PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2012, 39(4): 368~377
www.pibb.ac.cn

人类基因组信息量的扩增速率受重组的影响 *

刘国庆 [1, 2, 3] 罗辽复 [3]

([1] 内蒙古科技大学数理与生物工程学院，包头 014010；[2] 内蒙古科技大学生物工程与技术研究所，包头 014010；
[3] 内蒙古大学物理科学与技术学院，呼和浩特 010021)

摘要 减数分裂重组是基因组进化的重要驱动力，揭示基因组在与重组有关进化压力下的进化规律是基因组进化研究领域的重要课题．一系列证据表明基因组编码信息量在进化过程中随时间增加．重组率与自然选择效率成正比，因此，在进化过程中基因组信息量的增加速率可能会受到重组率的影响．本文定义表征编码信息量增加速率的信息参数，以人类基因组为研究对象，分析基因组信息量的增加速率与重组率的关系，发现二者显著正相关，表明重组可能是加快基因组信息量增长的重要途径．

关键词 重组率，编码信息量，信息增量，选择压力
学科分类号 Q61
DOI: 10.3724/SP.J.1206.2011.00261

减数分裂重组是指真核细胞减数分裂过程中同源染色体之间发生的遗传物质的交换事件．重组的作用是多方面的．从细胞过程来看，重组通过形成交叉(chiasma)确保减数分裂过程中的同源染色体的正确分离[1-2]．从进化的角度而言，减数分裂重组对基因组进化意义深远．减数分裂是有性生殖生物世代交替的转折点，减数分裂过程中发生的遗传重组从分子水平上为自然选择提供途径．如果同源染色体之间不能发生物质交换，每一条染色体所含有的遗传信息就只能被固定在特定等位基因上，当突变发生时有利的和不利的改变将很难分开．通过基因重排，重组使有利的和不利的突变分开，促进有利等位基因的扩散和扩展，同时在不影响其他连锁基因的条件下消除有害的等位基因[3]．重组还可能通过以下两种突变的方式影响基因组进化：a．重组可能具有诱变性而改变突变率[4]；b．重组过程中异源双链 DNA 分子的形成会诱导偏向 GC 的基因转换[5]．上述三种作用是重组影响基因组进化的根本途径，由于这些作用的存在，重组频率沿染色体上的不均匀分布[6]对遗传信息的组织和构建产生重要影响．例如，基因组 GC 含量[7]、二核苷相对丰度[8-9]、密码子偏好性[10]、重复元件的分布[11-12]、假基因的分布[13]、内含子的长度[14]、C_pG 岛分布[12]、基因的固定概率和蛋白质的进化速率[15-16]在基因组内的不均匀性与重组频率的不均匀性之间有不同程度的相关性．重组也有可能影响基因组编码信息量的扩增速率．

Luo[17-18]在大量实验发现和理论研究的基础上总结归纳，提出以信息流为主线的基因组进化方向的假说：在 DNA、RNA、蛋白质的相互作用下，通过序列复制、编码方式增加，以及基因在基因组间转移等机制，生物体基因组 DNA 的编码信息量在进化中随时间增加．这里的编码信息是指功能编码信息，它包含两方面的内容：a．蛋白质编码信息，其载体序列就是通常意义上的蛋白质编码序列； b．非蛋白质编码信息，是指基因表达调控信息．

揭示作用在基因组上的进化压力及其进化方向是基因组进化研究的核心内容．基因组编码信息量在进化过程中随时间增加的这一假说，从总体上描述了基因组进化的方向．其中编码信息量的增加速

* 国家自然科学基金(61102162, 90403010), 内蒙古自治区高等学校科学研究项目(NJ10098)和内蒙古科技大学创新基金(2009NC005)资助项目．
** 通讯联系人．
Tel: 0472-5954358, E-mail: gqliu1010@163.com
收稿日期：2011-06-09，接受日期：2012-02-13

率可能受到重组率的影响，因为重组率反映自然选择效率[19-20]，而选择效率与编码信息量的增加速率成正比．本文定义表征编码信息量增加速率的信息参数，并以人类基因组为研究对象，在全基因组范围内分析编码信息量的增加速率与重组率的关系，研究基因组在重组这一进化压力下的进化规律．

1 数据和方法

1.1 基因组序列

人类基因组全序列(human build 36, reference sequence, updated in 2006)取自 GenBank (ftp://ftp.ncbi.nlm.nih.gov)．基于 human build 36 人类基因组 Ensembl 基因的注释信息(包括基因转录起始、终止位点，外显子和内含子位点信息)是从 UCSC 基因组浏览器(http://www.genome.ucsc.edu)检索得到的．根据位点信息，从人类基因组全序列上摘取内含子序列、基因间序列和编码序列．这里，基因间序列指的是上一个基因的转录终止位点和下一个基因的转录起始位点之间的间隔序列．若注释文件中未给出基因的转录起始和终止位点，则上一个基因的终止密码子和下一个基因的起始密码子之间的间隔序列被选定为基因间序列．为避免冗余，遇见基因重叠或外显子可变剪接情况时只选取与第一转录本对应的内含子和编码序列；基因间序列的选取是“纯”而又非冗余的，即被选取的基因间序列之间不能有重叠区域，并且切掉它与重叠基因的重叠区．

1.2 重组率

人类遗传图谱数据[21]从遗传图谱服务器 MAP-O-MAT [22] (http://compgen.rutgers.edu/mapomat) 获取．此遗传图谱是合并人类遗传多样性研究中心(Centre d’Etude du Polymorphisme Humain(CEPH))基因型数据和 deCODE 家系基因型数据以及单核苷多态性数据的高密度遗传图谱，共有 14 759 个遗传标记．图谱中已给出每个遗传标记在 human build 36 人类基因组上对应的物理位置．重组率计算方法类似于文献[23]中所述，把每条染色体上遗传标记的遗传位置作为其物理位置的函数进行三次样条(cubic spline)拟合，并求出拟合函数在每个遗传标记物理位置处的一阶导数作为该位点的重组率(cM/Mb)．计算 5-Mb 非重叠滑动窗口的平均重组率时，粗略假定相邻遗传标记之间所有位点的重组率是相同的，则每个窗口的平均重组率为 5 Mb 个位点的重组率平均值．这样，我们得到 22 条常染色体和 X 染色体的性别平均重组率．Y 染色体不重组(假常染色体区域除外)．

1.3 信息增量

对于序列，定义多样性指标

$$H=-N\sum_i p_i\log_2 p_i=N\log_2 N-\sum_i m_i\log_2 m_i \qquad (1)$$

定义编码信息量[17-18]：

$$I_C=\log_2 s^N=N\log_2 s=H_{\max} \qquad (2)$$

其中为 $H_{\max}$ 多样性指标 H 的最大值，m_i 和 p_i 分别为序列中长度为 k 的特定信息符号 i 出现的频数和频率，s 为信息符号数目，N 为所有信息符号出现的总数，与序列长度 L 的关系为 $N = L_{-k}+1$．这里，我们称长度为 k 的信息符号为 k-mer．对于 DNA 序列，$s=4^k$．

定义微分信息量(或称信息增量)

$$h(x)=\frac{\mathrm{d}H(x)}{\mathrm{d}x} \qquad (3)$$

$h(x)$表示序列增长 $\mathrm{d}x$ 长度时序列的多样性增量为 $\mathrm{d}H(x)$，微分信息量是与序列长度(而非时间)有关的编码信息量的增加速率．

易证 $$h(x)\geqslant 0 \qquad (4)$$

序列的增长往往采取随机插入的方式，最有效的是片段重复[24]．为便于理论计算，我们采取简化模型，令序列的增长是一次一个碱基单方向增加的过程，则微分信息量表达式为：

$$\begin{aligned}h(x)_{x=N}&=\{\frac{\mathrm{d}H(x)}{\mathrm{d}x}\}_{x=N}=H(N+1)-H(N)\\&=(N+1)\log_2(N+1)-N\log_2 N+m_{iN}\log_2 m_{iN}\\&\quad-(m_{iN}+1)\log_2(m_{iN}+1)\end{aligned} \qquad (5)$$

其中位点依赖的多样性 $H(x)$为序列 0–x 位点间的多样性，m_{iN} 为序列第 N 位后的第一个碱基 i 在 0–N 位点之间的频数．当 N 很大时式(5)可化简为

$$\begin{aligned}h(x=N)&=N\frac{\ln(1+\frac{1}{N})}{\ln 2}+\log_2(N+1)-m_{iN}\frac{\ln(1+\frac{1}{m_{iN}})}{\ln 2}-\log_2(m_{iN}+1)\\&=\log_2(\frac{N+1}{m_{iN}+1})\end{aligned} \qquad (6)$$

信息增量这一指标表示基因组增长的过程中编码信息量的增加速率．从式(5)可知信息增量表示基因组序列增长之后和增长之前序列多样性的差值．多样性是与信息量相平行的从信息角度对状态空间的一种描述．Shannon 信息量是对信息源中状态不确定性或紊乱性的一种描述．如果我们关心的不仅是各个状态可能出现的概率(相对频率)，而且是它们的绝对频数，那么，Shannon 信息量就应换成 Laxton 多样性．可以证明多样性是信息量的 N

(各种状态出现的总频数)倍[25]. 我们定义的信息增量可以定量地表示序列增长前后的多样性增加程度，其数值越大，说明增长的那一段序列对多样性的贡献越大，这一点可以从式(6)看出来. 在本问题中涉及的是序列增长过程中信息符号(如 4 种碱基)绝对频数的变化，故应该用多样性. 用 h_k 表示以 k-mer 为信息符号时候的信息增量. 以 h_3 为例说明信息增量的计算方法：公式(6)中的 N 为计算位点上游序列(称之为 bath)中所有 3-mer 在单碱基步长计数中出现的总频数，m_{iN} 为计算位点的特定 3-mer 在 N 中出现的频数. 上游序列较长时 N 近似等于上游序列长度. 从染色体一端开始，每移动一个碱基，N 增加 1，直到一条染色体结束为止.

1.4 两类信息增量

用两种方法计算信息增量：a. 从染色体正链 5′端开始扫描，用公式(6)计算每一个位点上的信息增量，取每 5-Mb 窗口内的平均值作为该窗口所对应的信息增量. 计算信息增量时，随着计算程序在染色体上扫描过程的推进，bath 的长度一直增加下去，直到扫完一条染色体为止. 这样计算出来的信息增量叫作第一类信息增量. 注意，分别计算三种序列(编码序列、内含子和基因间序列)的第一类信息增量时，从计算位点上游的天然染色体序列统一计数公式(6)中的 N. 这种计算方法中，bath 为计算位点上游的全部序列，是三种序列的混合序列. b. 第一类信息增量的计算方法中，bath 是天然序列中计算位点上游的全部区域. DNA 序列是近似马尔科夫链，即将来发生的事情只与现在有关，而与过去无关. 相对于上一个进化事件而言的信息增量叫做第二类信息增量. 第二类信息增量的概念符合马尔科夫链的设想. 计算第二类信息增量时，bath 取每 5-Mb 窗口上游的固定长度的一段序列. 本文计算第二类信息增量时 bath 大小选为 50 kb. 第二类信息增量和第一类信息增量只在 bath 的选择上不同.

1.5 相关分析

研究表明人染色体重组率在 kb 尺度水平上沿染色体变化[26]，但是在 Mb 尺度上人类基因组重组率较保守[27]. 因此，在相关研究中以几个 Mb 大小的滑动窗口作为研究单元，能够避免重组率本身的变化速率较快的影响. 我们以 5-Mb 非重叠滑动窗口计算重组率和信息增量，并分析二者的相关性. 不同类型的序列进化规律可能有所差异，因此对编码序列、内含子和基因间序列进行了单独分析. 我们对全基因组序列还进行了整体分析，即不区分序列类型的情况下直接分析 5-Mb 窗口序列的信息增量和重组率的相关性. 对两个变量间的直线关系进行相关分析称为直接相关分析(或两两相关分析，pairwise correlation). 本文采用 Spearman 相关分析. Spearman 相关系数是对两变量等级或秩次(rank)间是否相关的一种测度，它用两变量的秩来进行直线相关分析，即先按两变量大小顺序，由小到大排上秩次，再看两变量的秩次间是否相关[28-29]. 为消除 GC 含量对信息增量和重组率之间相关关系的影响，采用了偏相关分析(partial correlation)[28-29]. 偏相关分析是指当两个变量同时与其他变量相关时，将其他变量的影响剔除，只分析指定两个变量之间相关程度的过程. 用偏相关系数来度量在其他变量都保持不变时指定的两个变量间的相关程度.

2 结　果

2.1 信息增量的计算结果

从每条染色体正链 5′端开始扫描计算每 5-Mb 窗口的第一类信息增量 $h_k(h_1, h_2,\cdots, h_8)$，图 1 中以 4 号染色体为例给出了 h_1 和 h_6 沿染色体的分布. 图 2 中给出了 4 号染色体第二类信息增量 h_1 和 h_6 (bath 大小均为 50 kb)的分布.

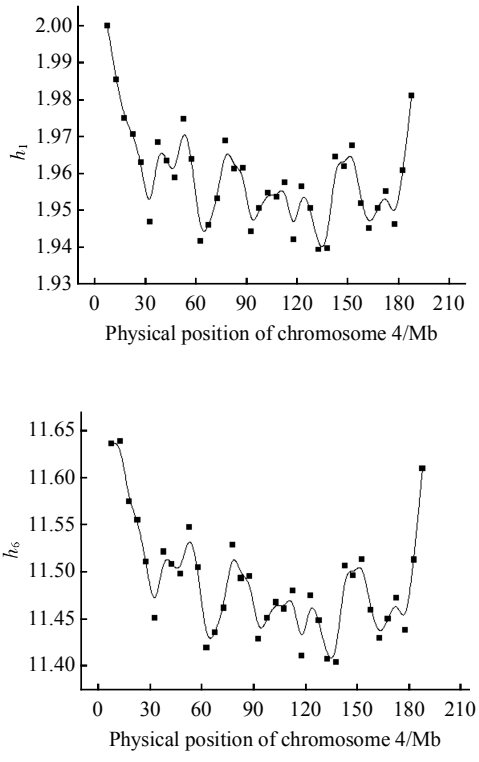

Fig. 1 The genomic distribution of the first type of information increment along human chromosome 4

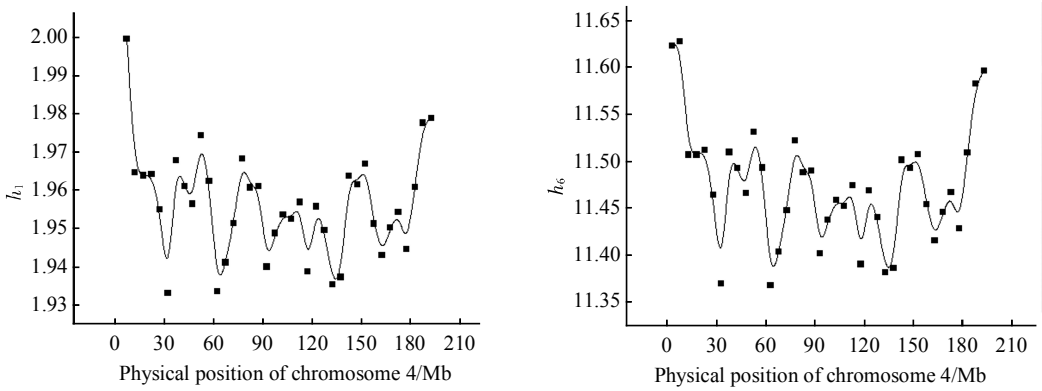

Fig. 2 The genomic distribution of the second type of information increment along human chromosome 4

从图 1 和图 2 可发现：不论是第一类还是第二类信息增量，其沿染色体的分布呈现出很强的不均匀性；染色体两端的信息增量较大，中部的较低，换言之，在基因组序列的增长过程中，染色体两端信息量的增长速率较快；同类信息增量的分布模式之间以及第一类和第二类信息增量的分布模式之间都很相似．其他染色体信息增量的分布趋势与 4 号染色体一致(数据未给出)．进一步分析表明，同类信息增量(如第一类信息增量 $h_1, h_2, \cdots, h_8$)之间、第一类和第二类信息增量之间均存在很强的线性正相关关系(表 1)，前者表明 h_1 中已包含了其他信息增量 h_k 的大部分信息，而后者可能是因为信息增量对 bath 大小的依赖性较弱造成的．

Table 1 Pearson correlations within and between the first and second type of information increment (h_k) averaged over 5-Mb windows for complete human genome

	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8
h_1	**0.98**	0.99	0.99	0.98	0.97	0.93	0.80	0.56
h_2	0.99	**0.97**	0.99	0.99	0.98	0.94	0.80	0.56
h_3	0.99	0.99	**0.97**	0.99	0.98	0.94	0.81	0.57
h_4	0.99	0.99	0.99	**0.96**	0.99	0.95	0.83	0.60
h_5	0.98	0.99	0.99	0.99	**0.95**	0.98	0.87	0.65
h_6	0.95	0.96	0.96	0.97	0.98	**0.93**	0.94	0.77
h_7	0.86	0.86	0.87	0.88	0.91	0.96	**0.89**	0.92
h_8	0.67	0.68	0.69	0.71	0.74	0.82	0.93	**0.82**

Data (in bold) in the diagonal line denote correlations between the first and second type of information increments. Data under the diagonal line denote correlations among the first type of information increments. Data above the diagonal line denote correlations among the second type of information increments. All the correlations in the table are significant ($P < 0.001$). Sample size n=583.

图 3 直观地显示了 5-Mb 窗口第一类和第二类信息增量之间的相关性，可见对于基因间序列、内含子和基因组整体来说，第一类和第二类信息增量之间存在较好的线性关系，且前者偏大于后者．由于少数派碱基对信息增量的贡献较大，第一类信息增量偏大于第二类信息增量就意味着，基因组中任意区域的碱基组分对上游的第一类 bath(相比第二类 bath 而言)而言更倾向于是少数派碱基．对于基因编码区来说，第一类和第二类信息增量之间线性关系明显减弱，表明编码区的信息增量对 bath 的依赖性增强．从图 3 还可以看出，在三种序列中，编码序列的信息增量较大．我们还发现，信息增量随 k 的增加而线性增加(图 4)．这是因为随着 k 的增加，k-mer 的数量增加，信息量增加．

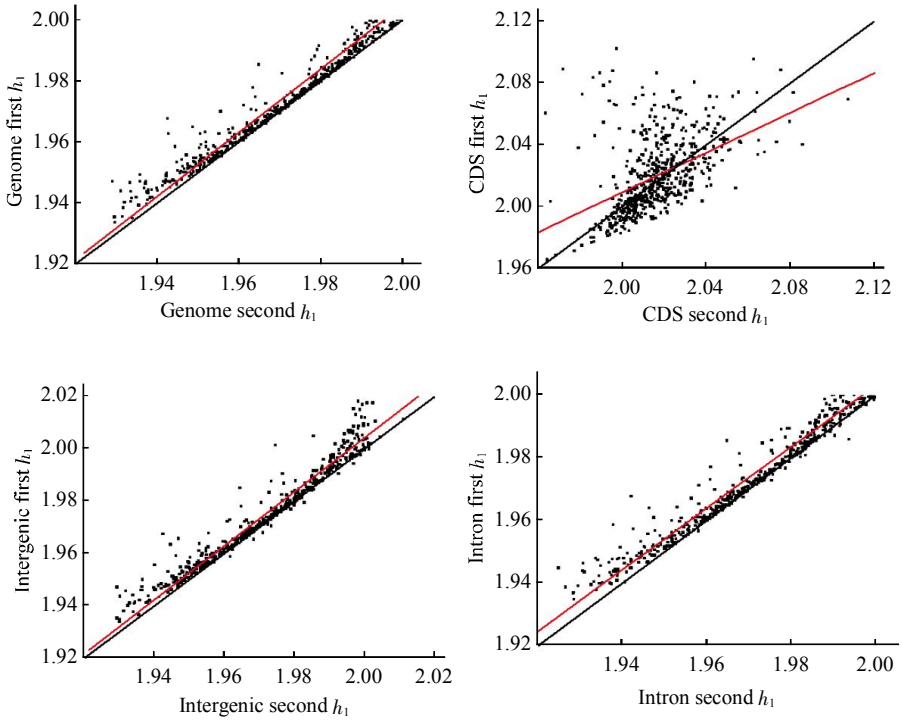

Fig. 3　Correlations between the first and second type of information increments averaged over 5-Mb windows along the genome

The red lines are least squares lines, and the black ones are reference lines on which the abscissa and ordinate are equal.

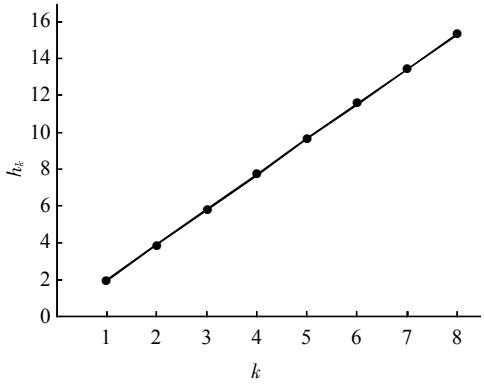

Fig. 4　Plots of the first type of information increments, averaged over 5-Mb windows along the genome, against *k*

2.2　信息增量与重组率

重组是基因组进化的驱动力. 群体遗传学理论认为，频繁重组的区域是选择效率高的区域[19-20]. 我们认为选择效率高的区域是编码信息扩增速率较快的区域. 若现在的染色体重组率数据能够很好地反映在漫长的进化过程中染色体增长时的重组频率信息，则有理由推断重组率和信息量的增加速率之间存在正比关系. 为验证这一点，本文分析了5-Mb 窗口的重组率和信息增量之间的关系，结果列于表 2 和表 3.

通过直接相关分析发现，对编码区、基因间序列、内含子和整个基因组序列来说，除了编码区的少数几个信息增量以外，其他第一类信息增量跟重组率均显著正相关，编码区的相关性稍弱(表 2). 在编码区、基因间序列和内含子序列中，编码区对信息增量的贡献是较大的，然而其信息增量受重组的影响却最弱(表 2)，显然这跟编码序列的表达相关的约束有关. 例如，在密码子使用、密码对使用、密码子第一和第二位点的高保守性(非同义突变率低)等功能约束下，编码区保持相对保守，不易受其他因素的影响. 第二类信息增量与重组率之间同样存在与第一类信息增量相似的关联性(表 3).

Table 2 Spearman correlation of recombination rate with the first type of information increment and local GC content for human genomic sequences in 5-Mb windows

		h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	GC
Genome	Pairwise	0.484***	0.497***	0.502***	0.505***	0.507***	0.503***	0.482***	0.426***	0.491***
	Partial	0.002	0.091*	0.120**	0.136**	0.146***	0.152***	0.166***	0.181***	
CDS	Pairwise	0.185***	0.225***	0.238***	0.240***	0.242***	0.247***	0.242***	0.219***	0.399***
	Partial	–0.175***	–0.116**	–0.087*	–0.074	–0.068	–0.067	–0.074	–0.100*	
Intergenic	Pairwise	0.452***	0.466***	0.472***	0.475***	0.477***	0.472***	0.444***	0.378***	0.451***
	Partial	0.053	0.133**	0.156***	0.170***	0.172***	0.167***	0.172***	0.186***	
Intron	Pairwise	0.437***	0.449***	0.450***	0.453***	0.456***	0.455***	0.436***	0.371***	0.447***
	Partial	–0.013	0.065	0.080	0.099*	0.119**	0.143***	0.171***	0.186***	

In the partial correlation analysis between recombination rate and information increment, the control variable is local GC content. Two-tailed significance: *P < 0.05; **P < 0.01; ***P < 0.001. Sample size n=563.

Table 3 Spearman correlation of recombination rate with second type of information increment and local GC content for human genomic sequences in 5-Mb windows

		h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	GC
Genome	Pairwise	0.471***	0.485***	0.492***	0.502***	0.514***	0.534***	0.545***	0.512***	0.491***
	Partial	–0.047	0.129**	0.169***	0.226***	0.262***	0.152***	0.286***	0.308***	
CDS	Pairwise	–0.033	0.035	0.060	0.068	0.081	0.124**	0.138**	0.098*	0.399***
	Partial	–0.093*	–0.024	0.007	0.022	0.030	0.049	0.076	0.098*	
Intergenic	Pairwise	0.442***	0.462***	0.475***	0.493***	0.517***	0.533***	0.422***	0.042	0.451***
	Partial	–0.072	0.125**	0.199***	0.274***	0.312***	0.323***	0.373***	0.377***	
Intron	Pairwise	0.443***	0.457***	0.466***	0.480***	0.493***	0.499***	0.378***	0.107*	0.447***
	Partial	0.004	0.103*	0.145***	0.195***	0.233***	0.277***	0.285***	0.258***	

In the partial correlation analysis between recombination rate and information increment, the control variable is local GC content. Two-tailed significance: *P < 0.05; **P < 0.01; ***P < 0.001. Sample size n=563.

分析重组率和信息增量之间的关系时选用的窗口大小为 5-Mb. 为了消除窗口选择的人为主观性，我们做了几个对照(表 4)，发现 1-Mb 和 10-Mb 窗口的结果与 5-Mb 窗口结果十分相似，趋势依然存在. 这表明，所发现的结果不是由于人为选择 5-Mb 窗口造成的.

Table 4 Spearman correlation between recombination rate and information increment for the human genome in 1-Mb or 10-Mb windows

Window	Type	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8
1-Mb	Type1	0.396***	0.397***	0.399***	0.401***	0.406***	0.414***	0.414***	0.383***
1-Mb	Type2	0.395***	0.397***	0.402***	0.415***	0.436***	0.464***	0.474***	0.437***
10-Mb	Type1	0.492***	0.507***	0.514***	0.519***	0.518***	0.510***	0.485***	0.413***
10-Mb	Type2	0.504***	0.521***	0.527***	0.538***	0.543***	0.543***	0.525***	0.487***

Two-tailed significance: *P < 0.05; **P < 0.01; ***P < 0.001. Sample sizes for 1-Mb and 10-Mb windows are 2882 and 308, respectively. “type1” and “type2” represent the first type of information increment and the second type of information increment respectively.

2.3 信息增量对 bath 大小的依赖性很弱

信息增量与所考虑位点上游序列的碱基分布有关，我们把它叫做“浴池”效应，意思和统计物理中的“热浴”(bath)相同，这就是将计算位点的上游序列称之为 bath 的原因. bath 的长短可能会影响信息增量的值. 为考察这一点，我们在 bath 长

度不同的情况下(bath = 50 kb, 500 kb, 5 Mb, window=5 Mb; bath =10 kb, 50 kb, 500 kb, window= 500 kb，window 为计算平均信息增量所用窗口)计算信息增量．结果表明窗口大小相同的条件下信息增量几乎不依赖于 bath 大小(相关系数达 0.98 以上，如图 5 和图 6)．信息增量对 bath 大小的依赖性很弱，这就消除了 bath 长度的粗略假定的任意性．

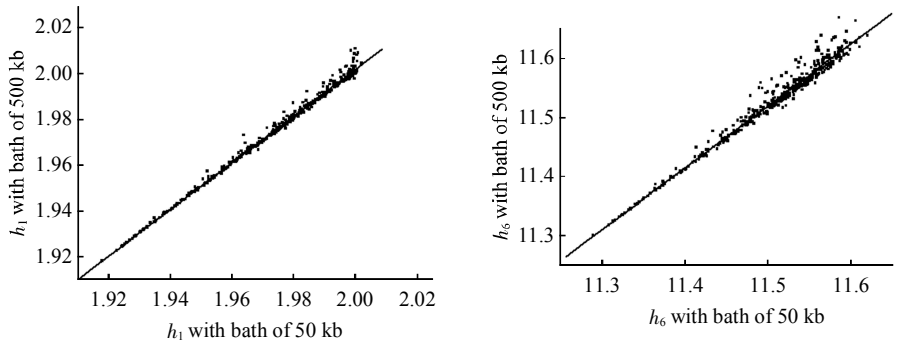

Fig. 5　Information increment is independent on bath size

Two examples using (h_1 and h_6) averaged over 500-kb windows are given.

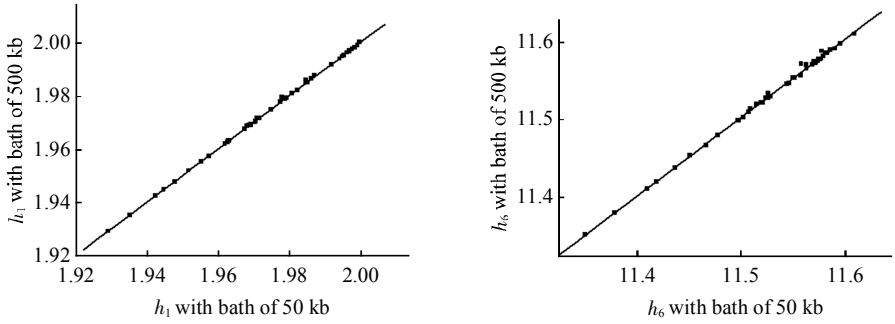

Fig. 6　Information increment is independent on bath size

Two examples (h_1 and h_6) averaged over 5-Mb windows are given.

计算结果表明(表 5)，信息增量之间的相关性随着 bath 的减小而减弱，换句话说，bath 越小，信息增量越受其大小影响．第二类信息增量之间的相关系数从 bath 为 50-kb 和 5-kb 之间的 0.95 降至 50-bp 和 10-bp 之间的 0.64．统计学上，相关系数达 0.8 以上的被视为强相关，因此，从表 5 可知 bath 小到 500 bp 以下的时候 bath 的大小对信息增量的影响较大．

Table 5　Pearson correlations among the second type of information increments (h_1) of different bath size

	50-kb	5-kb	500-bp	50-bp	10-bp
50-kb	1				
5-kb	0.95***	1			
500-bp	0.80***	0.85***	1		
50-bp	0.62***	0.66***	0.68***	1	
10-bp	0.41***	0.44***	0.43***	0.64***	1

The information increments of different bath size were computed for a 10-kb genomic region of human (chr1: 50 000～60 000 bp).

Two-tailed significance: *P < 0.05; **P < 0.01; ***P < 0.001．Sample size n=10 000.

3 讨 论

从信息增量的计算公式(6)可知，信息增量 $h(x)$ 随 $\frac{N}{m_{iN}}$ 的增大而增大，即若碱基 i 在 bath 中的含量少，则它对计算位点的信息增量贡献就大，这说明“少数派”最能产生多样性，而“多数派”碱基产生的多样性就低．一段序列包含较多的“少数派”碱基，这段序列的信息总增量就高．本文分析结果显示重组率与信息增量成正相关，表明重组是一种让“少数派”碱基表现出多样性的途径．重组可能会通过选择作用导致重组率和信息增量之间的正相关．例如，高重组区 Hill-Robertson 干涉较少，选择效率较高，从而导致这个区域的信息增长速率快．

重组过程中发生的偏向 GC 的基因转换会导致序列 GC 含量的增加[5,7]．对于人类基因组来说，全基因组平均 GC 含量约 42%，在重组相关的偏向 GC 突变效应的影响下序列 GC 含量向 50%漂变．因为 4 种碱基等概率分布时信息量最大，偏向 GC 的突变效应显然会导致序列信息量的增加．重组率与信息增量的相关性是不是由偏向 GC 的基因转换造成的呢？通过分析发现，GC 含量与信息增量之间的确存在较强的相关性(表 6)．偏相关分析结果(表 2、表 3)表明，控制 GC 含量时信息增量与重组率的关系变弱，甚至变得不显著($P > 0.05$)．然而，这并不能表明信息增量与重组率的相关性肯定由基因转换的 GC 偏向效应引起，因为 GC 含量和信息增量本身就是具有内在关联的两个参数．重组率与信息增量和偏向 GC 的基因转换皆相关，因此，从二元线性回归结合协方差分析可能会得到更好的结果，进而才能对重组率、信息增量和偏向 GC 基因转换之间的关系进行更全面的讨论．

Table 6 Pearson correlation between local GC content and information increment for the human genome in 5-Mb windows

	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8
Type1	0.984	0.976	0.970	0.963	0.951	0.915	0.808	0.613
Type2	0.997	0.992	0.988	0.982	0.969	0.925	0.795	0.560

All the correlations in the table are sighificant ($P < 0.01$). Sample size n=563. “type1” and “type2” represent the first type of information increment and the second type of information increment respectively.

本文考虑序列长度从 N 增至 N+1 时信息量的变化，这是最简单的模型，实际进化中是片段的组合导致序列的扩展，重组也是发生于片段间．故更符合实际的模型是：考虑两个片段的组合中信息量的变化．定义片段 a 和 b 的信息量分别为

$$H(a)=-N_a\sum_i p_{ai}\log_2 p_{ai}=N_a\log_2 N_a-\sum_i m_{ai}\log_2 m_{ai} \quad (7)$$

$$H(b)=-N_b\sum_i p_{bi}\log_2 p_{bi}=N_b\log_2 N_b-\sum_i m_{bi}\log_2 m_{bi} \quad (8)$$

当它们组合起来，信息量记为 $H(a+b)$

$$H(a+b)=(N_a+N_b)\log_2(N_a+N_b)-\sum_i (m_{ai}+m_{bi})\log_2(m_{ai}+m_{bi}) \quad (9)$$

易证 $H(a+b)\geqslant H(a)+H(b)$，定义 a 和 b 组合中信息增量

$$I(a,\ b)=\frac{H(a+b)-H(a)-H(b)}{H(a)+H(b)} \quad (10)$$

计算人类基因组每两个相邻的 50-kb 片段 a 和 b 组合中的信息增量 $I(a, b)$，并在 5-Mb 窗口对应的信息增量平均值和重组率之间进行相关分析，结果表明二者是显著正相关的(相关系数 r=0.13, P=0.002)．这表明微分信息量 $h(x)=\frac{\mathrm{d}H(x)}{\mathrm{d}x}$ 只是组合信息增量的一种特殊情况，同时还表明本文单核苷酸信息增量模型为实际基因组进化与重组率关系的研究提供了有价值的线索．

减数分裂重组是基因组进化的重要驱动力．通过本文分析发现，基因组信息量的增加速率与重组率正相关，表明重组可能是调控基因组信息量增长速率的重要因素．下一步我们将以基因组中进化水平不同的段落为研究对象，深入探究信息量的演化规律及相关的进化压力，以期在人类基因组中获得的结果能在其他物种中得到印证，揭示进化机制的普适性．

参 考 文 献

[1] Lynn A, Ashley T, Hassold T. Variation in human meiotic recombination. Annu Rev Genomics Hum Genet, 2004, **5**: 317-349

[2] Chowdhury R, Bois P R J, Feingold E, *et al*. Genetic analysis of variation in human meiotic recombination. PLoS Genet, 2009, **5**(9): e1000648

[3] Lewin B. Genes VIII. Upper Saddle River. NJ: Pearson Prentice Hall, 2004

[4] Lercher M J, Hurst L D. Human SNP variability and mutation rate are higher in regions of high recombination. Trends Genet, 2002, **18**(7): 337-340

[5] Galtier N, Piganeau G, Mouchiroud D, *et al*. GC-Content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics, 2001, **159**(2): 907-911

[6] Kong A, Gudbjartsson D F, Sainz J, *et al*. A high resolution recombination map of the human genome. Nat Genet, 2002, **31**(3): 241-247

[7] Meunier J, Duret L. Recombination drives the evolution of GC-content in the human genome. Mol Biol Evol, 2004, **21**(6): 984-990

[8] Liu G, Li H. The correlation between recombination rate and dinucleotide bias in *Drosophila melanogaster*. J Mol Evol, 2008, **67**(4): 358-367

[9] 刘国庆, 李 宏. 人类基因组中减数分裂重组对二核苷偏好性的影响. 科学通报, 2009, **54**(4): 448-456
Liu G Q, Li H. Chin Sci Bull, 2009, **54**(4): 448-456

[10] Singh N D, Davis J C, Petrov D A. Codon bias and non-coding GC content correlate negatively with recombination rate on the *Drosophila* X chromosome. J Mol Evol, 2005, **61**(3): 315-324

[11] Bartolome C, Maside X, Charlesworth B. On the abundance and distribution of transposable elements in the genome of *Drosophila melanogaster*. Mol Biol Evol, 2002, **19**(6): 926-937

[12] Jensen-Seaman M I, Furey T S, Payseur B A, *et al*. Comparative recombination rates in the rat, mouse, and human genomes. Genome Res, 2004, **14**(4): 528-538

[13] Liu G, Li H, Cai L. Processed pseudogenes are located preferentially in regions of low recombination rates in the human genome. J Evolutionary Biology, 2010, **23**(5): 1107-1115

[14] Comeron J M, Kreitman M. The correlation between intron length and recombination in *Drosophila*: dynamic equilibrium between mutational and selective forces. Genetics, 2000, **156**(3): 1175-1190

[15] Marais G, Charlesworth B. Genome evolution: recombination speeds up adaptive evolution. Curr Biol, 2003, **13**(2): 68-70

[16] Presgraves D C. Recombination enhances protein adaptation in *Drosophila melanogaster*. Curr Biol, 2005, **15**(18): 1651-1656

[17] Luo L F. Law of genome evolution direction: Coding information quantity grows. arXiv: q-bio/0808.3323v1; Frontiers of Physics in China, 2009, **4**(2): 241-251

[18] Luo L F. Information biology: hypotheses on coding information quantity. J Inner Mongolia University, 2006, **37**(3): 285-294

[19] Hill W G, Robertson A. The effect of linkage on limits to artificial selection. Genet Res, 1966, **8**(3): 269-294

[20] Kliman R M, Hey J. Reduced natural selection associated with low recombination in *Drosophila melanogaster*. Mol Biol Evol, 1993, **10**(6): 1239-1258

[21] Kong X, Murphy K, Raj T, *et al*. A combined linkage-physical map of the human genome. Am J Hum Genet, 2004, **75**(6): 1143-1148

[22] Kong X, Matise T C. MAP-O-MAT: Internet-based linkage mapping. Bioinformatics, 2005, **21**(4): 557-559

[23] Yu A, Zhao C, Fan Y, *et al*. Comparison of human genetic and sequence-based physical maps. Nature, 2001, **409**(6822): 951-953

[24] Hsieh L C, Luo L, Ji F, *et al*. Minimal model for genome evolution and growth. Phys Rev L, 2003, 90: 018101:1-018101:4

[25] Lu J, Luo L F, Zhang L R, *et al*. Increment of diversity with quadratic discriminant analysis-an efficient tool for sequence pattern recognition in bioinformatics. Open Access Bioinformatics, 2010, **2**: 89-96

[26] Crawford D C, Bhangale T, Li N, *et al*. Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat Genet, 2004, **36**(7): 700-706

[27] Serre D, Nadon R, Hudson T J, *et al*. Large-scale recombination rate patterns are conserved among human populations. Genome Res, 2005, **15**(11): 1547-1552

[28] 李春喜, 姜丽娜, 邵 云, 等. 生物统计学(第 3 版). 北京: 科学出版社, 2005
Li C X, Jiang L N, Shao Y, *et al*. Biostatistios. 3rd. Beijing: Science Press, 2005

[29] Hurlburt R T. Comprehending Behavioral Statistics. Belmont, California: Wadsworth Inc, 1994

The Growing Rate of The Information Quantity of The Human Genome is Modulated by Recombination*

LIU Guo-Qing[1, 2, 3)**], LUO Liao-Fu[3)]

([1)] *School of Mathematics, Physics and Biological Engineering, Inner Mongolia University of Science and Technology, Baotou* 014010, China;

[2)] *Institute of Bioengineering and Technology, Inner Mongolia University of Science and Technology, Baotou* 014010, China;

[3)] *School of Physical Science and Technology, Inner Mongolia University, Huhhot* 010021, China)

Abstract Meiotic recombination is driver of genome evolution. It is important to explore the rule of genome evolution under the control of evolutionary pressure linked to recombination. The coding information quantity of a genome (CIQ) always grows during evolution. Recombination rate is proportional to selection efficiency, thus the growing rate of coding information quantity of a genome (GR_{CIQ}) might be mediated by recombination rate. In this study, a parameter is defined to characterize GR_{CIQ}, and the correlation between GR_{CIQ} and recombination rate is analyzed in the human genome. The results show that there is a positive correlation between them, indicating recombination is likely to be an important pathway to increase GR_{CIQ}.

Key words recombination rate, coding information quantity, information increment, selection force

DOI: 10.3724/SP.J.1206.2011.00261

*This work was supported by grants from The National Natural Science Foundation of China (61102162, 90403010), The Research Program of Higher Education of Inner Mongolia Autonomous Region (NJ10098) and The Innovation Fund of Inner Mongolia University of Science and Technology (2009NC005).

**Corresponding author.

Tel: 86-472-5954358, E-mail: gqliu1010@163.com

Received: June 9, 2011 Accepted: February 13, 2012