微生物全基因组鸟枪法测序

罗春清,杨焕明

(中国科学院遗传与发育生物学研究所人类基因组中心,北京 100101)

摘　要:全基因组测序主要有二种策略,一种是分级鸟枪法测序,另一种是全基因组鸟枪法测序。微生物是一种十分重要的遗传资源,运用全基因组鸟枪法可以方便、快捷地完成其基因组的测序任务。本文对微生物全基因组鸟枪法测序中文库构建、插入片段的长短比例、反应投入量、拼接以及补洞等问题作了较细致的描述,有些步骤作了举例说明。

关键词:微生物;全基因组;鸟枪法测序

中图分类号:Q933　　**文献标识码**:A　　**文章编号**:0253－9772(2002)03－0310－05

Whole Microbial Genome Shotgun Sequencing

LUO Chun-qing, YANG Huan-ming

(*The Human Genome Center, Institute of Genetics and Developmental Biology, CAS, Beijing* 100101, *China*)

Abstract: Two strategies introduced for whole genome sequencing, one is clone by clone method, the other is whole genome shotgun sequencing, for microbes which are very important to us, whole genome shotgun sequencing method is very convenient. In this article we discussed the library construction、long－to－short－ratio of insert, 、total number of reads should be sequenced、assembly and gap filling technologies of the whole microbial genome shotgun sequencing method while some examples presented.

Key words: microbial; whole genome; shotgun sequencing

微生物个体小,作用大,有着非常重要的工农业和医学应用价值。自从 1995 年第一个微生物嗜血流感菌(*Haemophilus influenzae Rd*)[1]基因组的全序列完成全部测定以来,到 2001 年底已有总共 66 个微生物基因组(不包括病毒,以下同)完成了全序列测定并向国际公共数据库递交,另有 123 个微生物基因组已完成测序尚未递交或正在测序(http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/bact.html)。近年来,由于实验技术和生物信息学方法的不断进步,每年完成测序的微生物全基因组数更是呈现出线性增长的趋势(见图 1);这些已测序并递交的微生物主要包括病原微生物[1～7]、特殊环境下生活的微生物[8～12]以及重要的工业微生物等[13],其中,美国、日本和欧洲实验室完成的占绝大部分。

真正意义上的大规模测序得益于 20 世纪 80 年代末期荧光自动分析仪的发明[14,15],而鸟枪法测序(shotgun sequencing)策略为大规模测序提供了技术保障,该方法首先将一条完整的目标序列随机打断成小的片段,分别测序,然后利用这些小片段的重叠关系将它们拼接成一条一致序列[16]。80 年代初 Sanger 成功地用鸟枪法完成了一种 λ 噬菌体的全基因组序列测定[17],之后该方法又被成功地应用于更大一些的病毒 DNA[18]、细胞器 DNA[19,20]、以及细菌基因组 DNA 的序列测定[1]。基于鸟枪法的全基因组测序策略主要有 2 种:(1)分级鸟枪法测序(即 clone by clone 法测序)和(2)全基因组鸟枪法测序[16]。

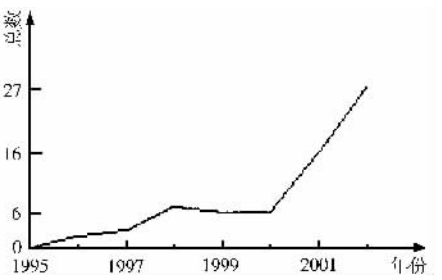

图 1　1995 年以来每年完成测序的微生物基因组数

Fig. 1　Microbial genomes sequenced every year since 1995

收稿日期:2002－03－01;**修回日期**:2002－04－11

作者简介:罗春清(1973－),男,湖北人,博士在读,专业方向:基因组学。Phone:010－80481355,E-mail:luochq@genomics.org.cn

分级鸟枪法测序是先构建微生物基因组的物理图谱,然后从物理图谱中挑选出一组重叠效率较高的克隆群,进行鸟枪法随机测序。国际合作的人类基因组计划就是采用的该方法。由于在测序中每个克隆都是相对独立的,这样计算机在处理时就相对容易一些,也减少了补洞阶段的工作难度。这种方法在对基因组较大的物种进行测序时有着优势,有利于不同的实验室之间开展合作,在早期的微生物全基因组测序中应用广泛[12,21,22]。全基因组鸟枪法测序直接将全基因组进行随机打断成小片段 DNA,构建质粒文库,然后测序。这种方法的优点是省去了复杂的构建物理图的过程。由于计算能力和拼接软件功能的不断提高,用这种方法对微生物全基因组进行测序已越来越普遍[9,23~26],甚至基因组较大的物种也使用这种方法,如果蝇(*Drosophila melanogaster*)全基因组测序[27]、Celera 公司进行的人基因组测序[28]等。以下主要介绍该策略。

1　测序步骤

先提取微生物基因组 DNA,纯化,用酶切或超声波的方法将 DNA 打断,电泳,分别回收不同大小的 DNA 片段,构建质粒文库、转化宿主菌,扩增培养,提质粒作为模板,做 PCR 测序反应,上测序仪;处理所有的测序数据,去除低质量和受载体污染的反应,拼接得到一组克隆群(contig);利用正反向信息等确定这些克隆群之间的位置关系,补洞(gap),最后得到一个完整的没有洞的基因组一致序列(见图 2)。

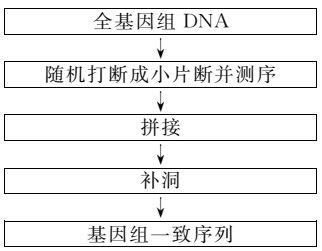

图 2　全基因组鸟枪法测序步骤

Fig. 2　Steps of whole genome shotgun

2　文库构建及测序

全基因组鸟枪法测序的第一个步骤是将全基因组 DNA 打碎,电泳,回收 1.5～3kb 的片段,和插入载体连接,构建质粒文库,目前构建文库的载体一般为 pUC18,也可以为噬菌体载体、复合载体等。将构建好的文库转化大肠杆菌(宿主菌),将大肠杆菌进行扩增培养就可以得到大量的的模板 DNA 可用于测序,由于插入位点两侧的序列相同,因此以两侧相同序列为引物对不同的模板进行 PCR 测序反应,实现了测序的规模化,这种方法叫做正反向末端测序法(pairwise end sequencing)[28]。由于插入片段的大小大致确定,因此同一个插入片段两侧的测序反应在一致序列上的位置就相对固定,呈对应关系,这种关系在以后的 scaffolds 构建过程中有很大的帮助作用。

一般插入片段越短,文库越容易构建,不过由于基因组中一些不可克隆区域的存在,如果这些区域的长度大于插入片段长度,则无法利用正反向反应确定这个区域两侧克隆群的位置关系,这样就需要一些插入片段较长、足以跨过这些不可克隆区域的克隆,利用正反向末端测序,确定这些区域的大小和其两侧克隆群的位置,因此,可视被测基因组中难以克隆区域的分布情况另外构建一个或多个插入片段较长的文库。James L. Weber 和 Eugene W. Myers 认为在全基因组鸟枪法测序中使用相对较长的插入片段有助于克服序列中的重复序列对拼接的影响,而且他们在对人类基因组鸟枪法测序的模拟拼接中认为,对于重复序列含量不同的基因组,长插入片段数量和短插入片段数量的比例(Long－to－Short Insert Ratio)应有很大的不同[29]。Andrew F. Siegel 等人针对全基因组鸟枪法运用正反向末端测序中不同插入片段的长度、比例构建了一个优化模型[28]。

3　全基因组大小预测

对基因组大小的预测,可以指导测序反应投入量,控制测序成本。假设文库随机,拼接后形成的片段沿染色体成泊松(Possion)分布,则[30]:

$$X = Ne^{-NW/G}$$

其中:X 为拼接后的克隆群数;G 为基因组大小;W 为插入片段的平均长度;N 为投入的反应数。因为理论上基因组大小 G 为一固定值,而实际上插入片段的平均长度 W 的值随反应数变化不大,因此 W/G 为一常数,X 随 N 的变化而变化,成函数对应关系。将 X 对 N 作图,可得一条曲线,曲线的顶点对应克隆群数目的最大值,在此点之后,反应数的增多只会使克隆群的总数降低,而不同大小的基因组,因为 G 值不同,会使曲线和其顶点位置发生改变。这样,先预估一个和基因组大小相近的 G 值,然后取不同的反应数拼接,我们将拼接后的克隆群总数对应反应数作图,绘成一个点,一组这样的点就构成了实际曲线。如果基因组的估计值 G 和实际的基因组大小相近,则实际曲线应和理论曲线吻合相对较好,否则应重新取值,直至到曲线最吻合为止,此时 G 值即为和基因组大小最相近的估计值。上述步骤可以通过编写程序自动完成。在我们对一种叫做螺旋藻(*Arthrospira spirulina*)的原核生物测序过程当中,对其基因组大小进行了预测,预测时投入反应数为 32 000 个,平均读长 427bp,覆盖率约为 2.4,从这 32 000 个反应中,随机抽取参与拼接的反应,每次拼接抽取的反应数按 1000 的倍数递增,先假定其基因组大小为 6.2Mb,将拼接后产生的克隆群数对应投入反应数作图,如图 3;两条曲线基本吻合,顶点在同一位置,证明实际基因组和所预测的大小相当。

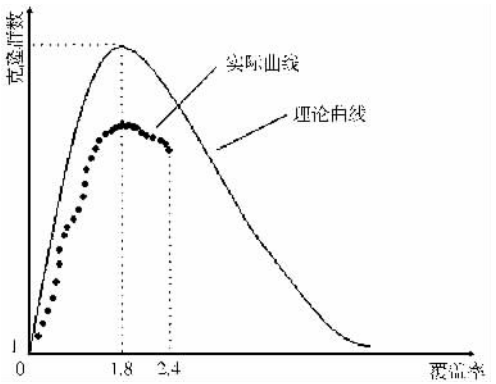

图 3 藻旋藻基因组大小预测

Fig. 3 Genome prediction of *Arthrospira spirulina*

4 拼 接

目前用于全基因组拼接的软件较多，例如 Phrap (Http://www.phrap.org)、CAP3[31]、assembler[32]、STROLL (http://genetics.med.harvard.edu/~tchen/STROLL/)等。其中 Phrap 和 Phred、Swat、Crossmatch、以及 Consed (Http://www.phrap.org)等一起构成了一个十分稳定的拼接软件包，在拼接基因组比较小，重复序列含量又较少的微生物基因组时比较实用。

一般测序微生物基因组需要达到大约 10 倍覆盖率(指高质量碱基)的测序量，这需要测定几万到十几万甚至更多的反应，视不同物种而定。这些反应中混杂了一些无用的序列，例如在转化时插入片段过短或没有插入片段、模板不纯等都可导致反应中污染载体和宿主菌基因组序列，这些污染的载体和宿主菌基因组序列在拼接前应屏蔽掉，污染较严重的可不参于拼接，另外对于一些质量较低的反应，也可以先去掉，这样可以减少参加拼接的反应量，不影响拼接效果，却又可以加快拼接速度；我们在做螺旋藻基因组拼接时，做过一个测试，用于测试的反应数共有 134 224 个，质量大于 20 的碱基覆盖率为 8.9，用于拼接的服务器为 Sun10000，占用一个节点，8 个 CPU，拼接软件为 Phrap，在计算机设置全部相同的情况下，我们作了 3 次拼接对照，其中的拼接时间和计算机资源占用情况见表 1：当然在反应数固定的情况下，拼接时间还和反应质量、重复序列含量、计算机当时的负载量、拼接软件的功能以及软件的编译情况等有关。如果采用在拼接前先屏蔽掉重复序列，拼接完后再将重复序列还原到一致序列中的方法，可加快拼接速度。

表 1 不同反应数拼接时间对照表

Table 1 Assemble times needed while number of reaction different

反应数(个)	拼接时间	最大占用内存
134 224	——	>4G
98 050	约 3 天	2～3G
71 043	9 小时	<2G

拼接后形成一组克隆群，这时一些拼接问题会在这些克隆群中出现，如有些区域覆盖率太低(例如，只有仅仅 1×)、一致序列的质量太低和拼接错误(如图 4)等，这些情况应当纠正，低质量区域应当重测，覆盖率低的区域可以增加一些该区域的测序反应，拼接错误区域(错拼区域)可以用 Consed 软件来手工编辑改正，或在错拼区域两边设计一对引物，进行 PCR 扩增，将产物测序并与错拼区域对照以确定拼接正确与否。低质量区和低覆盖率区域也可以放在全基因组测序的最后阶段来纠正。

5 Scaffolding 和补洞

当反应数达到一定的覆盖率之后，应可以结束随机测序阶段，当然由于建库质量不同，何时停止随机测序还应根据实际情况适时决定。然后依照正反向反应的对应关系，确定克隆群的前后位置关系，这样一组前后位置关系确定的克隆群就称为 scaffolds，构建 scaffolds 的过程叫做 scaffolding。通过 scaffolding 以后，会形成一组组前后位置关系已经确定的 scaffolds，而这些 scaffolds 最后还不能拼成一个统一的完整的基因组，它们之间还存在着洞，这些洞的位置关系不确定，被称之为 physical gap，而 scaffolds 内部的洞称为 sequence gap。在实际情况下，如果前期随机测序投入的反应比较多，则后期 scaffolding 中的 scaffolds 数量就会减少，这样需要设计来进行随机 PCR 反应的引物就少，然而由于一些重复序列和一些不可克隆成分的影响，当反应达到一定覆盖率后，增加反应也很难降低克隆群的总数了，或对降低克隆群总数影响不明显时，就可以进入 scaffolding 了(见图 5)。

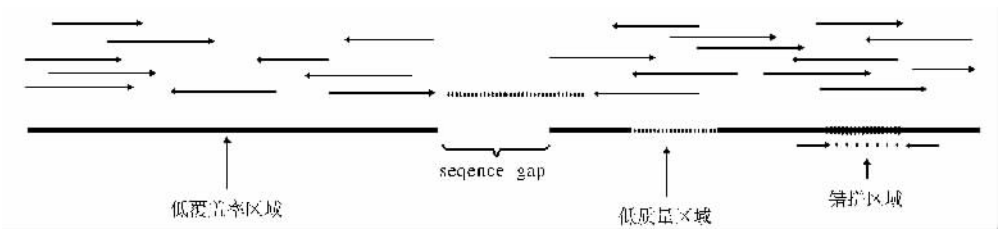

图 4 拼接中出现的一些问题示意图

Fig. 4 Some problems in contigs of assemble

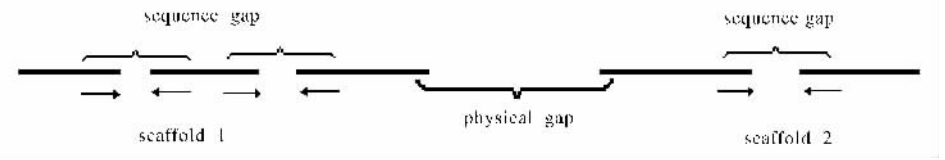

图 5　Scaffolds 示意图

Fig. 5　Relation of scaffolds

Sequence gap 可以用横跨洞的克隆 DNA 作模板,设计面向洞的延伸引物,作 PCR 测序反应来填补;而最后剩下的 physical gap 则可以在洞两边克隆群末端设计出面向洞的引物,然后将这些引物进行随机配对,以基因组总 DNA 为模板,进行 PCR 反应,以产物的有无来确定它们的位置关系。补洞阶段是微生物基因组测序阶段最为消费时间的阶段。理论上讲只要两个克隆群的前后位置关系确定,就可以通过 PCR 测序反应将它们连结起来。不过有些特殊区域如它们本身容易形成二级结构,因此造成了测序反应难以成功,或是测序序列质量低,不能满足拼接要求,特别是一些简单重复系列,不过这些系列在微生物中一般比较少。有人采用将这些难以测序的序列区域重新随机打断成更小片段再测序的方法,有效地克服了二级结构的问题,可以获得质量较高的测序反应[33]。

为了保证基因组的拼接正确性,最后可以有选择地用几种限制性内切酶对基因组进行电子酶切,同时与实际的酶切图谱作对照,如果两张图谱有不符的地方,则要重新拼接。最后按照人类基因组计划 DNA 一致序列的递交标准,基因组完成后的序列应该保持错误率低于 0.01%。

所有测序工作完成后,应该进行基因组的注释(annotation)工作,找出所有的可能的基因,对它们的功能进行预测,并可以在此基础上对微生物的代谢途径、进化等等进行研究。应该说这是基因组测序的目的。不过,由于前一阶段的测序工作是后一阶段工作的前提,因此对这一阶段的战略问题进行正确的决策,对降低研究费用和正确完成研究工作是至关重要的。

参 考 文 献(References):

[1] Fleischmann R D, Adams M D, White O, *et al*. Whole－genome random sequencing and assembly of *Haemophilus influenzae* Rd[J]. Science, 1995, 269: 496～512.

[2] Fraser C M, Gocayne J D, White O, *et al*. The minimal gene complement of *Mycoplasma genitalium*[J]. Science, 1995, 270: 397～403.

[3] Himmelreich R, Hilbert H, Plagens H, *et al*. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*[J]. Nucleic Acids Research, 1996, 24: 4420～4449.

[4] Tomb J F, White O, Kerlavage A R, *et al*. The complete genome sequence of the gastric pathogen *Helicobacter pylori*[J]. Nature, 1997, 388: 539～547.

[5] Stephens R S, Kalman S, Lammel C, *et al*. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*[J]. Science, 1998, 282: 754－759.

[6] Tettelin H, Nelson K E, Paulsen I T, *et al*. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*[J]. Science, 2001, 293: 498～506.

[7] Kunst F, Ogasawara N, Moszer I, *et al*. The complete genome sequence of the gram－positive bacterium *Bacillus subtilis*[J]. Nature, 1997, 390: 249～256.

[8] Kawarabayasi Y, Hino Y, Horikawa H, *et al*. Complete genome sequence of an aerobic hyper－thermophilic crenarchaeon, Aeropyrum pernix K1[J]. DNA Research, 1999, 6: 83～101.

[9] Nelson K E, Clayton R A, Gill S R, *et al*. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritime*[J]. Nature, 1999, 399: 323～329.

[10] Bult C J, White O, Olsen G J, *et al*. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*[J]. Science, 1996, 273: 1058～1073.

[11] Smith D R, Doucette－Stamm L A, Deloughery C, *et al*. Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics[J]. J Bacteriol, 1997, 179: 7135～7155.

[12] Ng W V, Kenneedy S P, Mahairas G G, *et al*. Genome sequence of Halobacterium species NRC－1[J]. Proc Natl Acad Sci USA, 2000, 97: 12176～12181.

[13] Bolotin A, Wincker P, Mauger S, *et al*. The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp. *Lactis* IL1403[J]. Genome Research, 2001, 11: 731～753.

[14] Hunkapiller T, Kaiser R J, Koop B F, Hood L. Large－scale and automated DNA sequence determination[J]. Science, 1991, 254: 59～67.

[15] Kaiser R J, MacKellar S L, Vinayak R S, *et al*. Specific－primer－directed DNA sequencing using automated fluorescence detection[J]. Nucleic Acids Research, 1989, 17: 6087～6102.

[16] Batzoglou S, Berger B, Mesirov J, Lander E S. Sequencing a genome by walking with clone－end sequences: a mathematical analysis[J]. Genome Research, 1999, 9: 1163～1174.

[17] Sanger F, Coulson A R, Hong G F, Hill D F, Petersen G B. Nu-

cleotide sequence of bacteriophage λDNA[J]. J Mol Biol, 1982, 162:729～773.

[18] Goebel S J, Johnson G P, Perkus M E, *et al*. The complete DNA sequence of *Vaccinia virus*[J]. Virology, 1990, 179:247～266.

[19] Oda K, Katsuyuki Y, Ohta E, *et al*. Gene organization deduced from the complete sequence of Liverwort *Marchantia polymorpha* mitochondrial DNA[J]. J Mol Biol, 1992, 223:1～7.

[20] Ohyama K, Fukuzawa H, Kohchi T, Shirai H, *et al*. Chloroplast gene organization deduced from complete sequence of livewort *Marchantia polymorpha* chloroplast DNA[J]. Nature, 1986, 322:572～574.

[21] Kawarabayasi Y, Sawada M, Horikawa H, *et al*. Complete sequence and gene organization of the genome of a hyper－thermophilic archaebacterium, *Pyrococcus horikoshii* OT3[J]. DNA Research, 1998, 5:55～76.

[22] Kaneko T, Sato S, Kotani H, *et al*. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis sp*. Strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein－coding regions[J]. DNA Research, 1996, 3:109～136.

[23] Hayashi T, Makino K, Ohnishi M, *et al*. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K－12[J]. DNA Research, 2001, 8:11～22.

[24] Fraser C M, Gocayne J D, White O, *et al*. The minimal gene complement of *Mycoplasma genitalium*[J]. Science, 1995, 270: 397～403.

[25] May B J, Zhang Q, Li L L, *et al*. Complete genomic sequence of Pasteurella multocida, Pm70[J]. Proc Natl Acad Sci USA, 2001, 98:3460～3465.

[26] White O, Eisen J A, Heidelberg J F, *et al*. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1[J]. Science, 1999, 286:1571～1577.

[27] Eugene W Myers, Granger G Sutton, Art L Delcher, *et al*. A whole－Genome Assembly of *Drosophila*[J]. Science, 2000, 287:2196～2204.

[28] Andrew F Siegel, Ger van den Engh, Leroy Hood, et al. Modeling the Feasibility of Whole Genome Shotgun Sequencing Using a Pairwise End Strategy[J]. Genomics, 2000, 68:237～246.

[29] James L Weber, Eugene W Myers. Human Whole－genome Shotgun Sequencing[J]. Genome Research, 1997, 7:401～409.

[30] Fraser C M, Fleischmann R D. Strategies of whole microbial genome sequencing and analysis[J]. Electrophoresis, 1997, 18, 1207～1216.

[31] Xiaoqiu Huang, Anup Madan. CAP3: A DNA Sequence Assembly Program[J]. Genome Research, 1999, 9:868～877.

[32] Sutton G, White O, Adams M, Kerlavage A. TIGR Assembler: A new tool for assembling large shotgun sequencing projects [J]. Genome Science & Technology, 1995, 1:9～19.

[33] Amanda A McMurray, John E Sulston, and Michael A Quail. Short－Insert Libraries as a Method of Problem Solving in Genome Sequencing[J]. Genome research, 1998, 8:562～566.

·会 讯·

第七届国际人类基因组大会在沪召开

国际人类基因组组织是一家独立的非赢利机构，迄今已有逾十年的历史了，国际人类基因组大会是该组织举办的全球性重要会议。2002 年 4 月 14～17 日，第七届国际人类基因组大会(HGM 2002)在上海国际会议中心隆重举行。来自国内的代表 450 人和来自国外的代表 750 人参加了盛会。国际人类基因组组织主席 Lap－Chee Tsui 和本地大会主席陈竺、强伯勤等主持了会议。

大会就人类基因组研究问题举行了几十场专题演讲，其中包括：单核苷酸多态性图谱的医疗应用(美国 Pui－Yan Kwok)；国际人类基因组测序项目和比较基因组学(美国 Eric Lander)；肿瘤基因组学及其临床应用(日本 Yusuke Nakamura)；乳腺癌和卵巢癌的遗传分析(美国 Mary－Claire King)；血友病的功能基因组学研究(陈竺)；人和小鼠的比较基因组学研究(美国 Mark Adams)；小鼠的功能基因组学研究(美国 Eddy Rubin)；绿河豚的基因组和比较基因组学研究(法国 Jean Weissenbach)；中国基因组多样性项目(金力)；单倍型遗传和人类进化(美国 David Cox)；复杂疾病中的基因组变异和图谱(美国 Aravinda Chakravarki)等。此外，于军还就中国完成水稻基因组序列框架图的报道向与会科学家简要介绍了中国科学家进行的测序工作。

大会审定录用的 680 篇会议论文参加了 Nature(国际自然科学周刊)出版集团(NPG)组织的 Poster 论文竞赛，其中有 5 篇论文获奖，包括英国、美国、俄国、中国台湾和中国大陆各一篇。北京大学人类疾病基因研究中心博士后张德礼等人的论文“人类新基因的电子克隆与实验确认”获得 Poster 奖。

大会开幕之前的 4 月 13 日上午，在上海交通大学举办了“21 世纪的基因科技和生物经济”论坛；下午在复旦大学举办了“基因组学和社会——东西方的对话”公开论坛。

(李绍武)