人类基因组北京区域(3pter-D3S3397): 序列完成图与分析*

中国人类基因组测序协作组**

(① 中国科学院北京基因组研究所, 北京 101300/浙江大学 James D. Watson 基因组研究院, 杭州 310008;
② 国家人类基因组南方研究中心, 上海 201203; ③ 国家人类基因组北方研究中心, 北京 100176)

摘要 "人类基因组计划" 旨在完整而且高质量地测定人类基因组序列, 将对生命科学研究产生深远的影响. 中国作为"人类基因组计划"的六个参与国之一, 承担了 3 号染色体 3pter-D3S3397 区域的测序任务, 该区域被称之为"北京区域". 北京区域的完成图序列全长 17.4 Mb, GC含量为 42%, 遗传重组率平均值为 2.14 cM/Mb. 在该区域鉴定了 122 个已知基因和 20 个预测基因, 并发现了 42607 个单核苷酸碱基多态性位点. 对北京区域序列完成图的综合分析表明: (ⅰ) 基因密度和GC含量在北京区域的分布与细胞遗传学图谱间成对应关系, 即G^-带GC含量高, 基因密度也高, 反之亦然; (ⅱ) 北京区域是 3 号染色体中平均遗传重组率较高的区域, 在亚端粒区的遗传重组率高达 6.14 cM/Mb; (ⅲ) 在GC含量较低的近端粒侧长达 1.1 Mb的基因贫乏区域极可能存在与乳腺相关的大基因; (ⅳ) 包括肿瘤、代谢综合症等在内的多种疾病相关基因定位于此区域. 以上分析充分表明, 北京区域在以医学应用为目标的深层次分子生物学研究中具有非常重要的作用.

关键词 人类基因组计划 北京区域 完成图 基因组概览 序列注释

人类基因组蕴含了大量与人类进化、生长和发育相关的信息. "人类基因组计划"(Human Genome Project, 简称 HGP) 旨在确定组成人类基因组的 30 亿个核苷酸的顺序, 无疑是自然科学史上的一个重大贡献. 与法国、德国、日本、英国和美国一道, 中国正式参与人类基因组计划, 并成为国际人类基因组测序协作组(International Human Genome Sequencing Consortium, 简称 IHGSC)六个会员国之一. 中国科学家完全支持人类基因组序列自由共享的原则, 并出色地按时完成了所承担的相关任务. 这些任务由中国人类基因组测序协作(Chinese Human Genome Sequence Consortium, 简称 CHGSC)组执行完成, 该

2005-05-25 收稿, 2005-07-18 收修改稿
* 中国人类基因组计划和中国人类基因组测序协作组由科学技术部、中国科学院和国家自然科学基金委员会共同资助
** 文后附作者列表

协作组由 3 个科研机构组成: 中国科学院北京基因组研究所(Beijing Genomics Institute, 简称 BGI)、国家人类基因组南方研究中心(Chinese National Human Genome Center at Shanghai, 简称 CHGC)和国家人类基因组北方研究中心(Chinese National Human Genome Center, Beijing, 简称 CHGB).

1999 年, 北京区域最初定为人类 3 号染色体短臂(3p)上从端粒(3pter)到遗传标记 D3S3610 的片段, 遗传距离估计约为 31 cM, 后来通过 CHGSC 的申请和 IHGSC 的统一安排, 该区域扩展到 D3S3397 标记处, 因此, 北京区域的遗传学距离估计为 39 cM. 从细胞遗传学的角度讲, 该区域位于 3pter 到 3p24.3 之间.

1999 年 9 月, 基于细菌人工染色体(bacterial artificial chromosomes, 简称 BAC)物理图谱而进行的大规模测序工作正式启动; 2000 年 5 月, 在 IHGSC 的协调和管理下, 人类基因组框架图序列完成, 共计 63.2 Mb 的原始序列数据被提交到公共数据库 GenBank, 包括填补序列间缺口和更详尽的序列注释分析在内的完成图工作, 一直持续到 2001 年 8 月. 与此同时, CHGSC 和项目资助方在杭州举行的第十届人类基因组测序之国际策略会议上正式宣布了北京区域完成图的完成. 截至 2004 年 7 月, 据 IHGSC 统计, CHGSC 对人类基因组完成图的贡献为 38.1 Mb.

尽管北京区域的初步分析都反映在IHGSC公布的人类基因组序列框架图中[1], 但是对其完成图序列进行全面分析仍然非常必要. 这些分析结果最终将整合到由人类基因组 3 号染色体测序协作组公布和发表的有关 3 号染色体的文章中, 对我们而言, 展现CHGSC在人类基因组计划中的工作和贡献, 全面分析北京区域的区域特征和特殊性, 是义不容辞的工作和任务.

1　材料和方法

1.1　图谱构建、测序和序列组装

依据公共数据库中的遗传标记和物理标记, 采用PCR方法从人类BAC文库RPCI-11 (http://bacpac.chori.org/home.htm)中筛选种子克隆, 将含有DNA序列信息的BACs定位于片段重叠群(Contigs)上. 基于BAC克隆的Contigs通过“BAC克隆末端测序步移法”(www.tigr.org)、“酶切指纹图谱鉴定法”(www.genome.wustl.edu)和荧光原位杂交(Fluorescence *in situ* Hybridization, 简称FISH)等不同策略进行延伸和验证. 通过构建插入片段大小范围为 1.0~3.0 kb的pU18质粒作为载体的BACs鸟枪法文库进行测序. 大部分序列数据在MegaBACE 1000 (Amersham)上进行测定, 其余为ABI 377 (Applied Biosystems)测序. 测序数据通过Phred-Phrap-Consed (http://www.phrap.org)软件包[2~4]进行分析处理, 并通过对每个BAC克隆的逐一分析进行缺口填补和提高序列质量.

1.2　序列注释和比较分析

一致序列(consensus sequence)的 GC 含量分析采用DNA_GC_Calculator (http://www.genome.iastate.edu/ftp/share/DNAgcCal/); CpG 岛通过 CpGPlot (http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/cpgplot.html)进行预测, 满足序列长度不小于 200 bp, GC 含量大于 50%, 且 CpG 岛期望数与观察数之比大于 0.6. 散在重复序列和简单重复序列分别通过 Repeat-Masker (http://repeatmasker.genome.washington.edu)和 Tandem Repeat Finder (http://tandem.bu.edu/trf/trf.html)[5]两个软件进行注释.

基因通过Ensembl自动化流程进行注释, 辅以人工校正. 对于Ensembl注释结果中的基因贫乏区, 我们采用开放阅读框预测软件和序列相似性搜索相结合的方法进行基因预测. 来自于RefSeq[6]和 H-InvDB[7]数据库的人类ESTs和cDNA序列通过BLAT[8]比对, 定位于人类基因组上. 与人类mRNA和cDNA序列相一致的序列被注释为“已知基因”. 通过运行FgeneSH[9]软件进行基因预测, 将FgeneSH预测基因与剪接的ESTs进行BLASTX比对, 若二者同源性≥97%且覆盖度≥90%, 则将FgeneSH预测基因注释为“预测基因”. 对于基因功能注释, 通过BLASTX搜索SwissProt[10], TrEMBL[10] 数据库及分析提取InterProScan[11]结果, 得到蛋白质和功能域的信息. tRNA基因预测使用tRNAscan-SE v1.23[12]软件, 在真核生物模式下选择默认参数, 阈值设为 20 bits. ncRNA 和 microRNA 基因的预测采用BLAST对Rfam数据库(6.1 版)[13]和miRNA Registry数据库(5.1

版)[14,15]进行穷尽搜索，满足所提交序列与公共数据库中的目标序列的同源性至少为 95%，且覆盖度超过 95%.

SNP 鉴定是将所有鸟枪法测序产生的序列(满足长度>200 bp)与北京区域的共有序列进行比对，基于以下标准进行筛选: (ⅰ) SNP 位点和其两侧各 5 个核苷酸的碱基质量要大于 Q20(准确率为 99%); (ⅱ) SNP 位点至少有 2 个以上的峰图序列文件支持和验证. 其他来自于公共数据库 dbSNP(http://www.ncbi.nlm.nih.gov/projects/SNP/, Build 123)的非冗余 SNP 也被定位到北京区域.

与其他哺乳动物间的比较分析中采用以下方法和软件: 用RepeatMasker来屏蔽重复序列，BLASTZ进行序列比较, axtBest/subsetAxt进行数据处理[16]. 所有分析和计算都是在曙光 2000，曙光 3000 (曙光信息产业有限公司), SUN 10K (SUN Microsystems Inc.), SGI 3800 (Silicon Graphics, Inc.) 和IBM P690 (IBM Corp.)等巨型计算机上运行.

2 结果

2.1 克隆图谱和序列图谱

与其他参与人类基因组计划的测序协作组一样，北京区域的物理图谱构建和测序采用逐级鸟枪法策略. 共筛选 401 个物理标记和遗传标记用来覆盖北京区域及其邻近区域的BAC克隆，平均密度约为 23 个标记/Mb. CHGSC的 3 个中心共进行 394 个BAC克隆的测序，产生 652646 个反应(reads). 参照IHGSC为HGP制定的标准[17]，序列质量均超过 99.99%的准确率. 经过填补缺口和排除错误组装后，北京区域的常染色体部分的最小并列途径(tiling path)由 135 个BAC克隆组成，非冗余物理序列长度为 17.4 Mb (图 1). 北京区域的序列数据可从以下网站下载: GenBank (http://www.ncbi.nlm.nih.gov/), UCSC (http://genome.ucsc.edu)和BGI ftp服务器(http://btn.genomics.org.cn/hgp/).

2.2 基因特征

我们对北京区域的高质量序列进行基因注释，一方面采用基因预测软件进行预测，同时还整合了EST, cDNA和蛋白质序列的实验证据. 在注释得到的北京区域的 158 个基因中，122 个基因被确定为已知基因，20 个为预测基因，16 个为假基因(表 1). 虽然北京区域的平均基因密度为 9.1个基因/Mb，但基因分布很不均匀，从 3p26.3 的 4.1个基因/Mb到 3p25.3 的 15.1 个基因/Mb变化不等. 基因密度如此大幅度的变化导致把北京区域分成了两个亚区: 0~8 Mb的基因低密度区和 8~17 Mb的基因高密度区(图 2). CpG岛一般出现在蛋白质编码基因的上游 2 kb和下游 1 kb附近. 在北京区域中，70.4%的蛋白质编码基因包含CpG岛，这个比例与先前对人类基因组框架图分析中，对CpG岛的估计值是基本一致的. 基因区域的长度是 8223518 bp，占北京区域的 47.3%，同先前已经发表的人类基因组第 6 号、7 号、14 号和 22 号染色体的基因区域所占比例(分别为 42.2%，46.5%，43.6%，42.4%和 51%)基本一致[18~22]. 北京区域的平均基因长度为 57912 bp; 平均外显子长度为 244 bp，外显子长度约占北京区域总长度的 3.23%. 最长的外显子来自于基因*SRGAP3*，长度为 5339 bp; 外显子个数最多的基因是*ITPR1*，共有 62 个外显子. 最长的基因是*GRM7*，有 12 个外显子，总长达 842 kb，该基因在大脑中表达，编码代谢型谷氨酸受体 7(mGluR7).

真核生物基因的可变剪切机制产生多个转录本，进而导致所编码蛋白质的结构多样性. 在人类基因组中，可变剪切现象比较普遍，保守估计约有 35%的基因存在可变剪切[23]. 在北京区域，约有 39%的基因发生可变剪切，平均每个基因有 1.5 个转录体. 基因*OGG1*(8-羟基鸟嘌呤糖基化酶)是一个典型的例子，有EST证据证明它有 8 个转录体. 基于定位到北京区域的EST数据集的分析，共发现 1301 个非冗余的供体/受体剪切位点，其中 1280 个(98.39%)为GT-AG，6 个(0.46%)是GC-AG，其他类型(例如: TG-AG, TT-AG, GA-GG)为 15 个(1.15%).

同蛋白质编码基因相比，非编码 RNA(ncRNA)基因很难预测并且少有实验证据. 这些非编码 RNA 基因依据其结构与所行使的催化以及调节功能的不同而进行分类，现在这些类别还在不断增加. 在北京

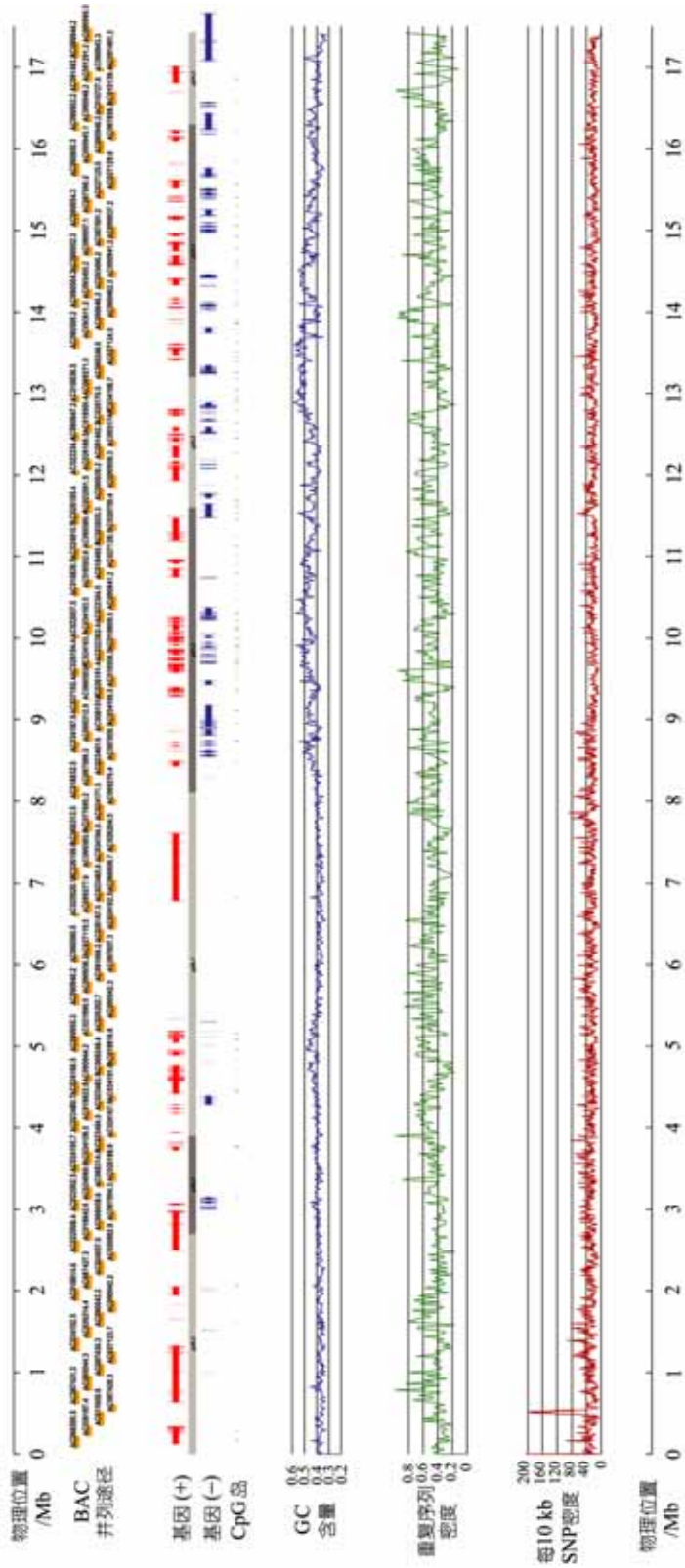

图 1　北京区域基因组概貌

表 1 北京区域的基因一览

基因分类	基因数目	基因长度/bp (%覆盖度)	CDS 长度/bp (%覆盖度)	基因平均长度/bp	外显子平均长度/bp	每个基因平均转录体数目	每个基因平均外显子数目
已知基因	122	7882429 (45.22)	524740 (3.01)	64610	240	1.57	10.97
预测基因	20	341089 (1.96)	26903 (0.15)	17054	359	1.05	3.55
蛋白编码基因	142	8223518 (47.18)	551643 (3.16)	57912	244	1.5	9.92
假基因	16	12046 (0.07)	11958 (0.07)	753	239	1	3.125
基因总数	158	8235564 (47.25)	563601 (3.23)	52124	244	1.45	9.23

区域所识别的27个非编码RNA基因中，有1个tRNA基因，6个信号识别粒子RNA(SRP_euk_arch)基因，7个核糖体RNA基因，9个小核仁RNA基因，其余的属于Y RNA和TPP/维生素B12核糖开关家族(cobalamin riboswitch families). 在北京区域没发现小RNA(miRNA)基因.

用InterProScan对北京区域编码的所有蛋白质进行注释，结果显示有70.3%的蛋白质被注释含有结构域(其中64.0%的蛋白质序列含有多个结构域). 出现最多的结构域是双向核定位信号(IPR001472)和免疫球蛋白样结构域(IPR007110). 基于InterProScan结果，利用基因本体数据库(Gene Ontology，简称GO)[24]的分类系统对基因进行功能分类. 59.2%的蛋白质编码基因能够依据GO进行分类，这其中有40.1%的基因属于分子功能类，38.3%的属于生物学过程类，21.6%的属于细胞成分类(图3).

2.3 北京区域基因组概貌

北京区域基因组的完成图概览如图1所示，平均GC含量为42%，同人类全基因组的平均GC含量基本一致，但GC含量区域特征明显，近端粒侧8.4 Mb区域的平均GC含量为38.5%，近着丝粒侧9 Mb区域的GC含量为45%. 重复序列家族SINEs, LINEs, LTR-反转座子和DNA转座子在北京区域的含量分别为13.97%，19.38%，8.67%和3.55%，相对应的人类全基因组中的含量分别为13.33%, 18.17%, 8.71%和3.53%. 北京区域SINEs，LINEs和DNA转座子的出现频率均高于人类基因组的平均水平(表2). 北京区域染色体带型(G带)包括从3p26.3到3p24.3(部分)的7个不同的带型(表3). 重复序列家族，GC含量和CpG岛在北京区域的分布符合已知的基因组趋势和特征，即G^-带GC含量高，CpG岛丰富，相应的基因密度也较高；G^+带GC含量低，CpG岛和基因密度均较低[25](表3). 基因组的另一趋势和特征，即高基因密度与SINE的含量相关，也在北京区域得到很好的验证.

重复序列突变率与它们存在于基因组中的年代正相关[1]. 我们将重复序列按不同的突变率归类，并和全基因组的结果比较. 在北京区域中，突变率大于16%的重复序列所占的比重比全基因组中的对应值高，特别是LINE2和DNA转座子. 北京区域中Alu的突变率峰值为11%，而人类基因组工作框架图的Alu的突变率的峰值仅为7%；LINE1的分布特征和全基因组较为相似，在4%的突变率处达到第一个峰值；LINE/L2和MIR在北京区域所占比例明显高于全基因组，而Alu和LINE1的含量则低于全基因组(图4)，因此推断北京区域与人类基因组其他区域相比，在进化上可能是较为古老的. 我们计算不同突变率的重复序列在北京区域的分布，选出含该突变率重复序列最多的区域为代表区域，发现它们与GC含量是正相关的(图5).

微卫星DNA(microsatellite)为短串联重复单位所构成的DNA序列总称，在人类基因组中出现频率很高[26]. 在北京区域共找到2090个不同的微卫星DNA，其中二核苷酸重复占29%，单核苷酸重复和四核苷酸重复分别占18.2%和18.8%. 二核苷酸重复中AC和AT含量较高，分别占59%和31%.

2.4 序列变异

来自于共同祖先的单体型是确定基于对不同人群单核苷酸多态性(SNP)位点的广泛收集[21]. 通常通过两种方法来发现SNP，即：(ⅰ) 比较不同单体型(不同个体或者同一个体的不同染色体)的装配后大片段的交叉序列；(ⅱ) 比较由不同个体组成的全基因

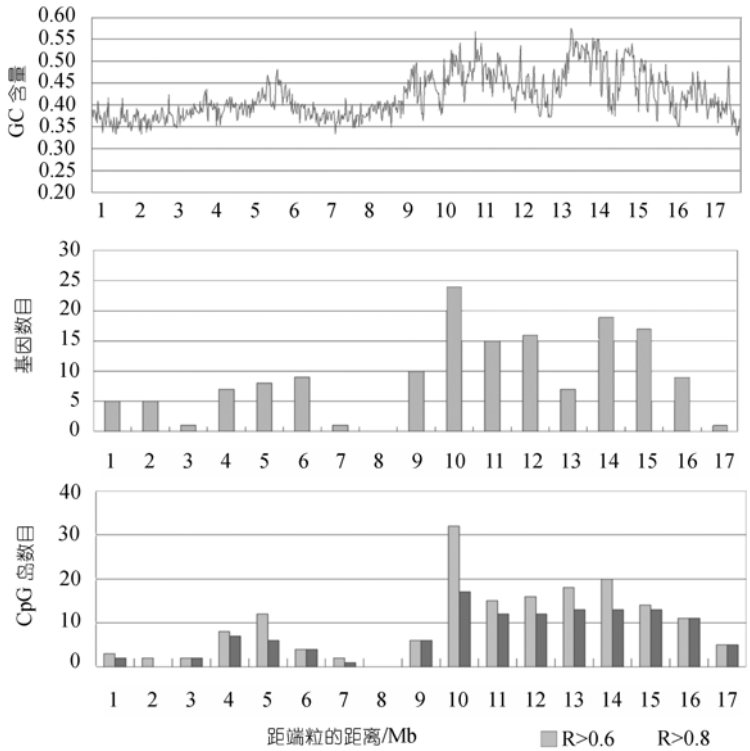

图 2　北京区域的基因分布及其与 GC 含量和 CpG 岛分布的相互关联

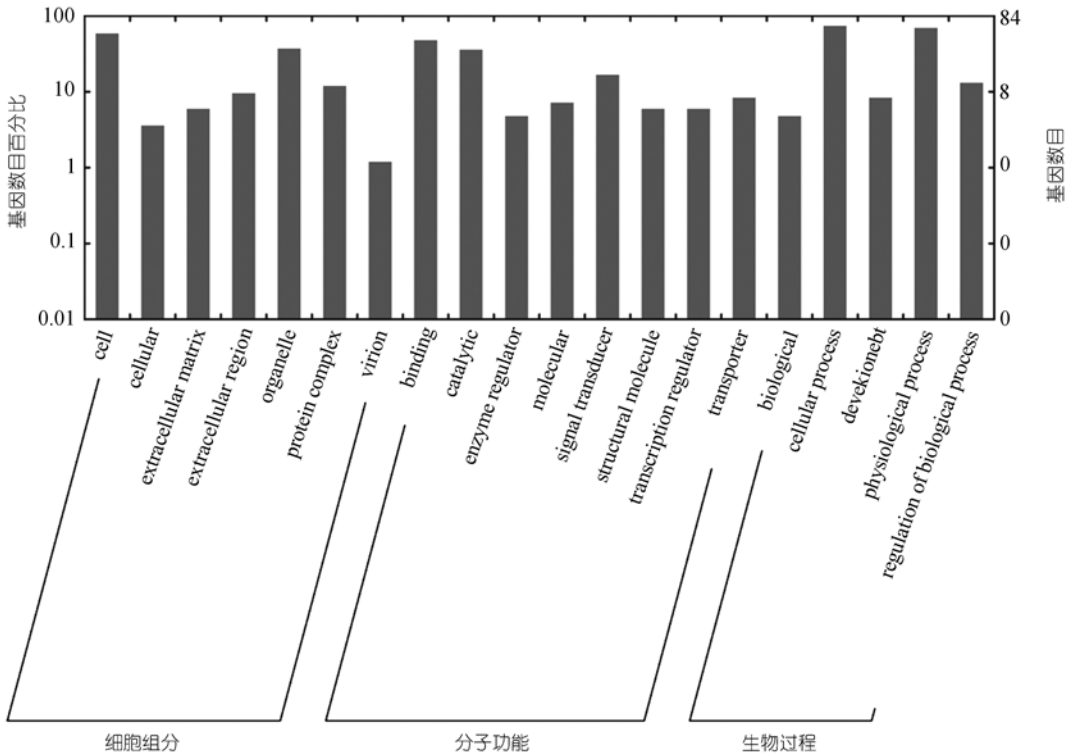

图 3　北京区域基因功能分类(GO)

表 2 散在重复序列在北京区域以及全基因组中的分布

重复序列家族	重复序列子类	北京区域			人类基因组		
		数目	序列长度/bp	百分比/%	数目(×100)	序列长度/Mb	百分比/%
SINE		11578	2434983	13.97	1718	383.4	12.5
	ALUs	6120	1639469	9.41	1152	302.3	9.85
	MIRs	5458	795514	4.56	567	81.1	2.64
LINE		5938	3378593	19.38	1345	588.5	19.18
	LINE1	3020	2611124	14.98	889	488.2	15.91
	LINE2	2527	676312	3.83	410	91.5	2.98
	L3/CR1	391	91157	0.52	46	8.8	0.29
LTR		3069	1511528	8.67	637	242	7.9
DNA		2588	618149	3.55	382	84.5	2.75
Total		23173	7943253	45.57	4083	1299	42.33

表 3 北京区域的细胞遗传图谱条带、GC 含量和基因密度之间的相互关系

细胞遗传图谱条带	G 带	位置(距端粒的距离/Mb)	GC 含量/%	基因数目	基因密度(基因数目/Mb)
3P26	3p26.3 (G^+)	1~2.7	36.93	11	4.1
	3p26.2 (G^-)	2.7~3.9	39.39	6	5
	3p26.1 (G^+)	3.9~8.1	39.23	19	4.5
3P25	3p25.3 (G^-)	8.1~11.6	45.51	53	15.1
	3p25.2 (G^+)	11.6~13.2	45.6	18	11.3
	3p25.1 (G^-)	13.2~16.3	45.65	46	14.8
3P24	3p24.3 (G^+)	16.3~17.4	37.83	5	4.5
Total		17.4	41.9	158	9.1

组鸟枪法文库测序得到的随机序列. 综合来自于dbSNP的已经被实验验证的 38475 个SNP信息, 在北京区域我们共识别 42607 个非冗余的SNP位点, 其中包括 4123 个我们通过序列比对分析得到的SNP. 目前北京区域的SNP密度为每 410 个碱基一个位点, 高于先前报道的人类基因组框架图的平均SNP密度(1/1300 bp)[1], 低于国际人类基因组单体型图计划(HapMap)的预测值 1/300 bp[27]. 图 6 展现了北京区域的SNP类型及其相对应的比例, 大部分SNP位于基因非编码区(UTR)或者位于蛋白质编码区但属于同义突变.

2.5 同源比对分析

哺乳动物的基因组结构被认为是非常保守的[22]. 随着小鼠、大鼠以及黑猩猩基因组框架图的完成, 使得开展物种间比较基因组学研究、基于比较结果进行基因预测和寻找高度保守区成为可能[20]. 借助于较长基因组序列的专业比对软件BLASTZ[28], 我们得到了北京区域与小鼠、大鼠和黑猩猩的同源区域(图 7).

3 讨论

3.1 序列图谱构建和缺口填补

北京区域的序列信息采用逐步克隆(clone-by-clone)测序法通过两个基本阶段而获得. 最初, 框架序列(draft assembly)由约 6X 覆盖度的鸟枪法 reads 拼接而成, 为了保证序列的高质量, 又继续测序达到 8X 覆盖度, 此时的序列准确率达到每 10000 个碱基少于一个错误. 在最终的北京区域完成图中, 实际的序列准确率已经超过了 99.99%.

采取逐步克隆法的基因组测序项目通常会遇到两种类型的缺口: 目标基因组区域由 BACs 所构成的克隆重叠群(clone contig)间的缺口和 BAC 内部质粒亚克隆间的缺口, 即序列缺口. 序列缺口由亚克隆质粒而不是 BAC 克隆序列本身作为桥梁进行缺口填补. 在北京区域框架图中, BAC 间的物理缺口共有 11 个, BAC 内部克隆间的序列缺口有几百个. 为了填补 BAC 间的缺口, 在北京区域的并列途径中共追加测

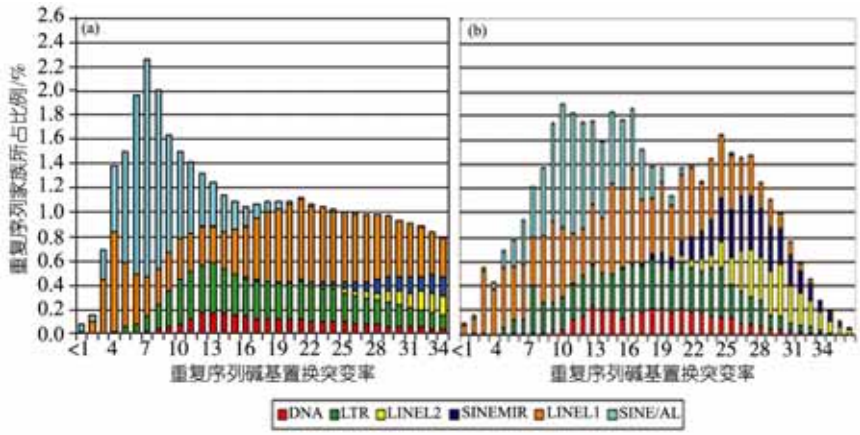

图 4　北京区域不同突变率的重复序列所占的比例

(a) 全基因组; (b) 北京区域

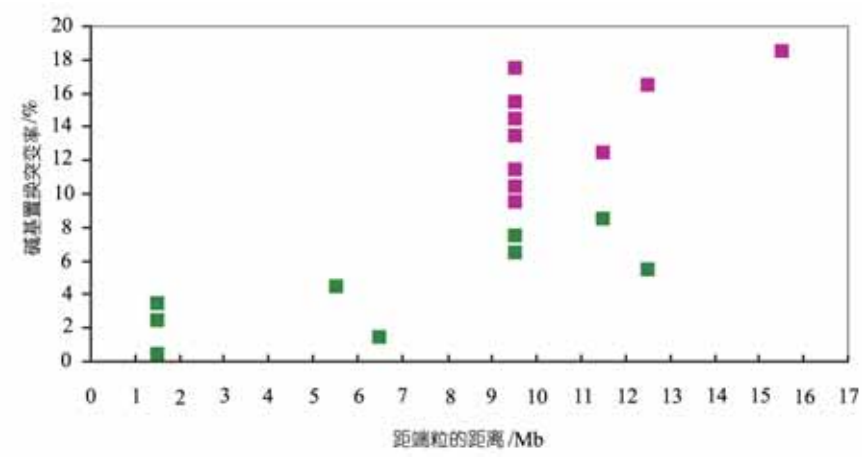

图 5　不同突变率的重复序列最高含量区域在北京区域的分布

低突变率(0~20%)以及进化中较新的重复序列被标出(不同的颜色代表不同的碱基突变率范围)

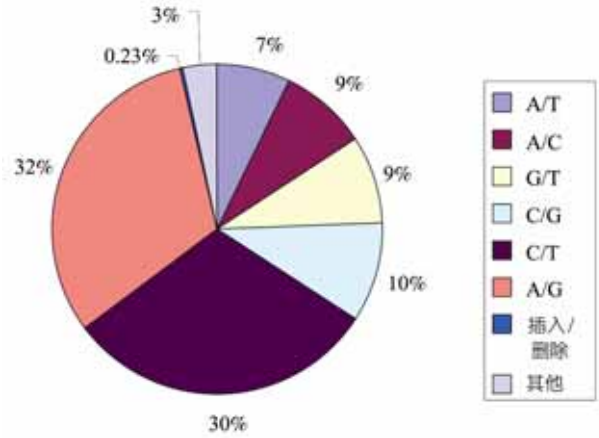

图 6　北京区域 SNP 类型

序了 39 个BAC; 测序共产生 220000 个克隆用以填补克隆内的序列缺口, 所采用的方法是据含有序列缺口的BAC克隆的末端序列设计引物, 直接以PCR反应进行筛选. 最后一个位于亚端粒区的物理缺口的填补付出了巨大的努力, 该区域大小最初估计为 100~150 kb. 根据Riethman等人[29]的方法, 我们构建了 3 个亚端粒区序列特异性的YAC, 包括 3p端粒区, 并进行了粘粒亚克隆测序. 在追加测序完成后, 仍然存在约 30 kb的缺口, 这个缺口是“技术允许缺口”, 如同HGP其他由于难以测序而遗留下来的缺口一样, 这些缺口将在IHGSC的资助下另行特殊测序和分析[30]. 北京区域完成图进一步被以实验和序列为基

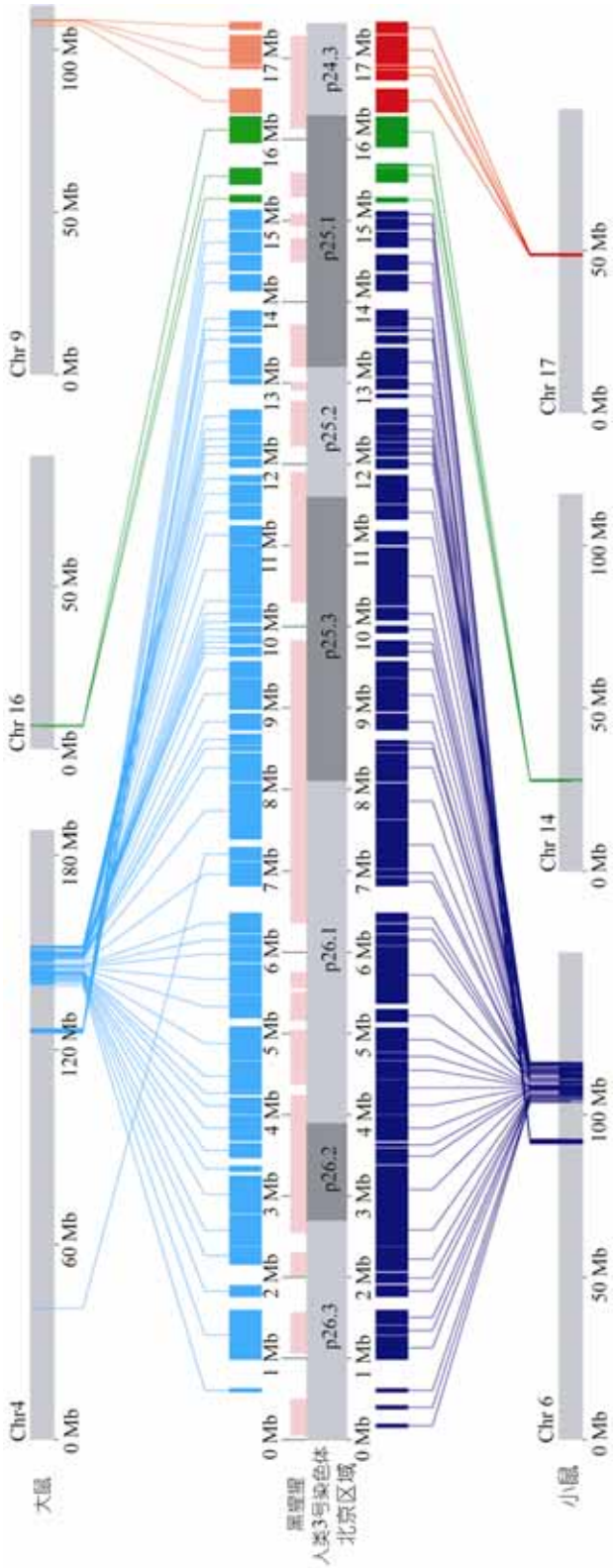

图 7 北京区域与小鼠、大鼠和黑猩猩同源板块区域示意图

础的限制性图谱(图 8)、遗传标记、全长 cDNA 和 EST 序列所验证和确认.

3.2 遗传重组率

遗传距离是同源染色体在有丝分裂时重组交换的频率, 两点间可观察到的重组的次数越多, 其遗传距离就越大. 遗传距离通过遗传图距单位——厘摩(centiMorgan, cM)来进行衡量, 1 厘摩相当于在一个世代中, 从一个遗传位点标记到第二个位点标记间发生交换的可能性是 1%. 遗传重组率(通常指 1 Mb 区域对应的遗传距离)在人基因组中分布并不均一[31,32](图 9), 依据Marshfield图谱[33], 人基因组平均遗传重组率为 1.36 cM/Mb, 最高遗传重组率为 17 号染色体 75~76 Mb区域的 7.60 cM/Mb. 而在Genethon图谱[34]中, 平均重组率是 1.30 cM/Mb, 最高值是X染色体的 22~23 Mb的 7.53 cM/Mb. 北京区域的平均遗传重组率几乎是全基因组平均水平的两倍, 基于Marshfield图谱和Genethon图谱的平均遗传重组率分别为 2.48 和 2.14 cM/Mb.

遗传重组率在染色体上的变化与染色体的位置密切相关. 一般而言, 染色体的近端粒区域的遗传重组率较高, 而染色体的中间部分重组频率较低; 大染色体近端粒区的遗传重组率高于小染色体近端粒区的遗传重组率[1]. 在北京区域 0~1 Mb的亚端粒区域内, Marshfield图谱的平均遗传重组率为 5.5 cM/Mb, 而Genethon图谱的平均重组率为 0 cM/Mb, 这缘于Marshfield图谱在端粒区域有较多的遗传标记, 从而能观察到更多的重组事件. 我们依据Genethon图谱和Marshfield图谱计算 3 号染色体的遗传重组率的分布(图 10), 发现北京区域作为 3 号染色体平均重组率最高的区域, 不仅高于 3 号染色体的中间区域, 而且还高于 3q的端粒区域. 3 号染色体的遗传重组率峰值为 6.1 cM/Mb, 出现在北京区域的 3~4 Mb部分. 对北京区域而言, 遗传重组率变动幅度很大, 2~3 Mb, 3~4 Mb和 8~9 Mb区域的平均重组率高达 4~6 cM/ Mb, 而 5~8 Mb和 10~17 Mb区域的重组率则较低.

依据Genethon图谱和Marshfield图谱, 北京区域的遗传重组率在全基因组范围内也是较高的, 一些亚区域的重组率高于 4 cM/Mb, 高于全基因组的平均水平(图 11). 造成遗传重组率差异的分子基础目前尚不明确. Gerton等人[35]根据他们对酵母*Saccharo- myces cerevisiae*减数分裂重组热点和非热点区域的作图发现: 遗传重组的相对发生率与染色体的GC丰富区相关. 然而, 北京区域的端粒侧的GC含量较低, 使得年轻的重复序列倾向于插入该区域[1], 年轻重复序列之间分歧程度小, 从而可能导致同源重组事件的增加.

3.3 北京区域的基因鉴定与大基因

(1) 基因鉴定: 目前基因鉴定有赖于 3 种基本方法: 序列相似性搜索、*ab initio*基因预测和实验数据支持. 第一种方法假定许多基本的生理生化过程在所有物种中都相当保守, 因此同源基因应该具有相似的功能, 特别是那些从相近物种进化而来的基因[24]. *Ab initio*基因预测软件从序列自身出发, 依其原理不同而分别侧重于序列的信号项(指短的序列模体)和内容项(指线性或种属特异性的密码子使用模式)之间的序列信息[36]. 两种最常用的基因预测软件——GenScan[37](强调内容项)和FgeneSH[38](强调信号项), 结合GeneWise[39]软件共同用于北京区域的注释. GeneWise将蛋白质序列相似性信息和*ab initio*基因发现算法进行整合, 以提高预测的准确性. 基因预测软件并不完全可靠, 它需要实验证据的支持, 而这些证据并不局限于基因功能的生物化学性质, 还包括已有的cDNA和/或EST序列. 在我们的分析中, 一条预测基因的核酸序列必须与其实验获得的转录体(cDNA或EST)达到 97%以上的同源性, 90%以上的覆盖度, 并且确认有两个以上的预测外显子, 才能被鉴定为一个新基因. 还有其他一些措施以保证基因预测的准确性, 包括排除由于序列污染而产生的虚假匹配、考虑 5′和 3′UTR区域中的转座子、以及排除与序列片段或假基因的比对结果等[39]. 读码框小于 50 个氨基酸的cDNA序列也将被排除. 最后, 我们对基因预测结果进行人工评价, 包括去除一些外显子预测片段、根据EST序列比对结果调整外显子边界、以及利用其他支持信息, 如CpG岛、GC含量和基因组比较等, 对基因进行验证.

(2) 大基因: 人类基因组的基因数目依据预测方法的不同而变化, 最初基于基因组序列草图估计的

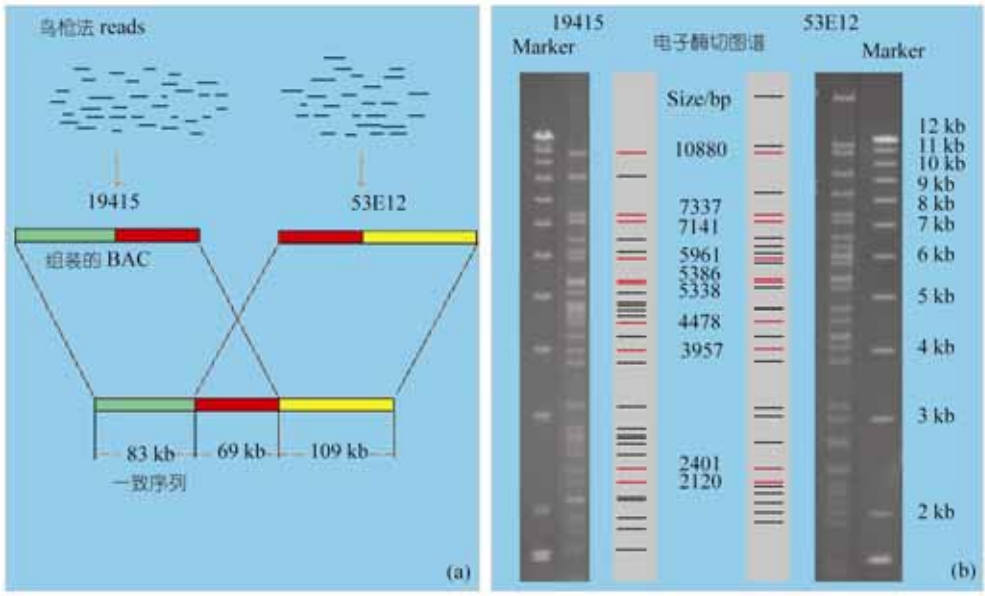

图 8 序列组装和验证示意图

依据序列间的重叠将组装后的 BAC 序列(RP11-19415 和 RP11-53E12)合并为一个一致序列(a), 一致序列可以通过对比基于实验数据的限制酶切图谱和计算机构造的电子酶切图谱得到验证(b)

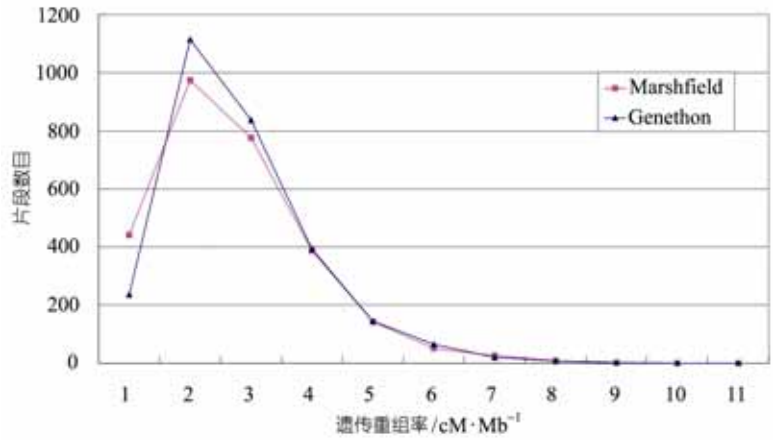

图 9 北京区域遗传重组率的分布(按 1 Mb 窗口计算)

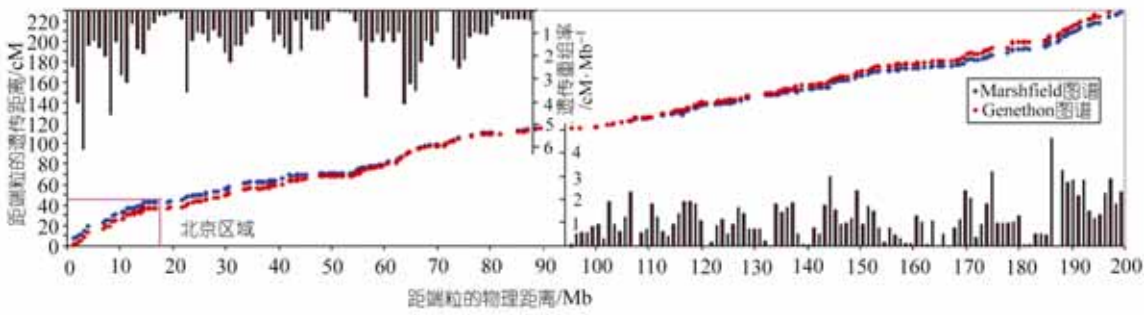

图 10 3 号染色体的遗传距离(基于 Marshfield 图谱和 Genethon 图谱)和物理距离分布

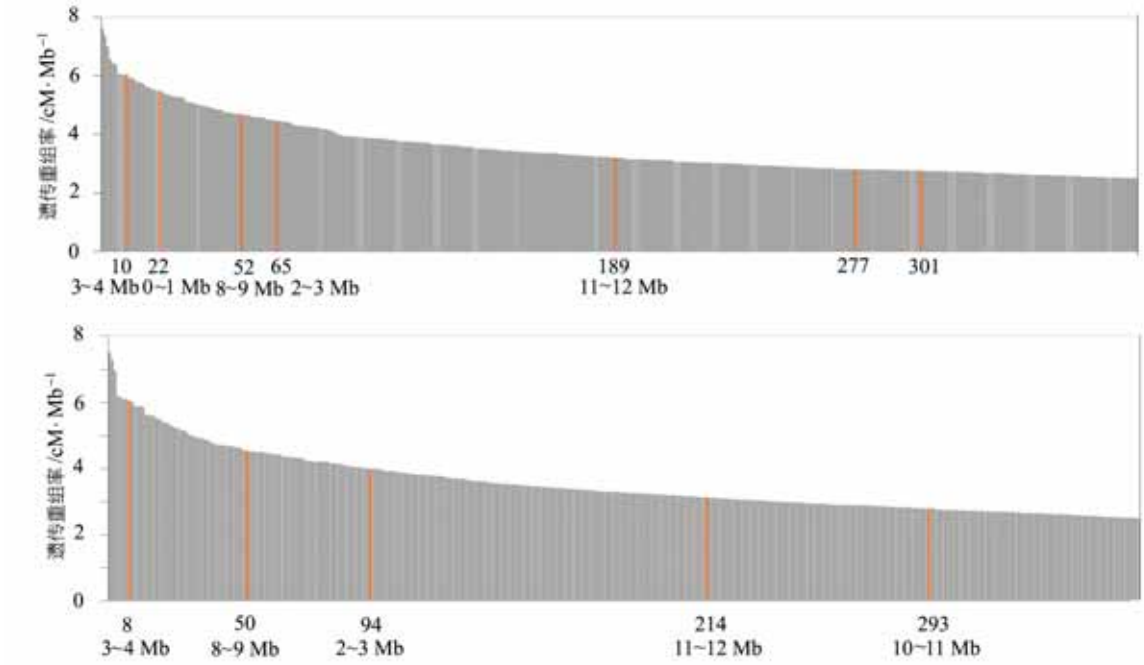

图 11　北京区域的高重组率区域在全基因组范围的排名, 分别基于 Genethon 图谱(上图)和 Marshfield 图谱(下图). 在 Marshfield 和 Genethon 图谱中分别有 4 个和 3 个区域的重组率大于 4 cM/Mb. 北京区域的亚区域的排名用橙色标出

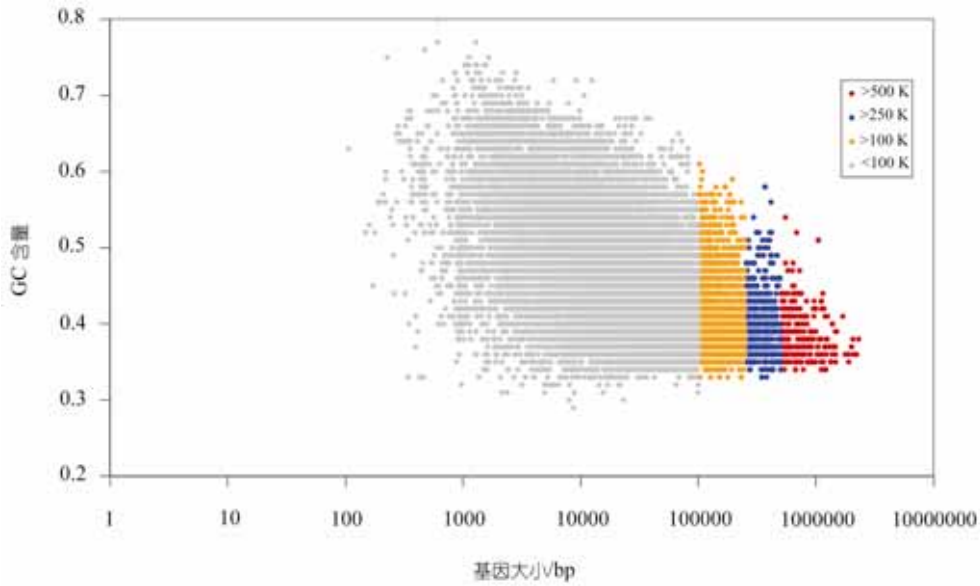

图 12　人类基因组中的 GC 含量与基因长度间的函数关系

基因总数为 3~4 万个[1], 目前基于序列完成图的估计为 2~2.5 万个[40]. 我们对UCSC的人类基因数据集 21869 个基因(http://www.ncbi.nlm.nih.gov, hg17, May 2004)进行统计分析, 结果表明人类基因的平均长度约为 58 kb. 如果以此作为人类基因的平均长度, 那只有 51%左右的基因组(58 kb × 25000 = 1450 Mb)是

转录的基因区域. 因此, 基因数目的差异不仅仅是由于保守的的基因鉴定流程所导致, 同时也是由于对人类基因平均长度的低估所造成的[41].

毫无疑问, 大基因确实存在于人类基因组中, 并且在哺乳动物基因组中是保守的. 一个最极端的例子是人类X染色体上长达 2.3 Mb的dystrophin基因[41]. 尽管大基因的数目在基因组中所占的比例很小, 如长度大于 100 kb, 大于 250 kb和大于 500 kb的基因数目分别占基因总数的 14.9%, 4.3%和 1.2%, 但是它们却分别占基因总长度的 62.4%, 35.1%和 16.8%. 此外, 对 21869 个已知人类基因的分析表明, 随着基因长度的增加, 基因倾向于占据AT含量高的区域, 高AT区域是大多数大基因存在的区域(图 12), 因此, 我们推断在人类基因组中有数百个大基因尚未发现, 特别是在所谓的“基因沙漠”区域中[41,42].

现有基因预测方法在大基因识别方面表现不佳, 原因之一, 大基因给以计算机为基础的基因预测方法带来统计方面的问题, 因为它们的内含子可长达数十万碱基[42], 而基因预测软件(如GenScan、FgeneSH)的模型和参数都来自对中小基因的总结, 因而缺少大基因的特征性参数资料[36,41]. 以往的分析显示GeneScan对于小基因的结果很好而对超过 100 kb的基因效果就相当不好了[41]. 原因之二, 大基因很难大量表达, 而且一般存在于分化组织和慢周期细胞中, 比如肌肉和大脑中, 因此在cDNA和EST数据库中很少有它们的表达序列信息. 其三, 利用一个较小的脊椎动物的基因组, 如: 河豚, 它与人类基因组具有相似的基因数目, 然而其基因组却只有人类基因组长度的 1/7[43]来分析大基因的尝试也不如预期的那样理想, 主要是由于河豚基因与人类基因的同源性很差, 而且缺乏大范围易于识别的同线性信息. 此外, 由于其他哺乳动物基因组与人类基因组高度相似, 使得几乎不可能从长为数十万碱基的内含子中鉴别出长度只有几百碱基的外显子, 除非对每一个大基因都从序列水平上进行仔细和彻底的检查.

北京区域的 122 个已知基因覆盖了 45.2%的基因组区域, 这个数字在全基因组的平均基因覆盖度范围内. 北京区域的“基因沙漠区”还有待于进一步探索以发现可能存在的大基因, 特别是基因 *CNTN6* 和 *CNTN4* 之间的 1.1 Mb 基因空白区. *CNTN4* 和 *CNTN6* 分别为长度 487 kb 和 311 kb 的大基因, 比较基因组分析已经在小鼠、大鼠和黑猩猩中找到了它们的同源基因, 这些基因在上述物种中具有相同的基因排列顺序和几乎相同的基因间距离(分别为 1.8, 1.1 和 1.2 Mb). 在人类基因组中, 这个区域存在 3 个剪切 EST 证据(AW858653, AW938304 和 AU122382); 用软件 FgeneSH 也在有剪切 EST 证据支持的基因组区域识别出一个长为 111 kb 的全长开放阅读框(Open Reading Frame, 简称 ORF). 目前在小鼠和黑猩猩 *CNTN6* 和 *CNTN4* 的基因间区域还没有发现任何基因, 但在大鼠基因组中可以预测到 3 个平均长度为 109 kb 的基因. GC 含量的计算结果显示这个 1.1 Mb 的基因沙漠区是富含 AT 的(37%), 因此, 我们推断在这个片段中可能存在一个或多个大基因.

3.4 生物学意义

(1) 基因功能: 北京区域含有 142 个蛋白编码基因(包括 122 个已知基因和 20 个新基因), 其中有 100 个(70.3%)编码蛋白含有已知的蛋白结构域或者属于已知的蛋白家族. 一个蛋白可能含有多个功能结构域, 参与复杂的生物过程和途径, 如基因 *RAF1* 参与调节肌动蛋白、病灶粘结和 MAPK 信号传导等三个途径.

我们依据基因本体(Gene Ontology, 简称GO)[24]对基因进行功能分类, 结果显示这些基因编码的蛋白质相当大一部分参与了信号传导、新陈代谢、蛋白质转录、翻译和转运等过程. 一些蛋白, 如*CHL1*, *IL5RA*, *ITPR1*, *ARL10C*, *GRM7*, *OXTR*, *CAMK1*, *IRAK2*, *Q86YP8*, *HRH1*, *PPARG*, *RAF1*, *WNT7A*, *SH3BP5*, *PLCL2* 和*OXTR*, 据报道在信号传导途径中发挥作用; 参与DNA转录和翻译的有: *BHLHB2*, *BRPF1*, *TADA3L*, *VHL*, *VGLL4*, *PPARG*, *HDAC11* 和 *NR2C2* 等蛋白; 各类转运蛋白包括: *TRNT1*, *ITPR1*, *SEC13L1*, *ATP2B2*, *SLC6A11*, *SLC6A1*, *SYN2*, *SLC6A6* 和*ZFYVE20*.

在脊椎动物进化过程中, 神经系统相关的基因大量增加[44]. 在北京区域存在相当多的神经系统相关基因, 如基因*GRM7* 编码L-谷氨酸受体, 是北京区

域中最大的基因(842 kb)[45]. *mGluR7*是其家族8个亚型中最保守的亚型, 在脑组织中广泛分布, 特别在脑皮质、海马和小脑等组织中表达水平较高[46,47]. *mGluR7* 在正常中枢神经系统中行使功能, 并与脑损伤的病理过程以及其他神经退行性疾病相关[48]. 北京区域其他与神经系统相关的基因还包括*SLC6A1*, *SLC6A6*, *SLC6A11*, *CHL1*, *CNTN4*和*CNTN6*, 分别属于钠/氯离子依赖的转运体家族和神经细胞结合蛋白家族. 钠/氯离子依赖的转运体家族是一个有 21 个成员的大家族, 负责神经递质的转运. *SLC6A1* 和 *SLC6A11* 都编码*GABA*转运体, 它们转录方向相同, 基因间距离为 55 kb, 所编码蛋白之间的相似性为54%, 推测它们可能是由一个早期的片段倍增所产生[49,50]. *SLC6A6* (氨基乙磺酸转运蛋白)位于 3p25.1 (G^-带), 与*SLC6A1*相距 3.4 Mb[18]. *CHL1*, *CNTN6*和 *CNTN4* 属于神经细胞结合蛋白家族, 基因长度均大于 200 kb, 转录方向相同, 且全部位于 3p26.3(G^+带)(表 4).

(2) 疾病相关基因: 据报道 3 号染色体短臂(3p)与多种肿瘤相关. 在多种肿瘤发生早期的原位组织或细胞系中存在 3p非随机、高频的杂合性丢失, 暗示这些区域可能存在抑癌基因(tumor suppressor gene, TSG). 定位在 3p25.3 的*VHL* 是已被证实的抑癌基因[51,52], 而且是导致Von Hippel-Lindau 综合症的致病基因. Von Hippel-Lindau 综合症是一种显性遗传的综合症, 表现为易患视网膜瘤、肾细胞癌和嗜铬细胞瘤等肿瘤[53]. *VHL*在多种不同的细胞中表达, 具有多种不同的功能, 目前已知的功能有: 调控缺氧诱导基因的表达, 对p27 稳定性的调控, 调控fibronectin的装配等[54~60].

除*VHL*外, 还有其他 9 个疾病基因被定位于北京区域: *FANCD2*, *CAV3*, *COLQ*, *BTD*, *ITPR1*, *OGG1*, *ATP2B2*, *RAF1* 和*XPC*. 范可尼贫血(FA)是一种常见的隐性遗传病, 8 个不同的基因都可导致范可尼贫血, *FANCD2* 是其中的一种. 8 种基因中仅有*FANCD2* 能够在果蝇和线虫中找到同源基因[61,62]. Limb-girdle muscular dystrophy (LGMD) 指一类在儿童期或成年期发病的肌营养不良, 它和杜氏肌营养不良显著不同. 患有LGMD 的病人通常表现为四肢肌肉无力和消瘦. LGMD 主要以常染色体隐性方式传递, 少数几种以常染色体显性方式传递. *CAV3* 的突变 和 1C型LGMD相关, *CAV3* 参与肌细胞膜的构建和跨膜信号的传导[63].

通过检索在线人类孟德尔遗传数据库OMIM (http://www.ncbi.nlm.nih..gov/Omim/), 我们在北京区域又发现了 17 个遗传疾病相关基因, 这些基因的突变可能导致机体对某些疾病的易感性或影响疾病的发展进程. 例如: *TIMP4* 是*TIMP*家族成员之一, 编码组织金属蛋白酶抑制剂. 在肿瘤发展过程中, 金属蛋白酶降解细胞外基质, 促进肿瘤细胞的侵袭和转移, 而*TIMP4* 则能够抑制金属蛋白酶活性. 体外实验证实*TIMP4* 的过量表达将显著抑制肿瘤细胞的侵袭能力[64].

5 种疾病通过细胞遗传学的方法被定位于北京区域(表 5), 我们可以通过定位候选克隆来获得这些疾病相关基因并进行基因突变分析[65]. 3p25 是心房隔膜缺陷(Atrioventricular septal defect 2, AVSD2)和Pancreatic endocrine肿瘤的致病区域. 3p 综合症是另一由于 3 号染色体短臂(3p25-3pter)末端片段缺失而引起的复杂疾病[66]. 由于包含多个基因的缺失, 3p综合症的临床表型较为复杂, 主要包括生长迟缓、智力低下、特殊颅面特征(小头, 小鄂、上睑下垂)和心脏隔膜缺陷. 我们可通过“表型解剖”策略, 将不同的表型定位于不同的染色体区段, 再定位候选克隆

表 4　北京区域中的神经系统相关基因

蛋白家族	基因符号	正负链	基因大小/bp	起始位置/bp	终止位置/bp
Neural cell adhesion	CHL1	+	212449	213650	426098
	CNTN6	+	310650	1109629	1420278
	CNTN4	+	486557	2588089	3074645
Sodium- and chloride-dependent transporter	SLC6A11	+	122230	10832917	10955146
	SLC6A1	+	46490	11009440	11055929
	SLC6A6	+	84,453	14419150	14503602

表 5 经细胞遗传学方法定位于北京区域，但未在分子水平识别出来的疾病列表

疾病名称	OMIM 编号	细胞遗传图谱位置	异常
MYMY	252350	3p26-24.2	Moyamoya disease 2
ST11	602011	3p25	Pancreatic endocrine tumors
AVSD2	606217	3p25.3	Atrioventricular septal defect, susceptibility to, 2
CMD1E	601154	3p25-p22	Cardiomyopathy, dilated, 1E
MFS2	154705	3p25-24.2	Marfan-like connective tissue disorder

来获得对应表型的基因[67]. 这种“表型解剖”策略，已在许多疾病相关基因研究中被成功地应用[68,69].

人类基因组序列的完成仅仅标志着我们了解人类自身生物学机制的长期努力的开始，这些努力将包括遗传变异研究、与人类疾病和健康相关的基因功能研究、以及新基因的识别和实验验证. CHGSC 是人类基因组单体型图计划(Human HapMap Project)的重要参与者，并将与世界范围的各个基因组研究机构紧密合作，共同致力于大规模科学研究项目，以促进与基因组研究的不断进展和应用.

致谢 感谢北京市政府、浙江省政府和杭州市政府给予本项目的财政支持；感谢乐清市政府和人民在项目初始阶段所给予的大力支持；感谢国内外的各位顾问、合作者和朋友；感谢广大民众的热情支持，同时也感谢人类基因组 3 号染色体测序协作组和国际人类基因组测序协作组的大力支持.

参 考 文 献

1 The International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature, 2001, 409: 860~921[DOI]
2 Ewing B, Green P. Base-calling of automated sequencer traces using phred. Ⅱ. Error probabilities. Genome Res, 1998, 8: 186~194
3 Ewing B, Hillier L M, Wendl C, et al. Base-calling of automated sequencer traces using phred. Ⅰ. Accuracy assessment. Genome Res, 1998, 8: 175~185
4 Gordon D, Abajian C, Green P. Consed: A graphical tool for sequence finishing. Genome Res, 1998, 8: 195~202
5 Benson G. Tandem repeats finder: A program to analyze DNA sequences. Nucleic Acids Res, 1999, 27: 573~580[DOI]
6 Pruitt K D, Maglott D R. RefSeq and LocusLink: NCBI gene- centered resources. Nucleic Acids Res, 2001, 29: 137~140[DOI]
7 Imanishi T, Itoh T, Suzuki Y, et al. Integrative annotation of 21037 human genes validated by full-length cDNA clones. PLoS Bio, 2004, 2(6): e162[DOI]
8 Kent W J. BLAT-The BLAST-like alignment tool. Genome Res, 2002, 12: 656~664[DOI]
9 Salamov A A, Solovyev V V. *Ab initio* gene finding in *Drosophila* genomic DNA. Genome Res, 2000, 10: 516~522[DOI]
10 Boeckmann B, Bairoch A, Apweiler R, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res, 2003, 31: 365~370[DOI]
11 Zdobnov E M, Apweiler Rolf. InterProScan——An integration platform for the signature-recognition methods in InterPro. Bioinformatics, 2001, 17: 847~848[DOI]
12 Lowe T M, Eddy S R. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res, 1997, 25: 955~964[DOI]
13 Griffiths-Jones S, Bateman A, Marshall M, et al. An RNA family database. Nucleic Acids Res, 2003, 31: 439~441[DOI]
14 Griffiths-Jones S. The microRNA Registry. Nucleic Acids Res, 2004, 32: 109~111[DOI]
15 Ambros V, Bartel B, Bartel D P. A uniform system for microRNA annotation. RNA, 2003, 9: 277~279[DOI]
16 Schwartz S, Kent W J, Smit A, et al. Human-mouse alignments with BLASTZ. Genome Res, 2003, 13: 103~107[DOI]
17 Felsenfeld A, Peterson J, Schloss J, et al. Assessing the quality of the DNA sequence from the Human Genome Project. Genome Res, 1999, 9: 1~4
18 Patel A, Rochelle J M, Jones J M, et al. Mapping of the taurine transporter gene to mouse chromosome 6 and to the short arm of human chromosome 3. Genomics, 1995, 25: 314~317[DOI]
19 Heilig R, Eckenberg R, Petit J L, et al. The DNA sequence and analysis of human chromosome 14. Nature, 2003, 421: 601~607[DOI]
20 Woolfe A, Goodson M, Goode D K, et al. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol, 2005, 3: e7[DOI]
21 Deloukas P, Matthews L H, Ashurst J, et al. The DNA sequence and comparative analysis of human chromosome 20. Nature, 2001, 414: 865~871[DOI]
22 Dunham I, Shimizu N, Roe B A, et al. The DNA sequence of human chromosome 22. Nature, 1999, 402: 489~495[DOI]
23 Mironov A A, Fickett J W, Gelfand M S. Frequent alternative splicing of human genes. Genome Res, 1999, 9: 1288~1293[DOI]
24 Ashburner M, Ball C A, Blake J A, et al, Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 2000, 25: 25~29[DOI]
25 Gardiner K. Human genome organization. Curr Opin Genet, 1995, 5: 315~322[DOI]

26 Kruglyak S, Durrett R T, Schug M D, et al. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc Natl Acad Sci USA, 1998, 95: 10774~10778[DOI]

27 The International HapMap Consortium. The international HapMap project. Nature, 2003, 426: 789~796[DOI]

28 Schwartz S, Kent W J, Smit A, et al. Human-mouse alignments with BLASTZ. Genome Res, 2003, 13: 103~107[DOI]

29 Riethman H C, Moyzis R K, Meyne J, et al. Cloning human telomeric DNA fragments into *Saccharomyces cerevisiae* using a yeast-artificial-chromosome vector. Proc Natl Acad Sci USA, 1989, 86: 6240~6244

30 Riethman H, Ambrosini A, Castaneda C, et al. Mapping and initial analysis of human subtelomeric sequence assemblies. Genome Res, 2004, 14: 18~28[DOI]

31 Purrello M, Alhadeff B, Whittington E, et al. Comparison of cytologic and genetic distances between long arm subtelomeric markers of human autosome 14 suggests uneven distribution of crossing-over. Cytogenet Cell Genet, 1987, 44: 32~40

32 Mohrenweiser H W, Tsujimoto S, Gordon L, et al. Regions of sex-specific hypo- and hyper-recombination identified through integration of 180 genetic markers into the metric physical map of human chromosome 19. Genomics, 1998, 472: 153~162[DOI]

33 Broman K W, Murray J C, Sheffield V C, et al. Comprehensive human genetic maps: Individual and sex-specific variation in recombination. Am J Hum Genet, 1998, 63: 861~869[DOI]

34 Dib C, Faure S, Fizames C, et al. A comprehensive genetic map of the human genome based on 5264 microsatellites. Nature, 1996, 380: 152~154[DOI]

35 Gerton J L, DeRisi J, Shroff R, et al. Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. Proc Natl Acad Sci USA, 2000, 97: 11383~11390

36 Wang J, Li S T, Zhang Y, Zheng H K, et al. Vertebrate gene predictions and the problem of large genes. Nature, 2003, 4: 741~749

37 Burge C, Karlin S, Prediction of complete gene structures in human genomic DNA. J Mol Biol, 1997, 268: 78~94[DOI]

38 Birn E, Clamp M, Durbin R. GeneWise and Genomewise. Genome Res, 2004, 14: 988~995

39 Fortna A, Gardiner K. Genomic sequence analysis tools: A user's guide. Trends Genet, 2001, 17: 158~164[DOI]

40 International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature, 2004, 431: 931~945[DOI]

41 Wong G K, Passey D A, Yu J. Most of the human genome is transcribed. Genome Res, 2001, 11: 1975~1977[DOI]

42 Wong G K, Passey D A, Huang, Y. et al., Is "junk" DNA mostly intron DNA? Genome Res, 2000, 10: 1672~1678[DOI]

43 Venkatesh B, Gilligan P, Brenner S. Fugu: A compact vertebrate reference genome. FEBS Lett, 2000, 476: 3~7[DOI]

44 Venter J C, Adams M D, Myers E W, et al. The sequence of the human genome. Science, 2001, 291: 1304~1351[DOI]

45 Wu Q, Dong W, Qi X. Genomic Structure of Metabotropic glutamate receptor 7 and comparison of genomic structures of extracellular domains of mGluR family. Chinese Science Bulletin, 2002, 47: 1330~1336[Abstract] [PDF]

46 Makoff A, Pilling C, Harrington K, et al. Human metabotropic glutamate receptor type 7: Molecular cloning and mRNA distribution in the CNS. Brain Res Mol Brain Res, 1996, 40: 165~170

47 Phillips T, Makoff A, Murrison E, et al. Immunohistochemical localisation of mGluR7 protein in the rodent and human cerebellar cortex using subtype specific antibodies. Brain Res Mol Brain Res, 1998, 57: 132~141[DOI]

48 Monaghan D T, Bridges R J, Cotman C W. The excitatory amino acid receptors: Their classes, pharmacology, and distinct properties in the function of the central nervous system. Annu Rev Pharmacol Toxicol, 1989, 29: 365~402[DOI]

49 Huang F, Shi L J, Heng H H, et al. Assignment of the human GABA transporter gene (GABATHG) locus to chromosome 3p24-p25. Genomics, 1995, 29: 302~304[DOI]

50 Borden L A, Dhar T G, Smith K E, et al. Cloning of the human homologue of the GABA transporter GAT-3 and identification of a novel inhibitor with selectivity for this site. Receptors Channels, 1994, 2: 207~213

51 Huebner K. Tumor suppressors on 3p: A neoclassic quartet. Proc Natl Acad Sci USA, 2001, 98: 14763~14765[DOI]

52 Kondo K, William G, Kaelin J R. The von Hippel-Lindau tumor suppressor gene. Exp Cell Res, 2001, 264: 117~125[DOI]

53 Friedrich C A. Von Hippel-Lindau syndrome. A pleomorphic condition. Cancer, 1999, 86: 2478~2482[DOI]

54 Duan D R, Pause A, Burgess W H, et al. Inhibition of transcription elongation by the VHL tumor suppressor protein. Science, 1995, 269: 1402~1406

55 Kibel A, Iliopoulos O, DeCaprio J A, et al. Binding of the von Hippel-Lindau tumor suppressor protein to Elongin B and C. Science, 1995, 269: 1444~1446

56 Gnarra J R, Zhou S, Merrill M J, et al. Post-transcriptional regulation of vascular endothelial growth factor mRNA by the product of the VHL tumor suppressor gene. Proc Natl Acad Sci USA, 1996, 93: 10589~10594[DOI]

57 Iliopoulos O, Levy A P, Jiang C, et al. Negative regulation of hypoxia-inducible genes by the von Hippel-Lindau protein. Proc Natl Acad Sci USA, 1996, 93: 10595~10599[DOI]

58 Pause A, Lee S, Lonergan K M, et al. The von Hippel-Lindau tumor suppressor gene is required for cell cycle exit upon serum withdrawal. Proc Natl Acad Sci USA, 1998, 95: 993~998[DOI]

59 Ohh M, Yauch R L, Lonergan K M, et al. The von Hippel-Lindau tumor suppressor protein is required for proper assembly of an extracellular fibronectin matrix. Mol Cell, 1998, 1: 959~968[DOI]

60 Kamura T, Koepp D M, Conrad M N, et al. Rbx1, a component of the VHL tumor suppressor complex and SCF ubiquitin ligase.

Science, 1999, 284: 657~661[DOI]
61 Timmers C, Taniguchi T, Hejna J, et al. Positional cloning of a novel Fanconi anemia gene, *FANCD2*. Mol Cell, 2001, 7: 241~248
62 Garcia-Higuera I, Taniguchi T, Ganesan S, et al. Interaction of the Fanconi anemia proteins and BRCA1 in a common pathway. Mol Cell, 2001, 7: 249~262[DOI]
63 Galbiati F, Volonte D, Minetti C, et al. Phenotypic behavior of caveolin-3 mutations that cause autosomal dominant limb girdle muscular dystrophy (LGMD-1C). Retention of LGMD-1C caveolin-3 mutants within the golgi complex. J Biol Chem, 1999, 274: 25632~25641[DOI]
64 Wang M, Liu Y E, Greene J, et al. Inhibition of tumor growth and metastasis of human breast cancer cells transfected with tissue inhibitor of metalloproteinase 4. Oncogene, 1997, 14: 2767~2774[DOI]
65 Collins F S. Positional cloning moves from perditional to traditional. Nat Genet, 1995, 9: 347~350[DOI]
66 Drumheller T, McGillivray B C, Behrner D, et al. Precise localisation of 3p25 breakpoints in four patients with the 3p- syndrome. J Med Genet, 1996, 33: 842~847
67 Angeloni D, Lindor N M, Pack S, et al. CALL gene is haploinsufficient in a 3p- syndrome patient. Am J Med Genet, 1999, 86: 482~485[DOI]
68 Dan M, Niebuhr A, Zhang X X. Critical region for cri-du-chat syndrome is located in the 5p15.2, a region flanked by DNA markers D5S713 and D5S18. Chinese Journal of Medical Genetics. 1997, 14: 263~267
69 Church D M, Bengtsson U, Nielsen K V, et al. Molecular definition of deletions of different segments of distal 5p that result in distinct phenotypic features. Am J Hum Genet, 1995, 56: 1162~1172

中国人类基因组测序协作组

分析和写作组:

王晶[1*], 张猛[1*], 吴清发[1*], 黄薇[2*], 沈岩[3*], 于军[1‡], 强伯勤[3‡], 陈竺[2‡], 杨焕明[1‡]

序列组装和完成图组:

白宝砚[4], 包维东[1], 卜东波[5], 陈斌[2], 陈冲[1], 陈丽虹[3], 陈润生[1,4,5], 陈竺[2], 从丽娟[1], 丁克越[3], 董伟[1*], 冯涛[1], 傅刚[2], 顾文漪[2], 何行[4], 胡松年[1*], 黄薇[2], 贾佳[2], 李松岗[1,7], 刘斌[1], 刘波勇[1], 刘昌宁[5], 刘涛[1,4], 吕纲[2], 牛宇欣[1], 齐晓华[1], 任双喜[2], 荣利[3], 沈岩[3], 盛海辉[2], 宋丽萍[1], 孙式微[5], 谈昕煜[3], 屠越峰[2], 王会峰[1], 汪建[1‡], 王晶[1], 王金华[3], 王俊[1*], 王亚妮[1], 黄家树[1,14], 吴清发[1], 熊慧[2], 杨焕明[1], 姚志建[3], 杨晓楠[1], 于军[1], 张力[1], 张祥林[2], 张以琳[1], 赵翼[5], 朱根凤[2], 朱晓鹏[4]

测序组:

包静月[1,8], 包其郁[1,8], 包维东[1], 毕世华[1], 边雪萌[1], 劳而斯伯伦[1,6], 蔡伦[5], 蔡天景[2], 蔡震[2], 曹婷[1], 曹玉竹[1], 陈宝献[1], 陈冲[1], 陈聪[1], 陈建龙[1], 陈洁[2], 陈俊宝[1], 陈丽霞[3], 陈殊[9], 陈同[2,10], 陈晓芸[2], 陈欣奕[9], 陈妍炯[1,11], 陈宜瑜[1], 陈竺[2], 程志华[12], 池澄[9], 崔红娟[1], 崔锦辉[1], 崔鹏[1,13], 丛彬[13], 戴理[1], 邓伟[4], 邓亚军[1,11], 丁浩[1], 董慧[12], 董伟[1], 董晓佳[3], 杜玉涛[1,13], 范红媛[1], 范克科[9], 方建秋[1], 冯国栋[1], 冯海燕[1], 冯洁[1], 冯小黎[1], 傅刚[2], 傅秋实[9], 高安武[1], 高继梅[1], 高泉[3], 高扬[1], 郜宜梅[1], 耿家宁[1,13], 龚光辉[1], 贡金英[3], 顾静莉[2], 顾军[3], 顾文漪[2], 顾晓程[7], 关巧宁[1], 桂琦[1], 郭道荣[1], 郭丽丽[13], 韩桂清[1], 韩旭荣[1], 黄薇[2], 贺凤英[3], 何佳颖[1], 贺林[12], 胡杰[1], 胡松年[1], 黄芳[1], 黄谷阳[2,12], 侯晓迪[9], 贾佳[2], 贾楠[1], 贾杏长[1], 蒋璐[1], 蒋延瑞[3], 蒋自由[1], 金业涛[1], 金泳三[1], 康慧[2], 康宁[3], KING Mary-Clare[2,14], 孔怡[1], 寇朝晖[3], 赖江华[1,11], 雷猛[1],李昌锋[1], 李晨吉[1], 李尔尧[3], 李飞飞[1], 李钢[1], 李家洋[1], 李继红[1], 李京[1], 李京飞[1], 李京湘[1], 李黎[1], 李丽丽[1], 李琳[1], 李蒙[12], 李楠[1], 李冉[1], 李声斌[11], 李淑萍[1], 李淑霞[1], 李双顶[11], 李双利[1,13], 李松岗[1], 李涛[1,11], 李伟[1], 李文杰[1], 李彦[1], 李燕妮[1], 李懿[1], 李志江[1], 廖劲松[1], 林伟[1], 凌伦奖[4], 林为[12], 刘斌[1], 刘波勇[1], 刘聪[9], 刘殿武[13], 刘海利[1], 刘景[1], 刘俊[1], 刘凯[1], 柳宁[10], 刘思奇[1], 刘伟[3], 刘新社[11], 刘燕华[1], 刘燕楠[1], 刘亦琳[1], 刘瑛[1], 刘裕[1], 刘玉[1], 刘占威[1], 陆红潮[5], 卢凌峰[2], 陆涛[3], 路甬祥[1], 罗春清[1], 吕纲[15], 马成[1], 马娇[1], 马清梅[3], 孟姗姗[1], 牟峰[1], 倪林[6], 牛宇欣[1], 潘交锋[1], 潘思

* 并列第一作者
‡ 通讯作者, E-mail: wangjian@genomics.org.cn; junyu@genomics.org.cn; chiangbq@ms.imicams.ac.cn; zchen@stn.sh.cn; yanghm@genomics.org.cn

塑[9], 亓秋辉[1], 齐晓华[1], 钱需方[12], 钱增敏[1], 强伯 勤[3], 乔振勇[12], 任双喜[12], 容利[7], 邵春燕[2], 邵玉 芬[1], 沈枫叶[12], 沈岩[7], 石宝辰[4], 史贵彬[3], 石宏方[1,13], SKOGERBO Geir[5], SMITH Michael[2,15], 宋丽萍[1], 宋虎平[1], 苏小虎[9], 孙洪[4], 孙家佳[1], 孙健东[1], 孙敏[1],孙涛[1], 孙伟[1], 孙永巧[1,13], 孙玉[1], 孙悦[1], 谈昕煜[7], 唐湘君[1], 陶然[1], 田朝光[1], 田薇[1], 田雁[1], 田勇[3], 田余庆[10], 佟景丽[1], 屠越峰[2], 万马[12], 王东[1], 王凤[1], 王光信[1], 王贵海[1], 王红娟[12], 王宏伟[7], 王浣洁[9], 王会峰[1], 汪建[1], 王坚 强[1], 王娟[1], 王俊[1], 王莉[1], 王立杰[1], 王丽娟[1], 王丽君[2], 王立群[12], 王世芬[1], 王 树[3], 汪文君[1], 王晓雷[1], 王晓玲[1], 王晓宁[10], 王雪 飞[1], 王学刚[1], 王艳[1], 王 莹[1], 王玉[16], 王媛媛[1], 魏树林[1], 黄家树[1,14], 吴春根[12], 吴东颖[1], 吴清发[1], 吴瑞明[1], 吴晓[9], 吴晓静[1], 郗艳[1], 席颍颖[3], 谢飞[1], 熊煜青[1], 许景义[5], 徐汝琴[1], 徐淑华[12], 徐薇[1], 徐研[9], 徐玉宁[3], 许祖元[1], 宣震宇[11], 薛蕊[1], 薛雅丽[1], 闫春霞[1], 严非[14], 严光美[2,17], 杨焕明[1], 杨睿[9], 杨树栋[1], 杨晓楠[1], 杨孔奎[1], 姚志建[7], 叶辰[1], 殷海风[2], 尤 波[1], 于兵[1,11], 于军[1], 袁开文[1], 袁义达[1], 岳淑洁[1], 曾一新[1], 翟东[1], 张波[1], 张彩民[1], 张凤梅[1], 张广 宇[1], 张国华[1], 张海清[1], 张洪波[1], 张晶分[5], 张晶 晶[1], 张久纯[1], 张兰芝[1], 张力[1], 张梅花[1], 张猛[1], 张鸣[12], 张如华[1], 张铁真[1], 张玮[12], 张祥林, 张晓良[1], 张秀清[1], 张玮[1], 张焱[3], 张夷[2], 张以琳[1], 张颖[1], 张勇[4], 张远森[1,13], 张玉志[1], 张志华[4], 赵红梅[1], 赵立见[1,13], 赵志静[1], 甄志成[7], 郑海波[1,11], 钟鸣[12], 周方楠[1], 周海霞[1], 周楠楠[1], 周鑫峰[3], 周炎[12], 周义[3], 朱炳颖[12], 朱波 峰[1,11], 朱峰华[2], 朱根风[2], 朱淼[1], 朱宁[7], 朱永革[1], 朱桢[1], 庄世涛[3]

项目领导小组:

陈竺, 傅刚, 顾军, 黄薇, 李家洋, 李载平, 刘谦, 强伯勤, 沈岩, 王贵海, 汪建, 徐新来, 许智宏, 杨焕明, 于军, 赵国屏

顾问委员会:

吴旻, 李载平, 王琳芳, 赵国屏, 柴建华, 王宇, 贺福初, 罗泽伟, 谈家桢

1 中国科学院北京基因组研究所, 中国北京 101300 /浙江大学 James D,Watson 基因组研究院, 中国杭州 310008
2 国家人类基因组南方研究中心, 中国上海 201203
3 国家人类基因组北方研究中心, 中国北京 100176
4 中国科学院生物物理研究所生物信息实验室, 中国北京 100080
5 中国科学院计算技术研究所智能信息重点实验室生物信息学实验室 ncRNA 研究小组, 中国北京 100080
6 丹麦奥尔胡斯大学人类遗传学研究所, 丹麦奥尔胡斯 8000
7 北京大学生命科学学院, 中国北京 100871
8 温州医学院, 中国温州 325003
9 中国人民大学附属中学, 中国北京 100080
10 东南大学医学院, 中国南京 210096
11 西安交通大学医学院, 中国西安 710049
12 上海交通大学 Bio-X 生命科学研究中心, 中国上海 200030
13 河北医科大学, 中国石家庄 050017
14 华盛顿大学医学遗传学系, 美国西雅图 206-543-1060
15 加拿大 BC 癌症研究中心基因组测序中心, 加拿大温哥华 2329
16 北京大学癌症治疗诊断中心, 中国北京 100035
17 中国科学院微生物研究所, 中国北京 100080