



生物医学大数据的现状与展望

宁康^{①*}, 陈挺^{②③*}

① 中国科学院青岛生物能源与过程研究所, 单细胞研究中心生物信息学团队, 青岛 266101;

② 清华大学信息科学与技术国家实验室, 生物信息研究部, 北京 100084;

③ 清华大学计算机科学与技术系, 智能技术与系统国家重点实验室, 北京 100084

* 联系人, E-mail: ningkang@qibebt.ac.cn; tingchen@mail.tsinghua.edu.cn

2014-08-26 收稿, 2014-11-26 接受, 2015-01-21 网络版发表

国家自然科学基金(30870572, 61303161, 61103167)和国家高技术研究发展计划(2014AA021502)资助

摘要 生物医学是一门新兴的前沿交叉学科, 它综合了医学、生命科学和生物学的理论和方法而发展起来. 近年来随着先进仪器装备与信息技术等越来越广泛和深入的整合到生物技术中来, 生物医学研究中越来越频繁的涉及到大数据存储和分析等信息技术. 大数据时代的来临对生物医学研究产生了重大影响. 其中, 一个重要发展趋势就是由假设驱动向数据驱动的转变. 数十年来分子生物学水平上的实验目的是获得结论或者是提出一种新的假设, 而现在基于海量生物医学大数据, 可以对海量数据的研究来探索其中的规律, 直接提出假设或得出可靠的结论. 随着先进的生物分析技术的不断推出和更新, 生物医学数据迅速积累. 基于此类大数据一些以往不能解决的问题将有望解决, 同时相关生物医学研究的新问题也层出不穷. 生物医学相关的大数据技术和相关应用主要包括: 基于高通量测序的个性化基因组、转录组和蛋白组研究, 单细胞水平基因型和表型研究, 人类健康相关微生物群落研究, 生物医学图像研究等. 相关生物医学大数据分析任务均具有着数据密集和计算密集的双密集性特点. 要充分地利用这些大数据解决一系列生物医学问题, 迫切需要高通量、高效率、高准确性的生物信息存储和分析策略. 本文总结和回顾生物医学大数据的生成、管理和分析相关的一系列问题, 其中重点讨论人体微生物群落、单细胞表型和基因型、生物医学图像等新近出现的生物医学大数据形式, 以及相关数据分析和应用前景等. 基于目前生物医学大数据的现状我们可以发现, 生物医学大数据的研究正处于蓄势待发状态: 适应于生物医学大数据的软硬件平台、大数据存储、大数据分析挖掘等方法等还不成熟, 制约着生物大数据的研究. 然而一旦相关研究获得突破并有所优化和应用, 将会全方位地支撑生物医学大数据的深入解构; 进而有助于对医学现象的趋势分析和预测, 服务于相关的遗传疾病研究、公共卫生监控、医疗与医药开发等广泛生物医学应用.

关键词

生物医学
大数据
微生物群落
单细胞
医学图像
数据挖掘

生物医学是应用生物医学信息、医学影像技术、基因芯片、纳米技术、新材料等技术的学术研究和交叉领域. 随着以“社会-心理-生物”为代表的大医学模式的提出和系统生物学的发展, 形成了现代系统生物医学^[1,2]. 面向生物医学的系统生物学研究是与21世纪生物技术技术和大数据技术密切相关的领

域, 是关系到提高医疗诊断水平和人类健康的重要研究领域.

随着生物分析技术和计算技术的快速发展, 生物医学产生了大量的数据. 21世纪以来, 随着高通量DNA测序的技术发展和逐步应用, 生命科学领域的数据量正在极速增长. 1977年实现了Φ-X174噬菌体

引用格式: 宁康, 陈挺. 生物医学大数据的现状与展望. 科学通报, 2015, 60: 534-546

Ning K, Chen T. Big data for biomedical research: Current status and prospective (in Chinese). Chin Sci Bull, 2015, 60: 534-546, doi: 10.1360/N972014-00895

全基因组测序；2000年，人类基因组草图被绘制完成^[3]。21世纪尤其是2010年以来，随着新一代测序技术的发展，更大数量级的基因组数据产出日渐增加（从GB、TB级到PB、EB级）：Illumina公司最新的推出的HISEQ X TEN测序仪3天内测序约1.8 TB的碱基数据(<http://investor.illumina.com/phoenix.zhtml?c=121127&p=irol-newsArticle&ID=1890696>)。大规模的基因组数据的分析和处理正在成为推动生命科学创新的重要源泉。

同时应指出的是生物医学大数据不仅仅来源于高通量的基因组和转录组测序。目前其他高通量组学数据，例如单细胞表型数据、动态生物医学图像等数据量也正在急剧增长。生命科学的快速进步，以及生物技术与信息技术的融合，使得大数据贯穿从基础研究到药物开发到临床诊疗到健康管理的所有环节。在基础研究领域，除高通量基因组和转录组测序产生的数据外，代谢组、蛋白质组等领域也正在极速增长，而细胞表型、代谢过程、致病基因等的分析都亟需将不同类型的数据加以整合和解构，从中挖掘出深刻而又非显而易见的生物学规律。

1 生物医学大数据研究的特点

1.1 大数据的重要性

高通量的研究思路和相关数据生产方式的飞跃是大数据产生的主要因素。大数据经历着从概念到小范围技术实践，最终到广泛接受并成为一个新兴研究方向的历程。2008年9月，*Nature*杂志率先出版了由Howe等人^[4]所发论文组成的“大数据专刊”，表明大数据的影响已触及自然科学、社会科学和工程学的各个领域。2009年10月，*The Fourth Paradigm: Data Intensive Scientific Discovery*^[5]一书的出版，显示出与大数据关系密切的数据密集型科学发现范式已被确立和广泛认可。2011年2月，*Science*杂志推出Overpeck等人^[6]所发表文章组成的“数据处理专刊”。2012年5月，联合国发布大数据政务白皮书*Big Data for Development: Challenges & Opportunities*^[7]，体现了大数据领域的研究计划在国家战略层面的重要性。2014年*Science*杂志推出“*Big biological impacts from big data*”^[8]等一系列评论，也明确无误地表明了生物学相关研究已进入大数据时代。

在大数据时代，庞大繁杂的数据以及对数据的

研究对社会、科技、经济的发展将发挥支撑促进作用。大数据本身是一种潜在的战略性资源，具有小规模数据无法匹及的趋势预测潜力，大数据的分析和应用才能将这些资源的效益真正释放出来。美国、欧盟等已在国家和地区层面开展了大数据研究和发展计划，将大数据研究提升到国家和国际重大战略层面。2013年5月9日美国总统奥巴马签署了一项行政命令，要求政府帮助公众和企业更容易取得政府持有的数据，从而促进美国的创新和经济增长。2013年7月习近平总书记在中国科学院考察时指出“大数据是工业社会的‘石油’资源”。谁掌握了大数据以及大数据的研究技术，谁就掌握了主动权。尤其是在生物医学等事关人类健康和命运的研究领域，对相关大数据的研究就是对健康领域未来的掌握。

1.2 生物医学大数据的特点

以高通量测序仪器、单细胞检测装备和实时动态图像系统为代表的新一代生物分析平台已经和正在为生物医学研究提供海量数据，而要充分利用蕴藏于海量数据中的深刻规律，大数据驱动的研究策略必不可少^[9]。大数据至少包含3层含义(3V)(图1)：数据量大(volume of data)，处理数据的速度快(velocity of processing the data)，数据源多变(variability of data sources)。这是那些依赖大数据进行分析和预测过程的重要特征^[10]。具体到生物医学大数据研究而言，大数据研究的3V特点体现如下：第一，生物医学数据量大。通常对于一个样本的人体基因组和转录组(多组织多时间点)测序数据量会分别超过100和30 GB(基于3 GB人类基因组和10~30倍测序深度)。考虑到一次试验中通常会涉及到数百个甚至上万个人体样本，相关的数据量产出十分巨大。第二，研究对于处理结果准确性和处理速度均有较高要求。如个性化医疗^[11]，就具有较高的时效性要求，而单细胞测序及诊断等^[12]，对突变位点和功能模块的鉴别准确性要求较高。第三，相关源数据来源多变且具有较大的异质性。同时生物医学数据的分析和解释通常会利用到NCBI系列^[13]等通用数据库以及UniProt (www.uniprot.org)等专业数据库。源数据和数据库的异质性，会导致数据缺失、数据矛盾等问题的普遍存在，成为相关大数据整合与分析的瓶颈。正是因为生物医学研究具有典型的3V特点，所以需要依靠大数据思维和数据分析策略对生物医学数据进行深入挖掘。

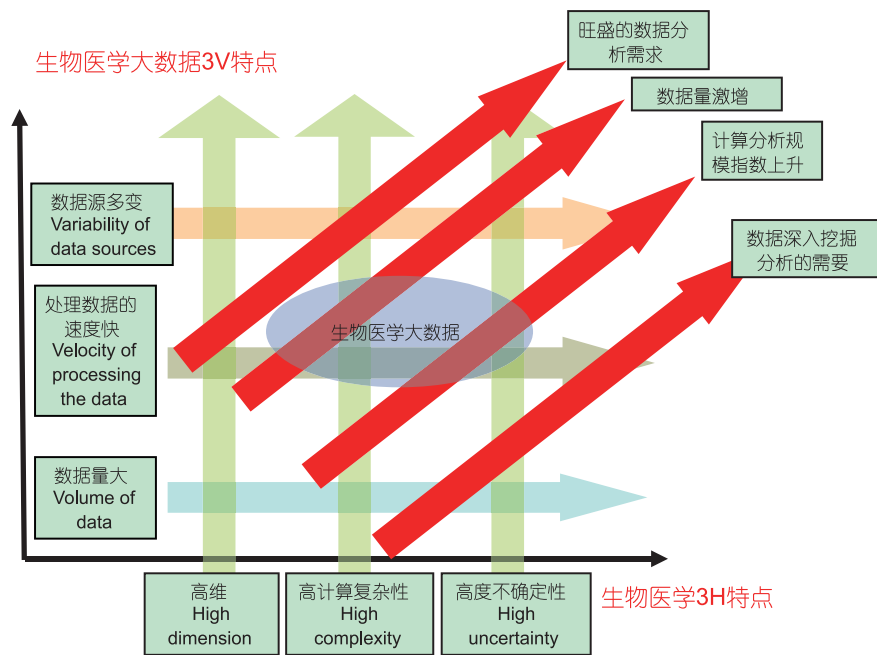


图1 (网络版彩图)生物医学大数据具有3V内涵和3H特点

Figure 1 (Color online) Research in biomedical big-data bears meaning of 3V and properties 3H

生物医学大数据和其他科学大数据一样,具有典型的“3H”特点,即高维(high dimension)、高度计算复杂性(high complexity)和高度不确定性(high uncertainty) (图1)。具体而言,第一,生物医学大数据在对于样本的多重分析角度、多组学数据和多样本量等方面均具有高维特点,需要对多维数据进行叠加、索引、学习。如最近谷歌基于数百天内监测数百万人的流感疫情数据建立FluTrend模型,并利用FluTrend为流感疫情的蔓延提供了一个清晰的图像,进而可以帮助医生能够有效地阻止流感疫情的蔓延(<http://www.google.org/flutrends>);另外美国在电子病历和大数据方面的推进^[14],收集到来自全美数千家医院数百万病人的各类型电子病历。这些高维度数据为发掘蕴含于高维数据中的深刻规律提供了基础,同时在数据整合与分析方面提出了挑战。第二,生物医学研究目标和过程的复杂性包括:不同组学数据的系统性整合需求、不同样本的比对需求、结果的统计验证等等,均需要基于大数据进行数据建模并归纳生物学规律。第三,生物医学研究中样本在来源、处理方法、存储格式上的差异性(heterogeneity)导致研究对象的高度不确定性和不吻合性,需要智能化的数据模型来加以深入分析。

与传统的逻辑推理研究不同,大数据研究是对

数量巨大的数据做统计性的搜索、比较、聚类 and 分类等分析归纳,进行相关性等分析。大数据研究一个重要发展趋势就是由假设驱动向数据驱动的转变。具体到生物医学大数据而言,数十年来分子生物学水平上的实验目的是获得结论或者是提出一种新的假设,而现在基于海量生物医学大数据,可以对海量数据的研究来探索其中的规律,直接提出假设或得出可靠的结论。生物医学大数据的“3H”特点将会是一把双刃剑:在大数据高速积累的同时,数据的差异性将会形成数据整合方面的瓶颈;但是一旦突破此一系列瓶颈,在大数据中蕴含的深刻生物学规律将会极大地促进对于人体健康的理解。

然而,和目前研究较为深入的互联网视频和社交网络等格式化数据相比,生物医学大数据也有其独特之处。首先,生物大数据处理需要复杂的信息提取计算。例如,基因组测序的原始数据是大量的DNA短片段(reads),不同的测序仪可产生长度从100碱基(bp)到10000碱基不同长度的短片段。这些DNA片段所代表的生物学的信息需要通过数据处理才能取得。若是对一个未知的物种的基因组测序,通常这些片段需要通过复杂的拼接算法(de novo assembly),才能将得到基因组的长序列;再通过复杂的统计模型,才能克服测序中的错误,确认基因组中每一个碱

基的可信度(base calling); 然后通过基因预测算法(gene prediction), 预测物种的基因; 最后利用序列比对算法(sequence comparison), 将物种的基因与其他物种功能已知的基因进行序列比较, 才能对这些基因进行功能注释(function annotation). 这些提取的信息才能提供后续的生物功能计算分析. 其次, 由于生命系统本身极其复杂, 例如, 通过对人类疾病的研究我们发现, 基因组中一个碱基的突变就可能对整个生物个体产生深远性的影响, 生物个体成长的过程的环境因素也直接或间接对生物体产生影响. 通过对单细胞的研究也让我们看到, 不仅仅每一个生物个体都与其他个体不同, 同一个体内的每一个细胞都与其他细胞不同, 同一个细胞在不同时间的状态也不同. 因此, 要完整地研究这样复杂的生命系统, 海量的生物医学数据样本也很难满足要求. 如果再考虑到生物医学数据获取的困难和高昂的代价, 因此生物医学大数据的分析就需要更多利用生命系统本身的规律和知识, 建立合理的假设和数学模型, 对数据进行分析 and 解释. 最后, 生物医学大数据的目标是科学的发现, 因此, 对于结果的验证和解释是必须的, 这也是它与其他大数据的不同之处.

基于以上介绍的生物医学大数据的含义和特点, 我们可以看出当前生物医学和大数据研究是相辅相

成、互相促进的, 其交叉结合共同促进相关基础和应
用的研究. 下面我们分别介绍驱动现有生物医学大
数据研究的应用, 生成相关生物医学大数据的先端
生物技术; 然后介绍具体的生物医学大数据, 以及相
关的大数据存储和云计算平台等计算技术; 最后介
绍基于生物医学大数据的生物医学研究新趋势.

2 生物医学大数据的典型应用

典型的生物医学数据包括癌症、个性化医疗等数
据, 其呈现形式包括功能基因组、单细胞、宏基因组
(又称元基因组)数据等. 所有这些数据存储于NCBI
或EBI等大型通用数据库中. 同时随着高通量测序技
术的发展和应用以及生物技术与信息技术的融合,
NCBI等大型通用数据库中生物医学数据类型和数据
规模不断增大^[15](图2).

2.1 现有大型通用生物医学数据库

现有生物医学大型通用数据库包括美国NCBI的
GenBank、欧洲的EBI、日本的DDBJ等. 针对于某些
特定数据或研究对象的数据库如Uni-Prot(蛋白数据
库)、MG-RAST(微生物数据库)也正在快速发展. 这些
都是从事生物信息数据的管理、汇聚、分析、发布等
工作的大型数据库. 近年来, 随着高通量测序技术的

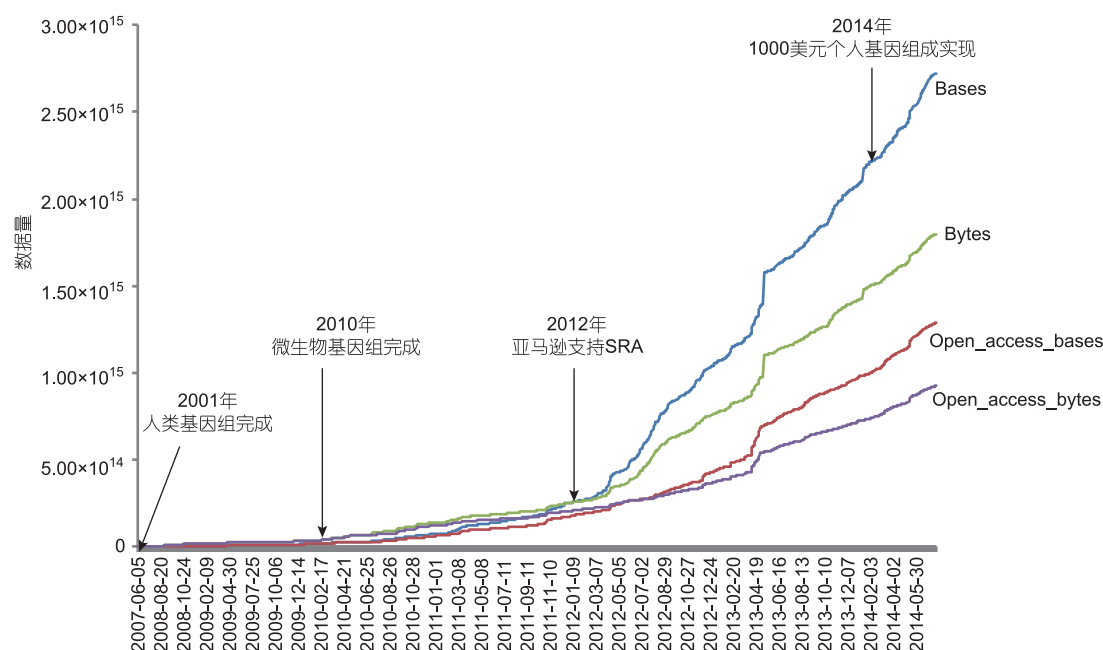


图2 (网络版彩图) GenBank SRA 数据库近年数据量增长情况

Figure 2 (Color online) The increase of data sizes for GenBank SRA database

发展等, 这些大型数据库数据量不断激增, 如表1所示。

2.2 个人基因组以及个性化医疗

2008年11月6日, *Nature*杂志刊登了“第一个亚洲人基因组图谱”论文, 封面名为“你的生命掌握在你手中”^[16]。“第一个亚洲人基因组图谱”的完成是医学方面的重要成就, 这意味着未来5~10年, 一个人只需要花很少的费用就可以拥有自己的基因组图谱。可以预见未来, 医生可以依据这个基因组图谱对病人进行更精确地诊断和治疗, 更可能在发病前就进行必要的干预。甚至连药物都可以根据这个基因图谱为一个人单独设计。可以说这是“你的生命掌握在你自己手中”, “个人基因组时代已经来临”的先兆。基因组图谱结合对基因表达调控等与医学有关知识, 可以对人类认识疾病的发病过程, 对疾病的抵抗性研究将带来新思路。有了“基因组图谱”不仅对疾病治疗有作用, 更重要的是在发病前人们就可以干预、预防这些疾病了。这样, 治病对人们来说将不再是千篇一律了, 甚至给病人的药物都有可能是根据每个人的特点专门设计的。因此个人基因组是个性化医疗的基础, 21世纪将是“个性化医疗”的时代, 在完成1万人的基因组图谱后, 真正的个人基因组时代将到来。到2014年, 1000美元的个人基因组已经实现^[17]。可以预计, 个人基因组图谱的绘制费用将越来越便宜, 并可能成为不少医院看病前的例行程序。

2012年斯坦福大学著名生物学家Snyder教授等研究人员^[18]使用个人的基因组图谱, 同时结合多个高通量生物技术定期监测人体的生理状态, 尝试个性化医疗的可行性。他们提出综合个人组学图谱的概念, 简称为iPOP(integrative personal omics profile)。iPOP对一个人进行长达14个月的医疗跟踪, 除了基因组测序之外, 期间通过血液样本对转录组, 蛋白质组, 代谢组, 微生物宏基因组, 以及个体自身抗体的

分布进行测量, 对包括2型糖尿病在内的各种医疗风险进行分析。分析结果展现个人的健康和疾病状态的各种分子成分和生物通路的广泛动态的变化。这项研究表明, 结合基因组图谱和多样动态的组学信息, 可以解释一个人的健康和疾病状态。在整个实验的过程中, iPOP产生了大量的高通量组学的数据, 在20个时间点监测了总共大约30亿个生物特征, 对这些数据进行复杂的计算分析。可以预见, 未来个性化医疗的广泛应用意味着需要对每一个人的基因组测序, 同时定期地检测、计算分析各样的组学数据, 根据分析的结果为每一个人提供个性化的防治、诊断、和治疗。大量的高通量组学的数据将会产生, 如何存储、分析和保护这些含有个人隐私的生物学大数据, 如何针对数据做医学上的解释诊断等等, 都是我们将要面临的挑战。

针对个性化医疗的浪潮, 目前系统生物学改变现代医疗系统的方式被总结为: 从以病征为主的疾病诊断和治疗, 向基于个体特征的精确治疗转变^[19]。特别是高通量的DNA测序和质谱仪技术的进步使得科学家和医疗人员能够对人体的细胞和组织、体液、身体的表皮、以及排泄物等采样, 非常准确地检测包括基因组、表观基因组(epigenome)、转录组、蛋白质组、新陈代谢组(metabolome)、免疫组(autoantibodyome)、微生物组(microbiome)、以及环境组(envirome)等在内的详细的多样组学信息。综合这些信息不仅能使我们对一个人的健康状况有全局的了解, 而且提供了一个新的途径能够个性化地检测健康状况和提供疾病的防治。如何设定样本采样和数据采集的标准^[20,21], 如何有效地利用和整合、计算分析这些数据将是未来最主要的挑战^[22]。

2.3 人体微生物群落研究

人体微生物群落存在于人的皮肤、口腔、胃、肠

表 1 现有生物医学相关典型数据库的数据量和项目数^{a)}

Table 1 Data sizes and project numbers for current typical databases related with biomedical researches^{a)}

数据库名称	数据库用途	数据量	项目数(估算)	链接地址
GenBank SRA	供公众自由读取的、带注释的DNA序列的总数据库	>3000 TB	>25000	www.ncbi.nlm.nih.gov/genbank
Uni-Prot	蛋白数据库	>1000 TB	>20000	www.uniprot.org
MG-RAST	微生物数据库	52.49 TB	18500	metagenomics.anl.gov
TCGA	癌症组学数据库	~3000 GB	~131000(样本数)	cancergenome.nih.gov

a) 截至 2014 年 8 月

道、血液等,与人体共生,对人的生理和营养有深远的影响。随着人类对于人体微生物群落研究的深入,越来越多以人体为宿主的体内和体外微生物群落,特别是人体肠道群落等,被广泛研究^[23]。据报道,约有100万亿个细菌分布在人体内外,细菌含量约为自身体细胞的9倍,其携带的基因数目大约是人类的1000倍。某些细菌甚至在人体生理机能中作用突出,比如某些细菌能够有效帮助人体构筑免疫系统^[24]有些细菌对促进食物消化不可或缺^[25];还有的可以防止病原体引发潜在病变^[26]。与其说宏基因组在研究微生物菌群,不如说在研究人类“自身”。已有研究称,人类许多疾病如疟疾,脑膜炎,败血症与致病菌有关,也有一些黏膜类疾病与菌群失调有关,甚至某些精神类疾病如抑郁症患者,其肠道内菌群都出现了异变^[27],可以说人体内外的微生物与人体健康息息相关。

目前在发炎性肠道疾病,肥胖症和2型糖尿病的等的病人身上发现,微生物和人体之的动态平衡关系遭到破坏。美国国家卫生院(NIH)在2008年投入超过1 亿美元的资金建立人体微生物基因组研究计划(Human Microbiome Project)[<http://hmpdacc.org/>],用于研究人体内微生物与人体健康的关系^[28]。这些微生物与人体共生(symbiosis),协助人体消化系统的运作,为人体提供必要的维生素,并在机体免疫方面发挥重要作用,能保护人体免受有害细菌的攻击。许多疾病的产生是由于这种共生的关系产生变化。如美国华盛顿大学的研究小组在2009 年发现肥胖病人的肠道微生物的多样性比正常人明显减少^[29]。近期的研究工作发现,在肠道中,微生物菌群在母亲怀孕的过程中自适应转变,帮助母体能产生更多的营养,同时可能导致母体在妊娠期增加体重和减少对葡萄糖耐受度^[30]。在免疫系统里,微生物群落被证明会影响人体的初始T细胞群,说明微生物和人体的免疫系统共同进化^[31]。对人体发育而言,科学家证明生命早期的抗生素暴露会影响脂肪组织,肌肉和骨骼的长远发展^[32]。在临床医学上,微生物群落的移植也成功地成为治疗clostridium difficile的主要方法^[33]。

我国对宏基因组的研究虽然起步较晚,但是进步却是非常快速。以华大基因为代表,通过与欧洲METAHIT CONSORTIUM 的合作,华大基因对124例欧洲人肠道的微生物宏基因组测序分析^[34],发现超过3百万的不同基因,大多数的基因是以前未曾研

究过的。在另一项对糖尿病人肠道宏基因组的研究中^[35],发现糖尿病人肠道微生物群落的多样性明显少于正常人肠道的微生物群落。

因此,可以预见在不远的未来,不论是人类生存环境还是人类自身健康研究领域,将会产生大量的宏基因组的测序数据。如何存储、计算分析这些数据,如何利用这些数据为人类生存环境监控和个性化医疗提供信息,对人类未来的发展至关重要。

3 生成海量大数据的先端生物技术

生物医学大数据的研究依赖于高通量、高质量的数据生成线段生物技术和相关仪器。目前生物医学大数据的来源主要有3种:(1) DNA测序仪器;(2) 高通量高精度质谱仪;(3) 高通量高精度表观型分析仪器。详见图3。

DNA测序仪器: 454, Illumina, PacBio等新一代测序技术的问世^[36](表2),带来生物医学领域的革命。新一代测序技术能够较经济地对基因组进行高效准确的测序^[37]。随着实验技术的成熟和数据分析算法的开发,新一代测序技术不仅大量应用在生物医学数据研究^[38],而且在一些复杂的医学研究项目中也得到实际应用^[39]。相关大数据数量级远超过太字节(TB)级别(表2)。

高通量高精度质谱仪: 蛋白质组学是继基因组测序计划后崛起的一门新兴学科,逐渐成为后基因组时代的研究前沿和热点领域。而代谢组学是继基因组、蛋白组后发展起来的一门学科,主要研究的是作为各种代谢路径的底物和产物的小分子代谢物,在疾病诊断、新药研发、毒理方面都有非常大的应用。近年来随着研究人员对蛋白组学、代谢组学的不断重视以及质谱技术的高速发展,高通量高精度质谱仪产生了越来越多的生物医学数据。相关大数据累积的数量级也已超过太字节(TB)级别(表3)。

高通量高精度表观型分析仪器: 首先,随着荧光蛋白标记等标记式检测方法、红外和拉曼等非标记式细胞检测方法、单细胞操纵等技术的发展,荧光流式细胞分选仪(FACS)、活体单细胞拉曼分选仪(RACS)等单细胞分析和操控平台日渐成熟^[40]。由于单物种群体或群落中的单细胞数量巨大,相关单细胞表观型数据量大且积累迅速。其次,单细胞操纵技术的成熟开启了单细胞测序序幕,相关的海量测序数据将会迅速被生成^[41]。最后,高分辨显微图像的海量生

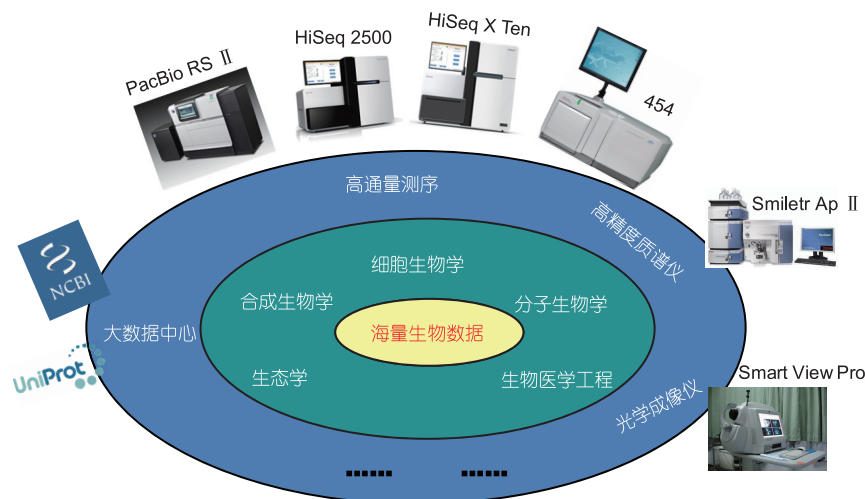


图 3 (网络版彩图)现有生成海量大数据的先端生物技术以及相关仪器和数据库
Figure 3 (Color online) Current advanced biotechnologies, equipments and databases that have generated biomedical big-data

表 2 当前新一代测序技术(PacBio, 454, Solexa)的重要技术指标之比较
Table 2 Comparison of key features of next generation sequencing technologies

基因组测序平台	PacBio	454	Illumina	Illumina
仪器型号	RS II	GS FLX Titanium XL+	HiSeq 2500 (Dual Flow Cell)	HiSeq X Ten (Dual Flow Cell)
读长(b)	8.5 K (平均值)	700 (中间值)	2x150	2x150
单次测序序列数		1 M	600 million	5.3~6 Billion
单次测序碱基数	375 M	700 MB	150~180 GB	1.6~1.8 TB
每次测序时间	<3 h	23 h	40 h	<72 h

表 3 当前新一代高通量高精度质谱仪重要技术指标之比较^{a)}
Table 3 Comparison of key features of high-throughput high-resolution mass spectrometers^{a)}

分析方式	分析仪器	类型	分辨率(K)	质量精度	动态范围(e)	数据量获取速度 (Mb/s)
四级杆技术(quadrupole)	QQQ(triple quadrupole)	离子流	1~2	<1 ‰	5~6	>50
离子阱(ion trap)	LIT	捕获(电子场)	1~2	<1 ‰	3~4	>10
离子飞行时间(ToF)	QToF	离子流	10~50	5~10 ppm	4	>10
环绕飞行技术(orbitrap)	FT	捕获(电子场)	7.5~240	500~10 ppm	4	>1

a) 数据量获取速度由每秒获得的质谱数据估算而来, 以 RAW 数据格式保存

成, 迅速地积累了更为海量的生物医学大数据^[42].

4 生物医学大数据的分析

生物医学大数据可以分为大数据存储和大数据分析两方面, 其中大数据存储服务于大数据的深入分析. 当今生物医学中的典型大数据包括各类基因组数据、宏基因组数据和单细胞数据以及生物医学图像数据等.

4.1 基因组数据分析

在高通量数据生成和系统化数据分析方面, 目前国际上对组学数据的高通量生成和系统化分析已经初步形成了若干通用流程. 在高通量基因组和转录组数据生成方面, 454, Solexa, PacBio等新一代测序技术的引入和推广, 配合高通量数据分析方法, 使更加细致、深入的基因组和转录组数据分析成为可能. 在标准化数据分析流程方面, 包括华盛顿大学的

Tophat-Bowtie-Cufflink 系列, 华大基因的 SOAP 系列^[43], 以及商业化的 CLCBio 系列(<http://www.clcbio.com>)等. 这些系统化分析流程整合了基因组、转录组和部分表观基因组等数据的分析, 极大地推动了生物系统的相关快速、标准化和深入的研究(在此不一一赘述). 随着高通量测序数据的快速积累, 更高水平上的基因组数据整合、挖掘与可视化等分析要求也在提高^[44~46]. 必须通过适应于大数据分析的软硬件系统优化、分析流程的整合、交互式可视化分析平台的建设等方法来实现.

4.2 蛋白质组数据分析

对于蛋白组学, 以高分辨多级串联质谱为代表的质谱分析技术日趋稳定; 通过收集海量的高分辨率一维质谱(MS)和二维质谱(MS/MS)数据, 一些大规模的蛋白组定性和定量分析工作也已完成^[9~11]. 目前蛋白质组学研究向着研究对象更全面(如全面的一级质谱数据独立获取(DIA)数据研究等^[47])和研究规律更深入(如整合不同组学数据进行机制性研究等^[20])的方向发展(在此不一一赘述). 尤为重要的是, 随着高通量数据生成和系统化数据分析方法的日臻成熟, 组学研究发展中的必然要求(如基于细胞内全组分多样性与相互作用的“全局性”分析要求、翻译后修饰等重要调控过程解析要求、表观型调控解析要求等)被提出, 不同层面组学数据之间的融合分析变得益发重要与紧迫.

4.3 宏基因组数据分析

生物学相关的微生物群落大数据分析和数据挖掘任务数量的指数型增长趋势. 目前, NCBI(<http://www.ncbi.nlm.nih.gov>), MG-RAST^[48]以及CAMERA^[49]中公开的宏基因组项目超过10000个, 包含高达数百TB的数据. 保守预计2014~2015年国际上每年会有上万个相关数据分析任务(每个任务中的样本数量从个位数到上千个), 因此其科研和应用市场需求十分旺盛. 其次, 每个微生物群落大数据分析项目的数据量也在增加: 宏基因组数据分析项目的平均数据量达到了10 GB~1 TB量级. 如此巨大的数据量对数据分析的效率和准确性也提出了较高要求.

宏基因组研究领域已经初步建立了如Green-
genes^[50], SILVA^[51]和RDP^[52]等大型特征序列数据库. 同时, 目前微生物群落研究领域存在一系列的

包括所有分析步骤的分析流程, 如面向宏基因组的Phyloshop^[53], QIIME^[54]等. 另外一些大型的微生物群落生物信息学研究网站正在快速发展, 如MG-RAST^[48]和CAMERA^[49], 这些网站通常包括大型的数据库和数据处理平台, 为微生物群落研究和成果的分享提供一站式解决方案. 在单细胞方法方面, 现有的单细胞表观型检测^[55]和单细胞测序的方法^[56]已经初步确定了其可行性, 揭示了其深远意义. 就软硬件架构而言, 目前单细胞研究方法还局限于数量较少的单细胞, 因此数据分析可以通过互相独立的CPU集群完成. 但是由于此方向发展速度极快, 已有迹象显示面向数百甚至上千个单细胞的数据分析时代已经来临^[57].

对于大规模的微生物群落数据(原始数据>1 TB), 单台超级计算机无法完成分析任务, 必须要CPU超级计算机集群才能够完成分析. 而随着微生物群落数据样本成指数增加, CPU超级计算机集群也无法同时分析多个大规模的微生物群落数据. 高效率地分析这些数据需要更加强劲的计算分析平台.

4.4 单细胞数据分析

相较宏基因组数据而言, 单细胞基因组数据分析中的单细胞数量使得其相关数据更为庞大. 如美国单细胞基因组中心SCGC已测定超过40万个单细胞基因组, 总数据量超过100 TB. 同时, 单细胞基因组测序由于受到MDA等DNA扩增技术的影响, 得到的测序深度具有高度不一致性(uneven distribution), 使得相关基因组分析难度更大. 再考虑到单细胞之间异质性和相关性(如基因结构, 基因表达等)分布情况不明的情况, 单细胞数据分析面临着巨大的挑战. 目前单细胞基因组数据分析方法不多, 主要包括改进的velvet^[58]等基因组拼装方法、单细胞基因表达差异化分析方法^[59]等, 还没有专门的方法来实现深入的单细胞异质化分析.

4.5 生物学图像数据分析

随着光学成像仪器和高精度细胞操作技术的日臻进步, 生物学图像相关数据急剧积累, 相关的图像处理技术也日新月异. 随着美国和欧盟在脑科学等生物医学研究领域的再次投入, 众多适应于TB级别的高通量高精度的2D和3D医学图像处理方法被提出^[60], 相关应用潜力也被逐步被认可^[61]. 2012年,

*Nature Methods*推出Baek等人^[62]所发表文章组成的生物图像处理专辑,系统性地总结了现有高通量生物图像处理的方法如通用的imageJ^[63]和针对生物医学图像的PhenoRipper^[64]等.2013年*BMC Bioinformatics*杂志开始整理和搜集生物图像数据和分析相关研究论文(<http://www.biomedcentral.com/bmcbioinformatics/series/BioimageInfo>),相关方法学文章层出不穷.生物医学图像数据处理发展迅猛,但目前还缺乏公认的、标准化的生物医学图像存储和处理平台.

5 生物医学大数据的储存、传输和高性能计算分析

通过对现有生物医学大型数据库(表1)的分析可以发现,新一代生物医学数据本身的特性(如其极端高通量、多来源性、异质性(heterogeneity)等)以及有效利用这些数据的方式(数据质量评价、新生成数据与历史性数据比较、分布式产生而集中性比较、数据的多角度多层次整合分析、信息几乎无限可扩展化等)迫切需要大数据研究方法的支持.而大数据研究方法主要有体现在3个方面:(1)在软硬件平台方面,需要利用云计算等计算体系实现快速、有效的数据分析;(2)在大数据存储方面,需要利用智能化存储系统存储和索引海量数据;(3)在大数据分析挖掘方面,需要通过信息化科研环境开发和测试新型处理算法.生物医学大数据的存储、传输、高性能计算分析和云计算等,是目前生物医学大数据在计算层面的主要需求和瓶颈问题.以下简要介绍相关现状(图4).

数据储存:对于擅长处理结构化数据的关联式数据库管理系统来说,是很难去处理生物医学大数据的.但是, NoSQL等新型数据库技术的出现为处理生物医学大数据提供了新的存储解决方案.如何优化新型数据库,更加智能化的存储生物医学大数据是目前的一大热门研究方向.

数据传输:高速互联网技术是目前可用和可靠的,并在未来相当长一段时间内解决生物医学大数据传输问题不多的软硬件解决方案之一.在另一方面基于高级数据压缩算法的数据传输将会进一步提高含量数据传输效率.

高性能计算:海量生物医学数据的HPC(high performance computing, 实时处理使高性能计算)成为必要.近年来,基于CUDA(compute unified device architecture, 统一计算设备架构)的GPU与CPU协作

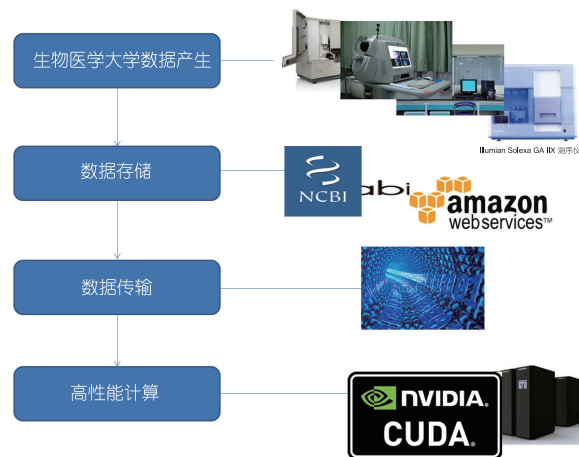


图4 (网络版彩图)现有生物医学大数据数据存储、传输和高性能计算分析流程与主要工具和平台

Figure 4 (Color online) Current data storage, data transfer and high-performance computation pipelines and platforms for big-data analysis for biomedical research

处理方法开始逐渐成为国内外高性能计算领域的研究热点,且GPU由于其强大的并行处理能力和对不同规模数据的自适应性,适合用来对多个基因组或宏基因组数据进行相似度计算与比对,数据库搜索等分析.基于GPU服务器的宏基因组序列处理将会显著提高分析效率^[65],且有助于进一步的数据挖掘.

云计算平台系统:现有生物医学大数据云平台主要利用主流云计算技术结合宽带科研网络、高性能计算和大容量基础设施构建;并通过研发面向生物医学数据特点、覆盖生物医学全领域的云计算技术打造全流程的生物医学研究信息化支持环境.国内外主要的生物医学大数据分析平台及应用有美国的Amazon Elastic Compute Cloud(EC2), Google Compute Engine和Microsoft HealthVault以及中国的浪潮云海等.这些云平台均有服务于生物大数据分析的序列比对、聚类、关联分析等功能.

6 大数据驱动的生物医学研究趋势

纵观各行各业和各个研究领域,人类已进入大数据时代.大数据研究将传统的先验知识驱动的研究方法转变为数据驱动的研究,符合一般的技术创新、发展、成熟规律.具体到生物医学大数据而言,以下几个方面将体现出其数据驱动研究的巨大力量(图5):

(1) 生物医学数据的分布式产生和高度集成化分析和数据建模,涉及到不同类型、不同尺度数据的深度整合:样本表型和基因型和元数据的整合,不同

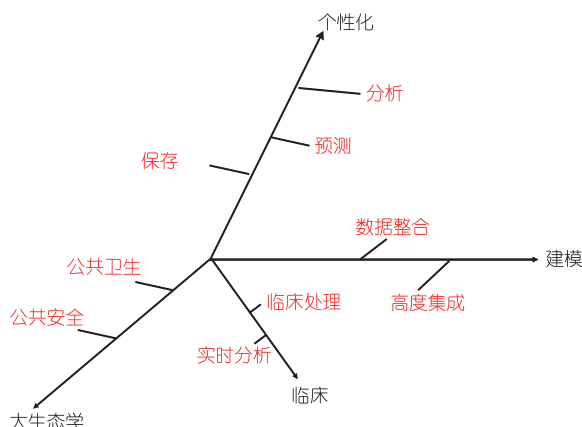


图5 (网络版彩图)大数据驱动的生物医学研究方向

Figure 5 (Color online) Major directions in big-data driven biomedical research

样本的整合,最后构建全方位数据模型.其中大数据整合将会是生物大数据的普适性难题,涉及到数据格式、数据矛盾、数据索引等一系列问题(图6).智能化的数据模型建模和分析将有助于解决以上问题,能够极大地深化对于为身故群落的理解,是生物医学数据研究的一大热点^[66].

(2) 生物医学数据的实时分析和临床处理,涉及到快速准确取样、数据挖掘和知识发现,以及临床处理或其他实时反馈,是生物医学数据研究的另一大热点^[67](图7).

(3) 生物医学数据的个性化分析、预测和安全保存(下一代电子病历)等具有极大的应用潜力.服务于

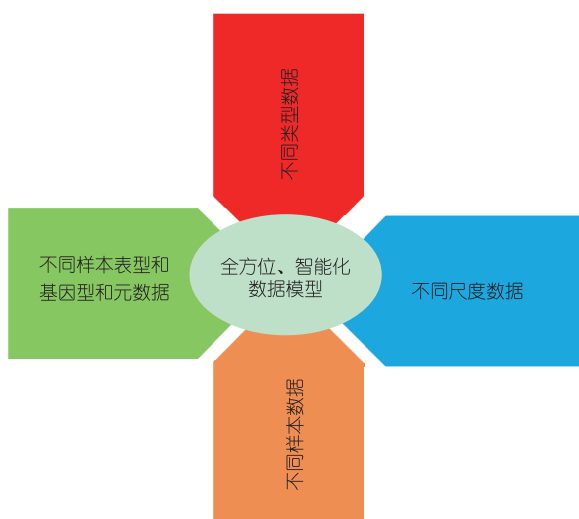


图6 (网络版彩图)生物医学大数据整合和建模

Figure 6 (Color online) Integration and modeling for biomedical big-data

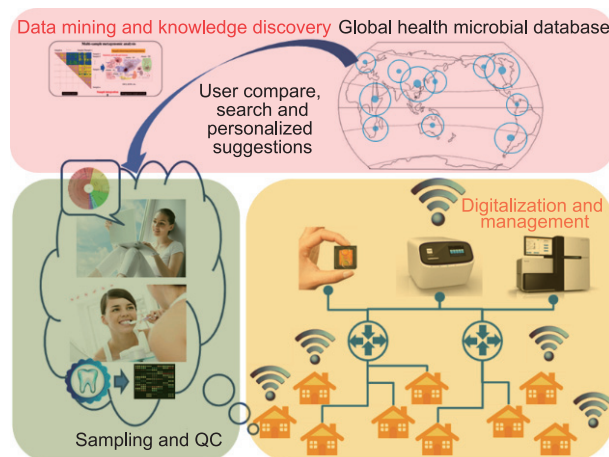


图7 (网络版彩图)生物医学大数据实时分析和反馈

Figure 7 (Color online) Fast process and feedback system for biomedical big-data

“下一代电子病历”建设的生物医学大数据搜集和分析处理,一方面将会搜集来自于海量样本的基因型和表观型数据,另一方面将会对相关数据进行整理和分析,进而提供个性化预测.同时对于个性化数据,将会涉及到海量数据的安全保存问题^[68].

(4) 人体微生物群落研究也是生物医学大数据研究的一个重要方向.医学大数据不光包括人体基因型和表型数据,而且还应包括人体微生物群落等数据和人体饮食、环境等元数据,构成大数据生态系统,共同服务于公共卫生和公共安全研究^[69,70].

人体微生物群落对人体健康的作用可以分为以下多个方面: (1) 通过宏基因组方法建立健全病原菌数据库,配合高灵敏度检测手段,对人体内外病原菌进行检测,找出患者结症所在. (2) 定期检测体内外微生物情况,结合宏基因组分析方法,提醒患者是否存在患病风险,及早干预,预防疾病发生. (3) 通过宏基因组方法,发现有益菌,通过一系列改造,植入人体,提高免疫力. (4) 面对日益紧张的公共卫生环境,及早检测城市环境中的有害菌,提出预防方案. (5) 面对日益突出的恐怖主义袭击,做好应对细菌武器袭击的预案,只有宏基因组学发展到一定的层次才能够快速发现细菌武器的致病菌,并及时做好控制和治疗工作. (6) 宏基因组学研究方法对于干细胞研究,癌症研究等基础医学研究互相借鉴,共同发展.

7 总结

综上所述,整合了大数据研究策略的生物医学

研究将会在更宽的范围和更深入的水平上对人类健康水平的提高做出更大的贡献。随着生物医学相关高通量数字化分析仪器的出现和应用,海量生物医学数据已经和正在被产生。为了充分利用海量数据中蕴含的深刻生物学规律,大数据分析系统必不可

少。适应于生物医学大数据的软硬件平台,大数据存储,大数据分析挖掘等方法的提出、优化和应用,将会全方位地支撑生物医学大数据的深入解构和相关研究对象的趋势分析和预测,进一步服务于相关的遗传疾病、公共卫生、医疗与医药等广泛生物医学应用。

致谢 本文的撰写得到了中国科学院青岛生物能源与过程研究所荆功超和孙政等的大力协助,同时感谢匿名审稿人对本文撰写提出的宝贵意见。

参考文献

- Engel G L. The need for a new medical model: A challenge for biomedicine. *Science*, 1977, 196: 129–136
- Antony P, Balling R, Vlassis N. From systems biology to systems biomedicine. *Curr Opin Biotechnol*, 2012, 23: 604–608
- Collins F S, Mansoura M K. The human genome project. *Cancer*, 2001, 91: 221–225
- Howe D, Costanzo M, Fey P, et al. Big data: The future of biocuration. *Nature*, 2008, 455: 47–50
- Hey A J, Tansley S, Tolle K M. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond: Microsoft Research, 2009
- Overpeck J T, Meehl G A, Bony S, et al. Dealing with data: Climate data challenges in the 21st century. *Science*, 2011, 334: 700–702
- Letouzé E. *Big Data for Development: Challenges and Opportunities*. New York: UN Global Pulse, 2012
- LI Y, Chen L. Big biological impacts from big data. *Science*, 2014, 12: 187–189
- Klauer S G, Dingus T A, Neale V L, et al. The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data. Technical Report, Highways; Safety and Human Factors; I83: Accidents and the Human Factor, US Department of Transportation, 2006
- Dieringer D, Schlötterer C. Microsatellite analyser (MSA): A platform independent analysis tool for large microsatellite data sets. *Mol Ecol Notes*, 2003, 3: 167–169
- Xu J, Wise C, Varma V, et al. Two new ArrayTrack libraries for personalized biomedical research. *BMC Bioinformatics*, 2010, 11: S6
- Kogan S C, Doherty M, Gitschier J. An improved method for prenatal diagnosis of genetic diseases by analysis of amplified DNA sequences. *N Engl J Med*, 1987, 317: 985–990
- Edgar R, Domrachev M, Lash A E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 2002, 30: 207–210
- Bates D W, Saria S, Ohno-Machado L, et al. Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Aff*, 2014, 33: 1123–1131
- Kodama Y, Shumway M, Leinonen R. The sequence read archive: Explosive growth of sequencing data. *Nucleic Acids Res*, 2012, 40: D54–D56
- Wang J, Wang W, Li R, et al. The diploid genome sequence of an Asian individual. *Nature*, 2008, 456: 60–65
- Hayden E C. Is the \$1,000 genome for real. *Nature*, 2014
- Chen R, Mias G I, Li-Pook-Than J, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*, 2012, 148: 1293–1307
- Chen R, Snyder M. Systems biology: Personalized medicine for the future? *Curr Opin Pharmacol*, 2012, 12: 623–628
- Kolker E, Özdemir V, Martens L, et al. Toward more transparent and reproducible omics studies through a common metadata checklist and data publications. *OMICS*, 2014, 18: 10–14
- Snyder M. iPOP and its role in participatory medicine. *Genome Med*, 2014, 6: 6
- Harvey A, Brand A, Holgate S T, et al. The future of technologies for personalised medicine. *N Biotechnol*, 2012, 29: 625–633
- Frank D N, Amand A L S, Feldman R A, et al. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci USA*, 2007, 104: 13780–13785
- Costello E K, Lauber C L, Hamady M, et al. Bacterial community variation in human body habitats across space and time. *Science*, 2009, 326: 1694–1697
- Frey-Klett P, Burlinson P, Deveau A, et al. Bacterial-fungal interactions: Hyphens between agricultural, clinical, environmental, and food microbiologists. *Microbiol Mol Biol Rev*, 2011, 75: 583–609

- 26 Sievert D M, Ricks P, Edwards J R, et al. Antimicrobial-resistant pathogens associated with healthcare-associated infections: summary of data reported to the National Healthcare Safety Network at the Centers for Disease Control and Prevention, 2009–2010. *Infect Control Hosp Epidemiol*, 2013, 34: 1–14
- 27 Messaoudi M, Violle N, Bisson J F, et al. Beneficial psychological effects of a probiotic formulation (*Lactobacillus helveticus* R0052 and *Bifidobacterium longum* R0175) in healthy human volunteers. *Gut Microbes*, 2011, 2: 256–261
- 28 Human Microbiome Jumpstart Reference Strains Consortium, Nelson K E, Weinstock G M. A catalog of reference genomes from the human microbiome. *Science*, 2010, 328: 994–999
- 29 Turnbaugh P J, Hamady M, Yatsunenko T, et al. A core gut microbiome in obese and lean twins. *Nature*, 2008, 457: 480–484
- 30 Koren O, Goodrich J K, Cullender T C, et al. Host remodeling of the gut microbiome and metabolic changes during pregnancy. *Cell*, 2012, 150: 470–480
- 31 Chung H, Pamp S J, Hill J A, et al. Gut immune maturation depends on colonization with a host-specific microbiota. *Cell*, 2012, 149: 1578–1593
- 32 Cho I, Yamanishi S, Cox L, et al. Antibiotics in early life alter the murine colonic microbiome and adiposity. *Nature*, 2012, 488: 621–626
- 33 Van Nood E, Vrieze A, Nieuwdorp M, et al. Duodenal infusion of donor feces for recurrent *Clostridium difficile*. *N Engl J Med*, 2013, 368: 407–415
- 34 Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 2010, 464: 59–65
- 35 Qin J, Li Y, Cai Z, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 2012, 490: 55–60
- 36 Mardis E R. The impact of next-generation sequencing technology on genetics. *Trends Genet*, 2008, 24: 133–141
- 37 Margulies M, Egholm M, Altman W E, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 2005, 437: 376–380
- 38 Warren R L, Sutton G G, Jones S J, et al. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, 2007, 23: 500–501
- 39 Mavromatis K, Ivanova N, Barry K, et al. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods*, 2007, 4: 495–500
- 40 Li M, Xu J, Romero-Gonzalez M, et al. Single cell Raman spectroscopy for cell sorting and imaging. *Curr Opin Biotechnol*, 2012, 23: 56–63
- 41 Ideker T, Galitski T, Hood L. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet*, 2001, 2: 343–372
- 42 Abraham J. Biomedical microanalysis—putting it to work now in diagnostic pathology. *Scan Electron Microsc*, 1979: 171–178
- 43 Li R, Li Y, Kristiansen K, et al. SOAP: Short oligonucleotide alignment program. *Bioinformatics*, 2008, 24: 713–714
- 44 Adams D J, Berger B, Harismendy O, et al. Genomics in 2011: Challenges and opportunities. *Genome Biol*, 2011, 12: 137
- 45 Floratos A, Smith K, Ji Z, et al. geWorkbench: an open source platform for integrative genomics. *Bioinformatics*, 2010, 26: 1779–1780
- 46 Ondov B D, Bergman N H, Phillippy A M. Interactive metagenomic visualization in a web browser. *BMC Bioinformatics*, 2011, 12: 385
- 47 Chapman J D, Goodlett D R, Masselon C D. Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. *Mass Spectrom Rev*, 2013, 33: 452–470
- 48 Meyer F, Paarmann D, D'souza M, et al. The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 2008, 9: 386
- 49 Seshadri R, Kravitz S A, Smarr L, et al. CAMERA: a community resource for metagenomics. *PLoS Biol*, 2007, 5: e75
- 50 Desantis T Z, Hugenholtz P, Larsen N, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol*, 2006, 72: 5069–5072
- 51 Pruesse E, Quast C, Knittel K, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res*, 2007, 35: 7188–7196
- 52 Cole J R, Chai B, Farris R J, et al. The Ribosomal Database Project (RDP-II): Sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res*, 2005, 33: D294–D296
- 53 Shah N, Tang H, Doak T G, et al. Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics. *Pac Symp Biocomput*, 2011, 165–176
- 54 Caporaso J G, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*, 2010, 7: 335–336
- 55 Wu H, Volponi J V, Oliver A E, et al. *In vivo* lipidomics using single-cell Raman spectroscopy. *Proc Natl Acad Sci USA*, 2011, 108: 3809–3814
- 56 Navin N, Kendall J, Troge J, et al. Tumour evolution inferred by single-cell sequencing. *Nature*, 2011, 472: 90–94
- 57 Hou Y, Song L, Zhu P, et al. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell*, 2012, 148: 873–885
- 58 Zerbino D R, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*, 2008, 18: 821–829
- 59 Guo G, Huss M, Tong G Q, et al. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev Cell*, 2010, 18: 675–685

- 60 Peng H, Tang J, Xiao H, et al. Virtual finger boosts three-dimensional imaging and microsurgery as well as terabyte volume image visualization and analysis. *Nat Commun*, 2014, 5: 4342
- 61 Eliceiri K W, Berthold M R, Goldberg I G, et al. Biological imaging software tools. *Nat Methods*, 2012, 9: 697–710
- 62 Baek H C, Ivan C B, Jennifer A B, et al. OMERO. searcher: Content-based image search for microscope images. *Nat Methods*, 2012, 9: 633–634
- 63 Schneider C A, Rasband W S, Eliceiri K W, et al. NIH image to imageJ: 25 years of image analysis. *Nat Methods*, 2012, 9
- 64 Rajaram S, Pavie B, Wu L F, et al. PhenoRipper: Software for rapidly profiling microscopy images. *Nat Methods*, 2012, 9: 635–637
- 65 Su X, Wang X, Xu J, et al. GPU-meta-storms: Computing the similarities among massive microbial communities using GPU. In: *Proceedings of the 7th International Conference on Systems Biology (ISB)*, Huangshan, 2013, 2013: 69–74
- 66 Burgun A, Bodenreider O. Accessing and integrating data and knowledge for biomedical research. *Yearb Med Inform*, 2008: 91
- 67 Abenstein J P, Tompkins W J. A new data-reduction algorithm for real-time ECG analysis. *IEEE Trans Biomed Eng*, 1982: 43–48
- 68 Loukides G, Gkoulalas-Divanis A, Malin B. Privacy-preserving publication of diagnosis codes for effective biomedical analysis. In: *Proceedings of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine (ITAB)*, Corfu, 2010
- 69 Romero L M. Physiological stress in ecology: lessons from biomedical research. *Trends Ecol Evol*, 2004, 19: 249–255
- 70 Boinski S, Cropp S J. Disparate data sets resolve squirrel monkey (*Saimiri*) taxonomy: Implications for behavioral ecology and biomedical usage. *Int J Primatol*, 1999, 20: 237–256

Big data for biomedical research: Current status and prospective

NING Kang¹ & CHEN Ting^{2,3}

¹ *Computational Biology Group of Single Cell Center, Qingdao Institute of Bioenergy and Bioprocess Technology, Chinese Academy of Sciences, Shandong, Qingdao 266101, China;*

² *Bioinformatics Division, TNLIT, Tsinghua University, Beijing 100084, China;*

³ *Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China*

At the frontier of cross-disciplinary sciences, biomedical research combines theory with methods, and biomedical sciences with computation. The recent in-depth integration of advanced equipment and information technology in biotechnology has led to an explosion of data collection, and thus there is a great need for data storage and analysis. Furthermore, the big data era is impacting greatly on biomedical research. In particular, research is transforming from hypothesis-driven to data-driven investigations. For decades, molecular biology research has been hypothesis driven, but the availability of massive biomedical data now allows researchers to directly explore the regularity contained in the data, make assumptions, and draw conclusions. With the fast accumulation of biomedical data, many problems that were unsolvable in the past can now be solved by carefully designed data analysis methods. At the same time, many new problems in biomedical research have emerged. Examples of big data technologies and applications include personalized genomics, transcriptomic and proteomic studies, genotyping and phenotyping of single cells, microbial community research, and biomedical imaging. All these applications are both data intensive and computation intensive, and thus advanced storage and analysis strategies characterized as being high throughput, high efficiency and high accuracy, are urgently needed to process these massive biological data. In this article, we summarize and review several aspects of biomedical big data (data generation, management, and analysis) and focus on data analysis and the application prospects of newly emerging data including human microbiota, the phenotype and genotype of single cells, and biomedical imaging. We conclude that biomedical big data is gaining momentum, although current hardware and software platforms for data-driven analysis remain a significant hurdle. We expect that as big data analysis breaks through this bottleneck, the in-depth research of biomedical big data will make a more significant contribution to clinical diagnosis and treatment.

biomedical research, big data, microbial community, single-cell, bio-imaging, data mining

doi: 10.1360/N972014-00895