

人类基因组 1 号、X-Y 染色体 CDS 区多态性统计分析*

夏侯文浩¹, 何 洪², 江庆澜³, 何 森¹

(1. 中山大学生命科学学院, 广东 广州 510275)

(2. 中山大学教务处, 广东 广州 510275)

(3. 广州市第一人民医院中心实验室, 广东 广州 510180)

摘 要: 依托 GenBank 数据库资源, 分别对 1 号、X-Y 染色体上 CDS 区的 SNPs 分布、类型和密度进行了初步分析。统计结果发现, AG (GA) 和 CT 类型占优, 分别各占 30% 以上; 类型 AC、AT、GT 和 CG 统计频率在相同的数量级上, 分别各占 8% 左右; 并且这种分布频率不会随不同的染色体而变化。但是, SNPs 在不同染色体上基因内的分布差异大。SNPs 在 Y 染色体上的密度最小, 在 1 号染色体上的密度最大。一般来说, SNPs 的分布位置主要集中在内含子区, 但是, 在 Y 染色体上, 外显子区的 SNPs 频数明显升高。

关键词: 人类基因组; CDS 染色体; 多态性

中图分类号: Q754 **文献标识码:** A **文章编号:** 0529-6579 (2006) 02-0123-03

随着人类基因组计划测序工作的完成, 人类基因组的进一步数据挖掘工作已经成为新的研究热点。人类每个个体具有 99.99% 的相同基因, 任何两个不同个体之间大约每 1 000 个核苷酸会有一个不同, 这被称之为单核苷酸多态性 (SNP)。每个人都有自己的一套 SNPs, 它对“个性”起着决定的作用。

相关研究面临的一个挑战, 就是在基因组水平上, 如何找到连锁不平衡和单倍型模块的特征, 并且在此基础上, 识别出与复杂疾病相关的突变。SNPs 可以用于研究人类基因组中的连锁不平衡, 定位疾病相关基因; 还可以用于探究人类的进化历史和族群的变化^[1-2]。近期的研究工作在含扩展单倍型模块的基因组区域内, 发现了可观的连锁不平衡变化, 这些区域为重組热点区域; 与 GC 含量一样, 基因组中 SNPs 的分布存在很大的变化^[3]。一般性的结论是, 在人类基因组中, 平均每 1 000 bp 碱基存在一个 SNP; 所以, 在整个人类基因组中大约有 300 万个 SNPs。近期的估计是, 人类基因组大约有 1 000 万个 SNPs^[4]。

研究发现, SNPs 与内含子可能存在某种关联。内含子大约占人类基因组的 1/4, 而基因组大小和内含子大小间也存在着广泛的相关性^[5]。

研究确认了 X 染色体上有 1 098 个蛋白质编码基因, 这些基因中只有 54 个在对应的 Y 染色体上

有相应功能的等位基因; Y 染色体比 X 染色体小得多, 大约有 78 个基因。X 染色体是研究人类疾病的重要突破口, 目前已知有超过 300 种疾病定位于 X 染色体上, 包括红绿色盲、血友病、各种智障和杜兴氏肌营养不良症等等^[6]。近年来对 Y 染色体 SNPs 的分析, 使得在人类进化、人类种群的演化和迁徙领域取得了一系列重要成果。对 1 号染色体的研究报道目前尚属少见。本文希望为进一步开展遗传病和人类进化等研究提供初步的理论基础。对于 SNPs 的研究, 可以带来人类种族观念、基因治疗等方面的重大变革。

1 数据和方法

本文研究使用的序列数据来自 GenBank 的人类基因组数据库。本文选择人类基因组的 1 号、X-Y 染色体为研究对象, 研究对象包括常染色体和性染色体的样本。Y 染色体也是最小的染色体, 大小只有 1 号染色体的 1/10。X 染色体是 Y 染色体的六倍。研究显示在 Y 染色体上的遗传物质比其他染色体上的年青。LINE 元件和 MAIR 家族在 Y 染色体上都比在常染色体上年轻。1 号染色体是次大的染色体, 最大的是 2 号染色体。但是这两个染色体大小差异只在 5% 以内。对于 1 号染色体的研究比较少, 目前研究工作多集中于最大、最小染色体。实际上, 1 号染色体上面的基因密度也较高。

* 收稿日期: 2005-09-18

基金项目: 广东省科技计划项目资助 (2003A3080503)

作者简介: 夏侯文浩 (1978 年生) 男, 硕士研究生, 通讯联系人: 何森, E-mail: h24@mail.sysu.edu.cn

利用 Perl语言为工具，从 GenBank整理和获取人类基因组的研究数据，编写 CGI程序，以及设计开发统计分析程序。文中分别就 1号、X Y染色体上 CDS区的 SNP类型、SNP密度和位置分布等进行了统计分析。

2 结 果

2.1 SNP类型 在 1号染色体上，SNP的主要类型是 CT AG AC CG AT和 GT CT和 AG出现的频数占绝对优势，分别为 70 074和 67 702。CG AC和 AT的出现频数在一个数量等级上，可以归属为一类；GT和 TG的出现频数在一个数量等级上，可以归属为另一类。

在 X染色体上，SNP的主要类型是 CT AG CG TG AC和 AT CT和 AG出现的频数占绝对优势，分别为 15 032和 14 591；CG AC TG和 AT的出现频数在一个数量等级上，可以归属为一类；GA GC和 GT的出现频数在一个很低的数量等级上，可以归属为另一类。

在 Y染色体上，SNP的主要类型是 CT GA GC GT和 AC CT和 GA出现的频数占绝对优势，分别为 595和 557；GC和 GT的出现频数接近，可以归属为一类；AC和 AT的出现频数较低，可以归属为另一类。

2.2 SNP分布 分别计算 3条染色体总 CDS区外显子和内含子中的 SNP数量，以期初步了解 SNP的基本分布情况。

计算表明，所有染色体上的 SNP分布主要集中在内含子区。但是，在 Y染色体上的 SNP在外显子区所占比例相比较而言，明显高于其他染色体。结合考虑内含子在总长度所占比例，其他染色体上的 SNP的分布与序列性质的相关性小，而与序列长度的相关性大，只是在 Y染色体上出现了在外显子区分布比率增加的情形。

2.3 SNPs密度 统计结果显示，在 1号染色体上共有 2 048个 CDS区域，含有 20 7221个 SNPs，平均一个 CDS区域内含有 101个 SNPs；可见 SNP的发生频率是十分高的。

在 X染色体上有 736个 CDS区域，含有 46 116个 SNPs，平均一个 CDS区域内含有 63个 SNPs，大大低于 1号染色体上的 SNP密度。

在 Y染色体上，有 73个 CDS区域，包含 1 747个 SNPs，平均一个 CDS区域内含有 24个 SNPs。只有 1号染色体上的 SNP密度的 1/5左右。

表 1 1号、X Y染色体上的 SNP类型
Tab 1 Types of SNPs in Chromosome of No. One X Y

	SNP类型	频数	百分率 /%	累计百分率 /%
1号染色体	AC	17196	8.3	8.3
	AG	67702	32.7	41.0
	AT	14555	7.0	48.0
	CG	17257	8.3	56.3
	CT	70074	33.8	90.1
	GA	2366	1.1	91.3
	GC	571	0.3	91.6
	GT	7552	3.6	95.2
	TG	9948	4.8	100.0
	小计	207221	100.0	
X染色体	AC	4059	8.8	8.8
	AG	14591	31.6	40.4
	AT	3551	7.7	48.1
	CG	4276	9.3	57.4
	CT	15032	32.6	90.0
	GA	379	0.8	90.8
	GC	97	0.2	91.0
	GT	90	0.2	91.2
	NC	1	0.0	91.2
	TG	4040	8.8	100.0
	小计	46116	100.0	
Y染色体	AC	135	7.7	7.7
	AT	92	5.3	13.0
	CT	595	34.1	47.1
	GA	557	31.9	78.9
	GC	211	12.1	91.0
	GT	157	9.0	100.0
	小计	1747	100.0	

表 2 1号、X Y染色体上内含子和外显子区的 SNP数量

Tab 2 The Numbers of SNPs in Genes of Chromosome of No. One X Y

	位置类型	SNP频数	百分率 /%	累计百分率 /%
1号染色体	外显子	8709	4.1	4.1
	内含子	202713	95.9	100.0
	总计	211422	100.0	
X染色体	外显子	2135	4.5	4.5
	内含子	44925	95.5	100.0
	总计	47060	100.0	
Y染色体	外显子	178	10.1	10.1
	内含子	1590	89.9	100.0
	总计	1768	100.0	

3 问题与讨论

通常 SNP的分布主要集中在内含子区，在 Y染色体上的 SNP在外显子区分布比率显著偏高。

SNPs 在不同染色体上的基因内分布差别大，在 Y 染色体上的密度最小，在 1 号染色体上的密度最大。

从分子层面上，SNPs 的发生类型是相对独立的，遵循一定的发生概率规则，而其密度则可以随着染色体的不同有大的变化，这可能与染色体年龄和 CDS 区的长度有关。结合考虑内含子在总长度所占比例，其他染色体上的 SNP 的分布与序列性质的相关性小，而与序列长度的相关性大；在 Y 染色体上，SNPs 可能与外显子存在某种相关性。从 SNP 的角度来看，在各个染色体上的分布变化不大，从而，利用 SNPs 作为基因组相关信息标记可能会更加可靠。

参考文献:

[1] DURET L. Statistical analysis of vertebrate sequence

reveals that long genes are scarce in GC-rich isochores [J]. J Mol Evol 1995 40 308—317.

[2] HURST L D. Evolutionary genomics: reading the bands [J]. Bioessays 2000 22 105—107.

[3] SACCONI S. Correlations between isochores and chromosomal bands in the human genome [J]. Proc Natl Acad Sci 1993 90 1929—1933.

[4] ZOUBAK S. The gene distribution of the human genome [J]. Gene 1996 174 95—102.

[5] GARDNER K. Base composition and gene distribution: critical patterns in mammalian genome organization [J]. Trends Genet 1996 12 519—524.

[6] ROSS N T. The DNA sequence of the human X chromosome [J]. Nature 2005 434 325—337.

CDS Region Polymorphism Analysis of No. 1, X, Y Chromosomes of Human Genome

XIAHOU Wen-hao, HE Hong, JIANG Qing-lan, HE Miao

- (1. School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China;
- 2. Division of Educational Administration, Sun Yat-sen University, Guangzhou 510275, China;
- 3. Core Lab, the First Municipal People's Hospital of Guangzhou, Guangzhou 510180, China)

Abstract: The polymorphisms of CDS regions of the No. 1, X, Y chromosomes of the human genome were analyzed in this paper. The main information was mined, which included the density of the SNPs, the type of the SNPs and the distribution. There are several possibilities that the SNPs took place, the main situations were that AG (GA) and CT hold the main quantities above 30%; AC, AT, GT and CG with the same class hold quantities about 8%. The distribution does not change in different chromosomes. The SNPs were mainly found in the introns, but the SNPs ratio is higher in the exons of chromosome Y than in other chromosomes.

Key words: human genome; CDS; chromosome; polymorphism