traditional and big - data systems for better public health [J]. Journal of Infectious Diseases, 2016, 214 (suppl 4): S399 - S403.
10 Simonsen L, Gog J R, Olson D, et al. Infectious Disease Surveillance in the Big Data Era: towards faster and locally relevant systems [J]. Journal of Infectious Diseases, 2016, 214 (suppl 4): S380 - S385.