BGI华大

Initial sequencing and analysis of thehuman genome

陈逸希 曾镜瑀 曹晨音 任宇晗

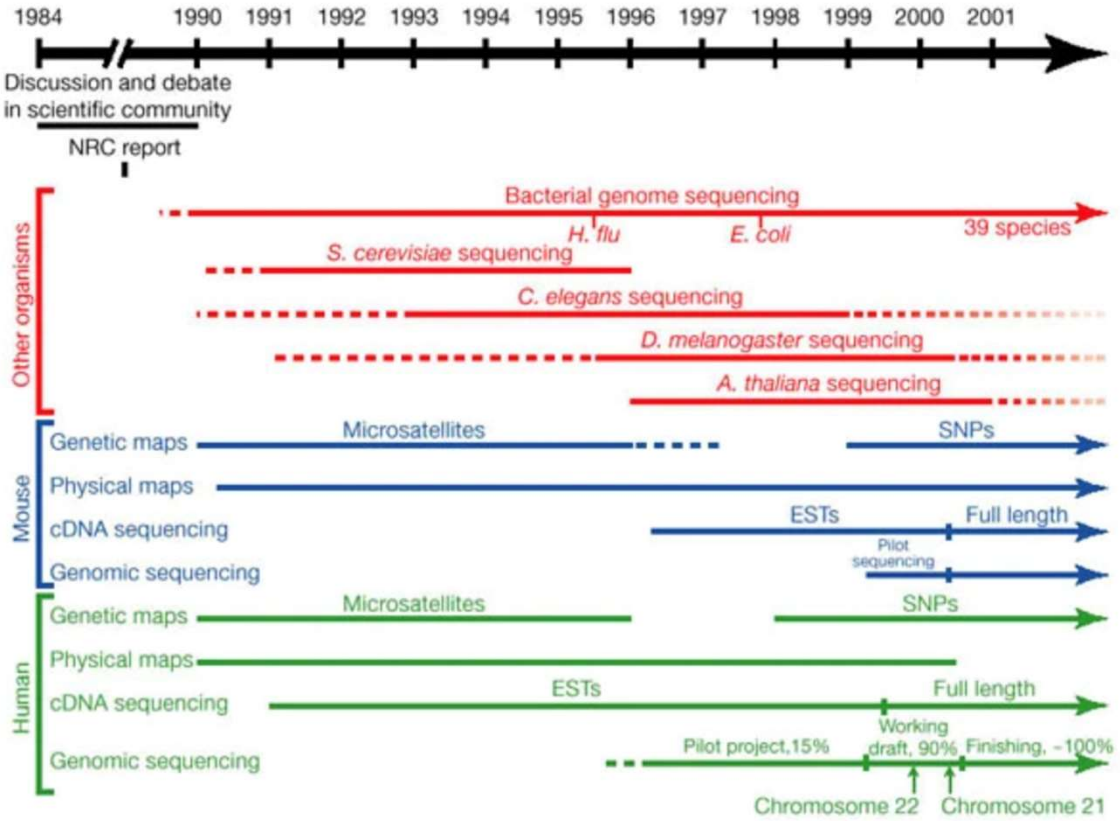
华大生命科学研究院
BGI · research

main content

- Background to the Human Genome Project
- Generating the draft genome sequence
- Broad genomic landscape
- Repeat content of the genome
- Gene content of the human genome
- Segmental history of the human genome
- Applications to biology and medicine and the next steps

BGI华大

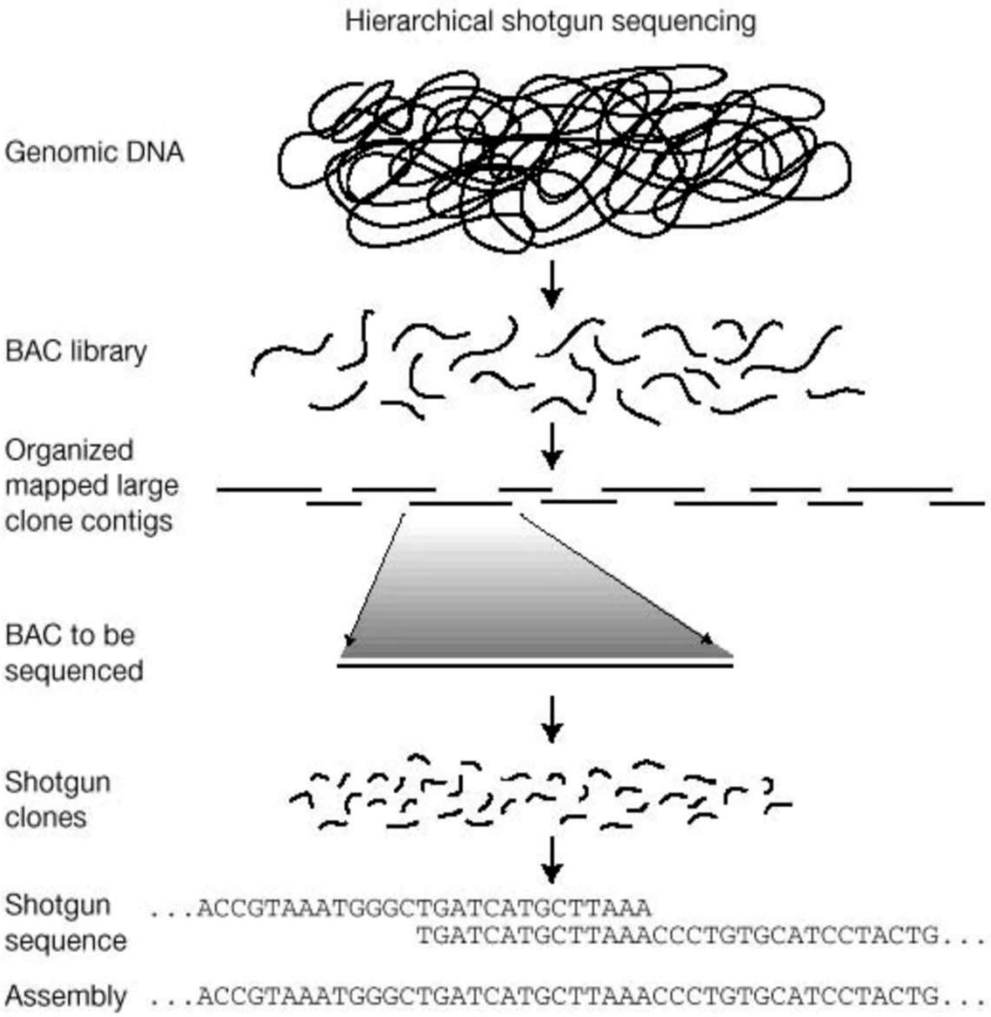
Background to the Human Genome Project



大规模基因组分析的时间线。1990 年对几种非脊椎动物模型生物（红色）、小鼠（蓝色）和人类（绿色）的选定工作组成部分。SNP，单核苷酸多态性；ESTs，表达的序列标签。

BGI华大

Background to the Human Genome Project



- 有两种方法可以对大型重复序列丰富的基因组进行测序：
- 全基因组鸟枪法测序方法
- 分层霰弹枪测序方法：
- 分层霰弹枪测序先将目标基因组片段化，克隆成大片段克隆载体，构建文库；此处显示了 BAC 载体。然后将文库中的基因组 DNA 片段组织成一个物理图谱，并通过随机鸟枪策略选择和测序单个 BAC 克隆。最后，组装克隆序列以重建基因组序列。

BGI华大

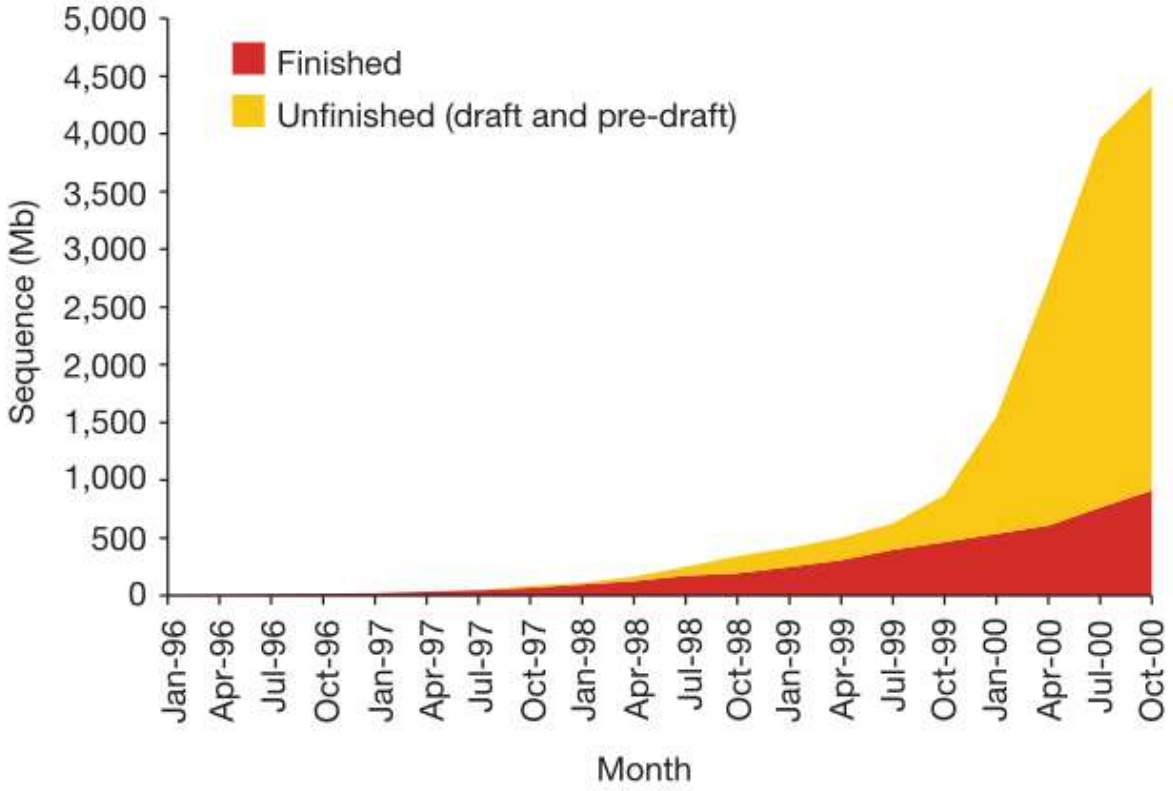
Generating the draft genome sequence

Table 1 Key large-insert genome-wide libraries

Library name*	GenBank abbreviation	Vector type	Source DNA	Library segment or plate numbers	Enzyme digest	Average insert size (kb)	Total number of clones in library	Number of fingerprinted clones†	BAC-end sequence (ends/clones/ clones with both ends sequenced)‡	Number of clones in genome layout§	Sequenced clones used in construction of the draft genome sequence		
											Number‖	Total bases (Mb)¶	Fraction of total from library
Caltech B	CTB	BAC	987SK cells	All	*Hin*dIII	120	74,496	16	2/1/1	528	518	66.7	0.016
Caltech C	CTC	BAC	Human sperm	All	*Hin*dIII	125	263,040	144	21,956/ 14,445/ 7,255	621	606	88.4	0.021
Caltech D1 (CITB-H1)	CTD	BAC	Human sperm	All	*Hin*dIII	129	162,432	49,833	403,589/ 226,068/ 156,631	1,381	1,367	185.6	0.043
Caltech D2 (CITB-E1)		BAC	Human sperm	All									
				2,501–2,565	*Eco*RI	202	24,960						
				2,566–2,671	*Eco*RI	182	46,326						
				3,000–3,253	*Eco*RI	142	97,536						
RPCI-1	RP1	PAC	Male, blood	All	*Mbo*I	110	115,200	3,388		1,070	1,053	117.7	0.028
RPCI-3	RP3	PAC	Male, blood	All	*Mbo*I	115	75,513			644	638	68.5	0.016
RPCI-4	RP4	PAC	Male, blood	All	*Mbo*I	116	105,251			889	881	95.5	0.022
RPCI-5	RP5	PAC	Male, blood	All	*Mbo*I	115	142,773			1,042	1,033	116.5	0.027
RPCI-11	RP11	BAC	Male, blood	All		178	543,797	267,931	379,773/ 243,764/ 134,110	19,405	19,145	3,165.0	0.743
				1	*Eco*RI	164	108,499						
				2	*Eco*RI	168	109,496						
				3	*Eco*RI	181	109,657						
				4	*Eco*RI	183	109,382						
				5	*Mbo*I	196	106,763						
Total of top eight libraries							1,482,502	321,312	805,320/ 484,278/ 297,997	25,580	25,241	3,903.9	0.916
Total all libraries								354,510	812,594/ 488,017/ 100,775	30,445	29,298	4,260.5	1

BGI华大

Generating the draft genome sequence



GenBank高通量基因组序列(HTGS)中人类序列的总数量

红色为已完成测序序列和黄色为未完成测序序列

BGI华大

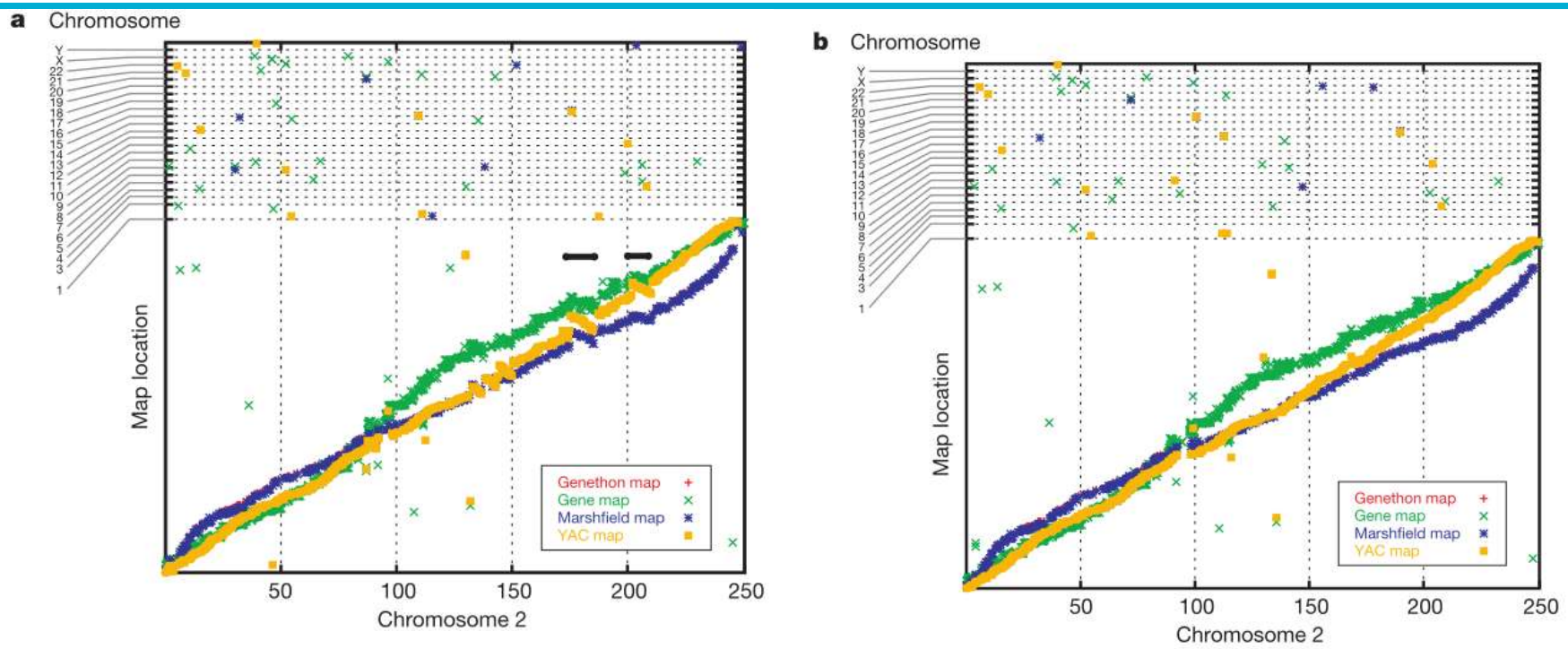
Generating the draft genome sequence

Table 4 Plasmid paired-end reads

	Total reads deposited*	Read pairs†	Size range of inserts (kb)
Random-sheared	3,227,685	1,155,284	1.8–6
Enzyme digest	2,539,222	761,010	0.8–4.7
Total	5,766,907	1,916,294	

BGI华大

Generating the draft genome sequence



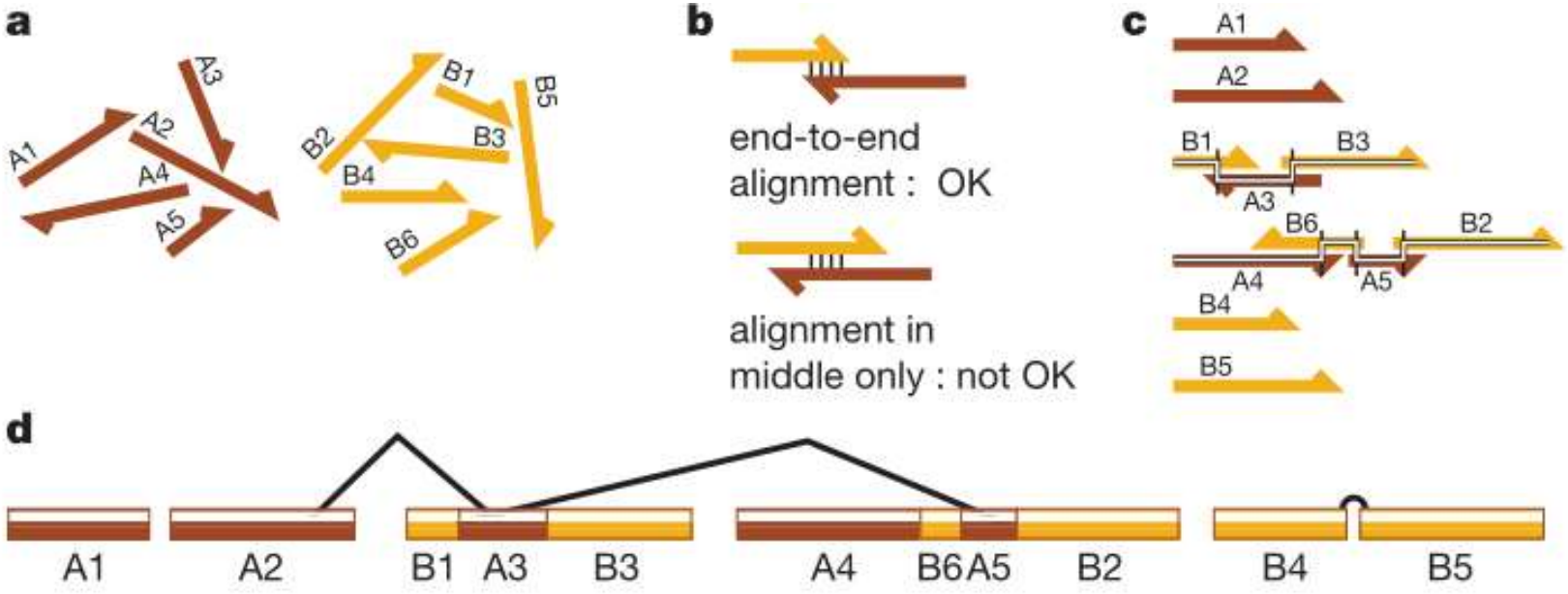
已知的基因组图谱(Genethon遗传图谱和Marshfield遗传图谱、GeneMap辐射杂交图谱和 YAC辐射杂交图谱)上标记的位置与它们在2号染色体草图序列上的位置相对照。水平单位是Mb。数据呈对角线分布，表明不同map上数据集的顺序和方向基本一致(两个遗传地图完全重叠)。

a.相对于其他map，有两个分段在早期版本草案序列中被倒置

b.在信息被用来重新定位这两个倒置的片段后，相同的染色体

BGI华大

Generating the draft genome sequence

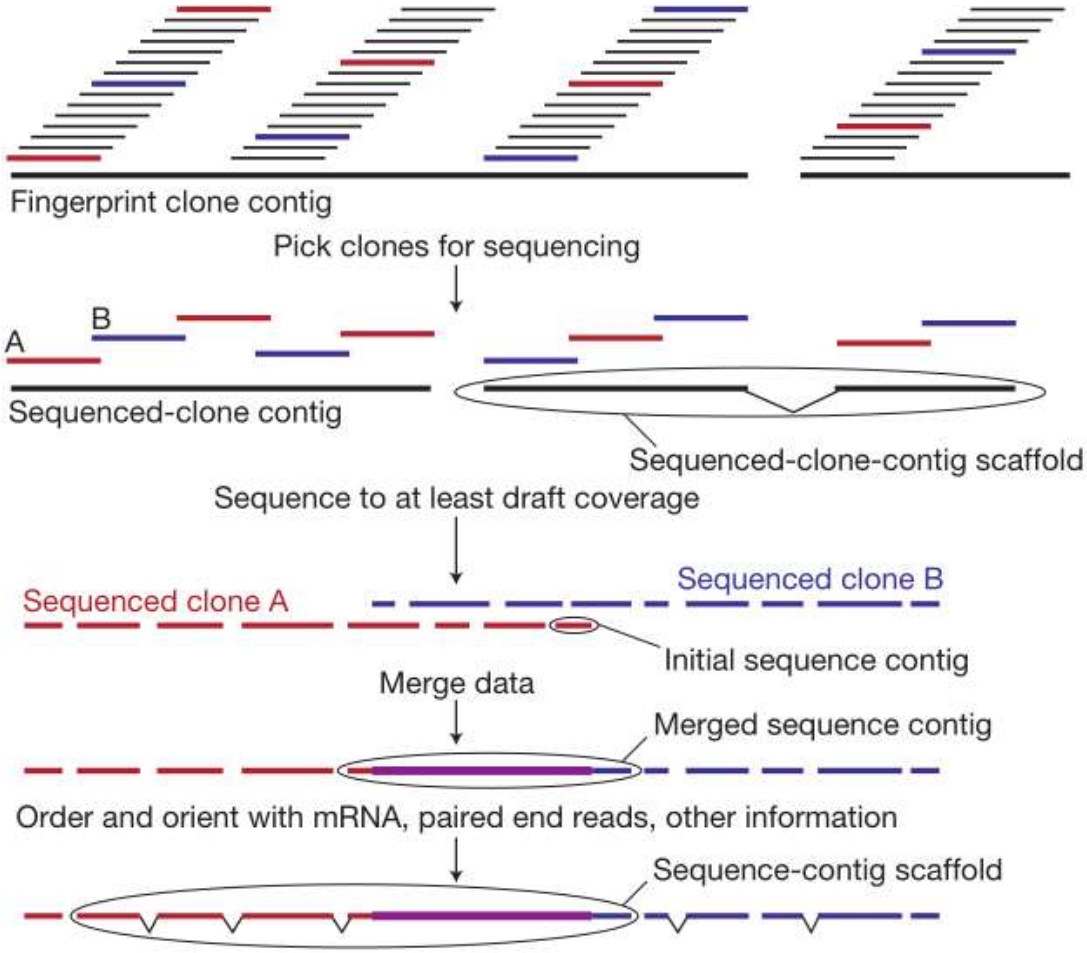


将已测序的单个克隆组装成基因组草图的关键步骤(a-d)

A1-A5为A克隆的原始序列，B1-B6为B克隆的原始序列

BGI华大

Generating the draft genome sequence

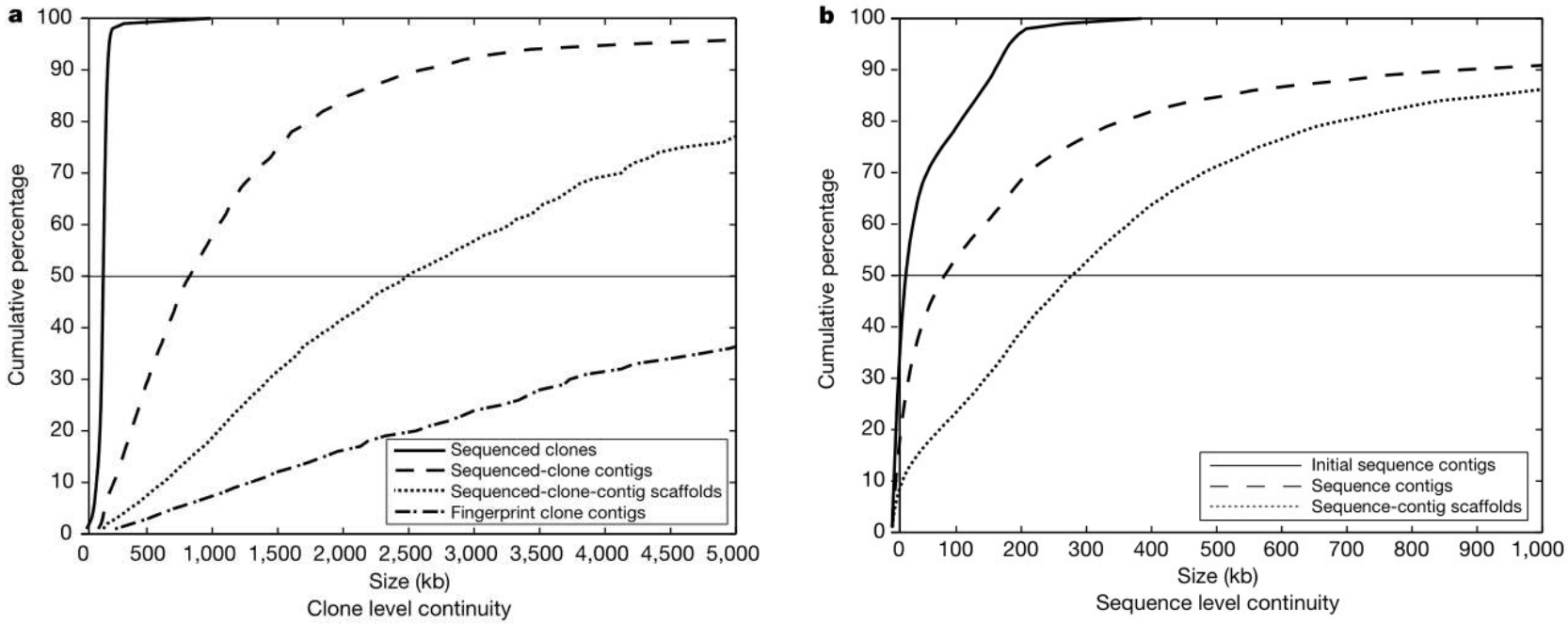


- 首先利用计算机程序FPC分析大插入克隆的限制性内切酶的酶切模式，组装“指纹克隆contig”。
- 然后选择克隆进行测序。要选择的克隆，其所有的限制性内切酶片段(除两个载体插入连接片段外)必须与contig中每一侧至少一个相邻的片段相同。
- 这些克隆被测序后，这个集合就是一个“测序-克隆contig”。当从指纹克隆contig中选择的所有克隆都测序后，已测序的克隆contig将与指纹克隆contig相同。未测序前，一个指纹克隆contig可能包含多个序列克隆contig。
- 在对单个克隆(例如，A和B)进行测序并绘制覆盖率草图后，使用GigAssembler对数据进行分析(图6)，从初始序列contigs生成合并序列contigs，并将它们连接起来形成序列-contig-scaffold。

BGI华大

Generating the draft genome sequence

N50：衡量基因组质量，当相加的长度达到Contig总长度的50%时，最后一个加上的Contig长度即为Contig N50



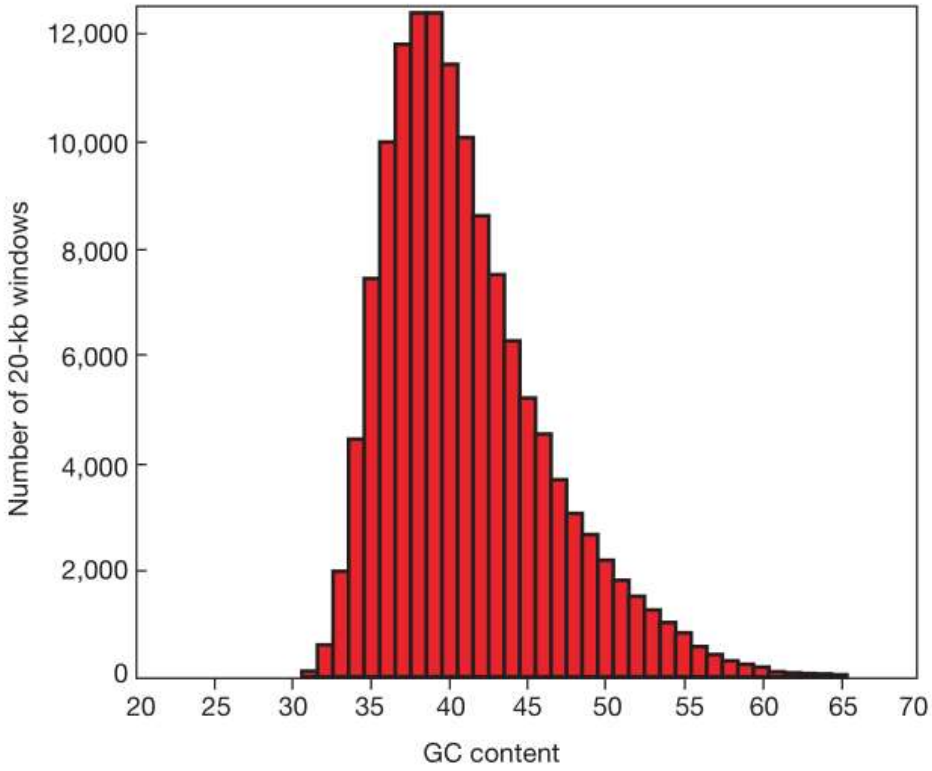
克隆水平邻连续性和序列连续性的分布，连续性逐级升高

a.克隆水平的连续性。这些克隆的大小分布很紧凑，N50为160kb。序列克隆的contigs代表了下一层次的连续性，并由mRNA序列或成对的BAC端序列连接，从而生成了序列克隆的contig scaffold

b.序列水平连续性。原始序列contig具有低连续性，N50 为21.7 kb。合并后，序列contigs的N50大约为82 kb。连接后，形成N50约为274 kb的序列scaffolds

BGI华大

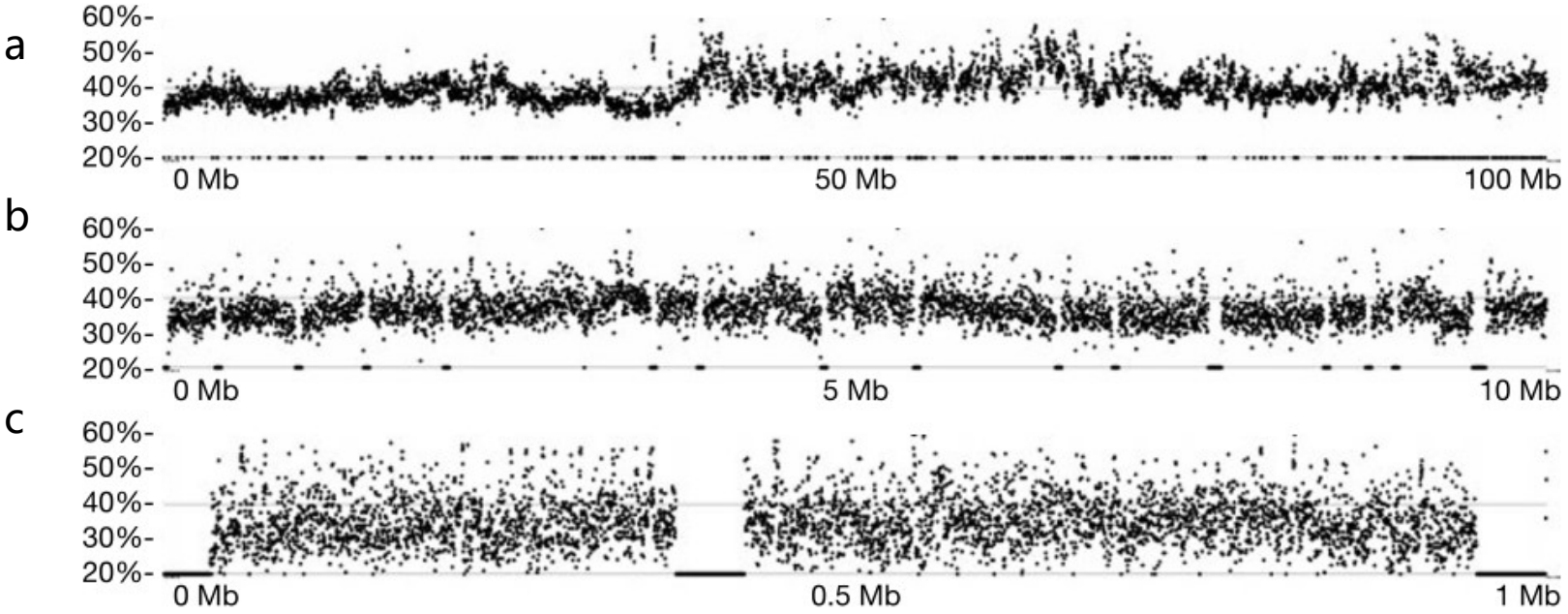
Broad genomic landscape



基因组序列草图中20kb窗口的GC含量直方图

BGI华大

Broad genomic landscape



不同尺度上GC含量的变化：

a.分析的整个100 mb区域的GC含量

b.前10 Mb的GC含量

c.前1 Mb的GC含量

BGI华大

Broad genomic landscape

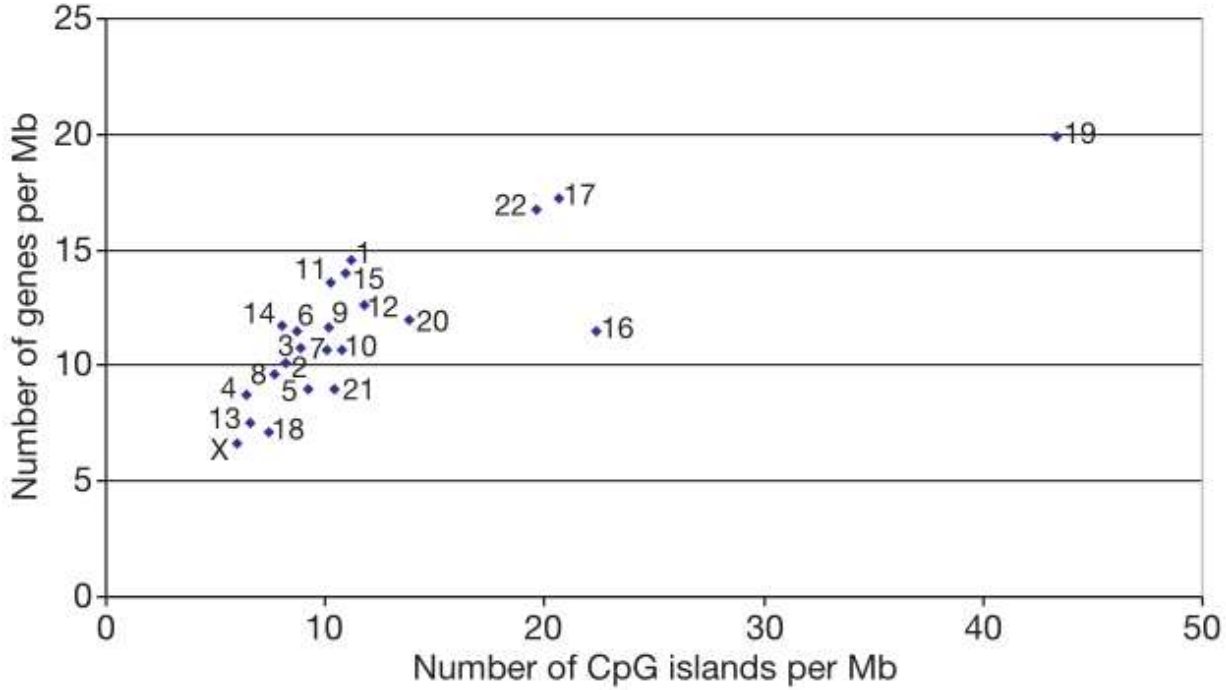
Table 10 Number of CpG islands by GC content

GC content of island	Number of islands	Percentage of islands	Nucleotides in islands	Percentage of nucleotides in islands
Total	28,890	100	19,818,547	100
>80%	22	0.08	5,916	0.03
70–80%	5,884	20	3,111,965	16
60–70%	18,779	65	13,110,924	66
50–60%	4,205	15	3,589,742	18

- 在基因组草图中，对每个二核苷酸进行评分(GC为+17，其他为-1)，并确定最大评分片段，确定潜在CpG岛
- 评估每个片段，以确定GC含量>=50%，长度>200的CpG岛数

BGI华大

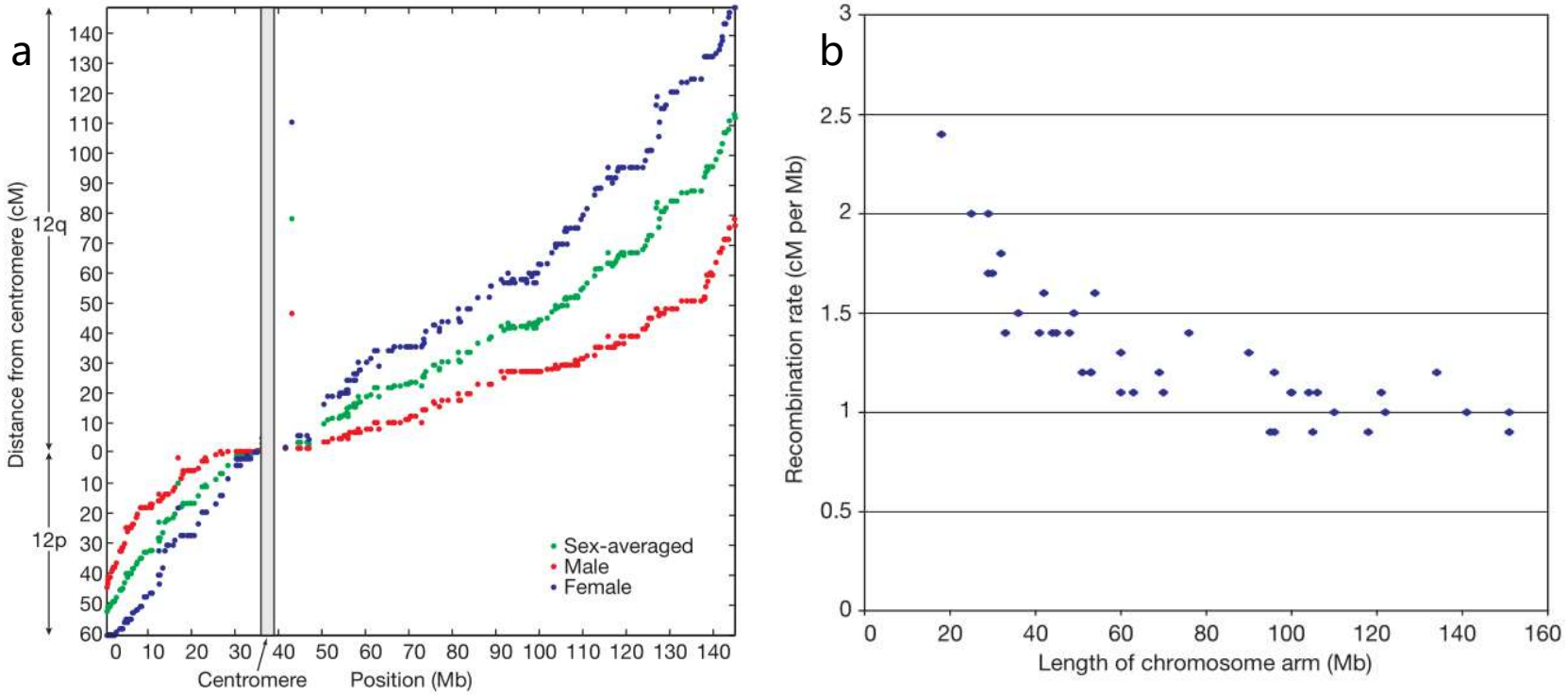
Broad genomic landscape



- 每条染色体每Mb的CpG岛数，与每Mb的基因数相比较
- 染色体16、17、22，特别是19明显的异常

BGI华大

Broad genomic landscape



a.沿着12号染色体遗传图谱的cM距离与基因组序列草图中的Mb位置相对照。染色体两端增加的斜率反映了端粒附近每Mb重组率的增加。相反，在着丝粒附近较平坦的斜率表明重组减少。

b.染色体臂的长度（Mb）和重组率（cM/Mb）的关系。随着染色体臂长度的减少，平均重组率显著上升。

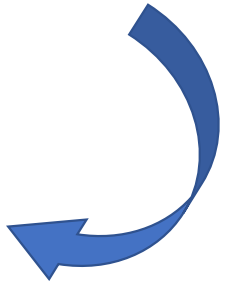
BGI华大

Repeat content of the human genome

问题：基因组大小与生物体的复杂性没有很强的相关性

重复序列

- 转座子衍生序列（散在重复序列）
- 加工假基因
- 简单重复
- 片段重复
- 串联重复



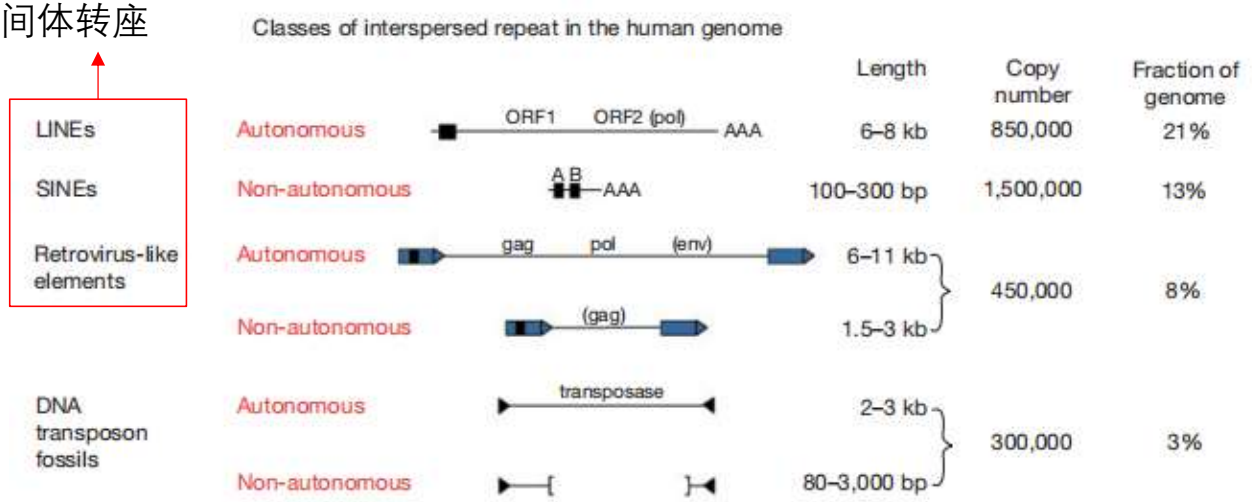
标记：作古生物学标记，提供进化事件和驱动力的线索

动因：异位重排、新基因产生与修饰、整体GC调节。影响染色体整体的结构和动力学性质。

BGI华大

散在重复序列类别

通过RNA中间体转座



检测重复序列的工具：RepeatMasker
SINE、LINE、LTR和DNA转座子约占序列全长的13%、20%、8%和3%。

LINEs：逆转录过程易中止导致非功能插入。人类只有LINE1有活性。
SINEs：没有编码蛋白的片段，包括启动子区域包括tRNA起源和7SL起源两种。SINE具有LINE的3末端并和LINEs共用一套转座机制。在人类基因组中只有7SL起源的Alu具有活性。
LTRs：包含所有必要的转录调控元件。自主元件可以编码蛋白酶、逆转录酶、Rnase H和整合酶。转座是由tRNA启动的。
DNA转座子：具有末端反向重复序列，自主型会编码一个转座酶，该转座酶在反向重复序列附近结合，并通过剪切粘贴机制介导转座。但该转座酶没有顺式偏好性，导致DNA转座子寿命较短

转座子元素采用不同的策略来确保它们的进化生存。包括水平传播和垂直传播

年龄分布

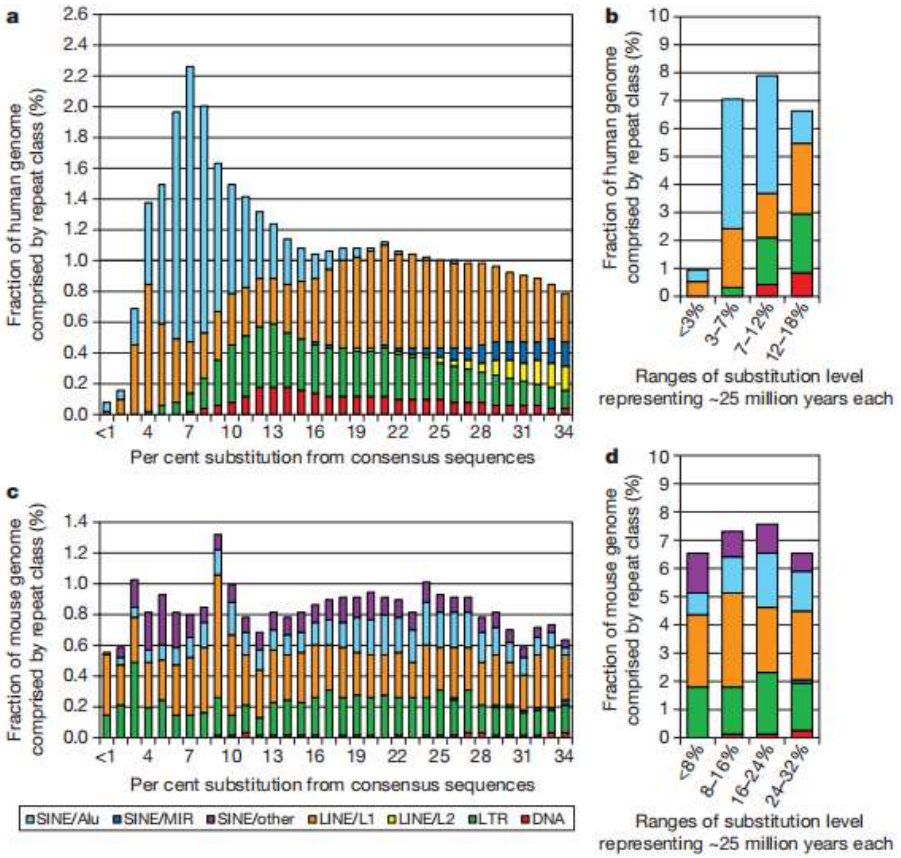


图1. 人与小鼠重复片段碱基替换水平分布

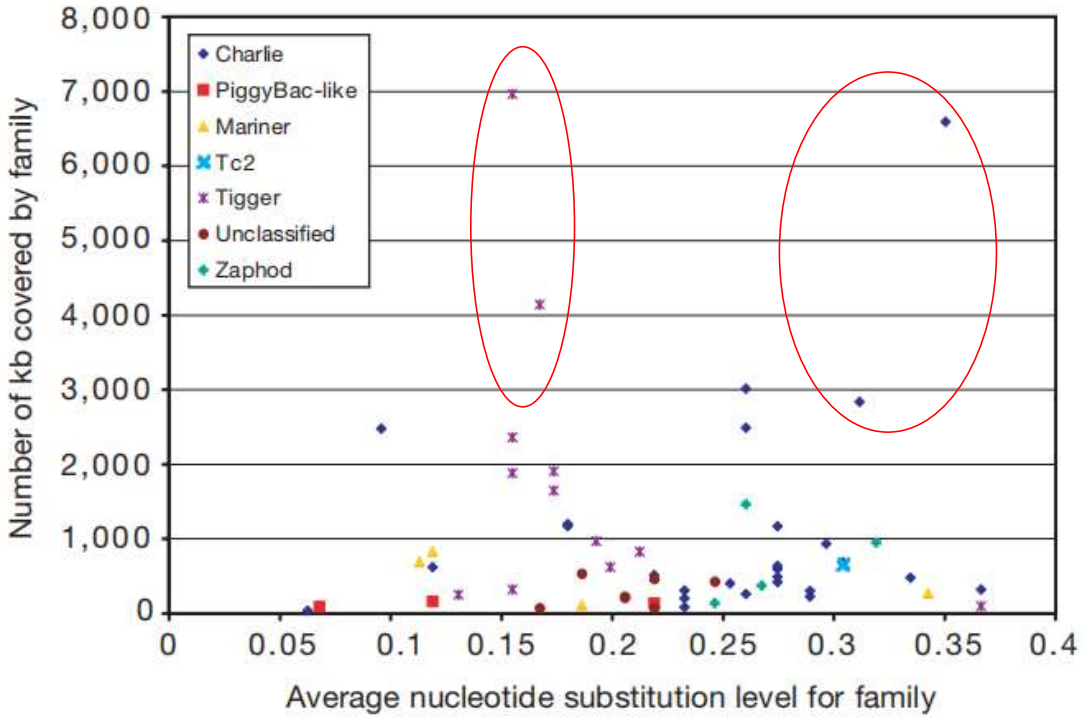


图2. 人DNA转座子碱基替换水平分布

结论：1. 从脊椎动物基因组中清除非功能序列的速度非常慢；2. LINE和SINE的寿命很长；3.DNA转座子有两个峰。由于DNA转座子可以产生大规模的染色体重排，很可能是该活动参与了物种形成事件；4. 序列草图中识别带有功能的全长LTR拷贝只有3个，可能已经快要消失；5.所有转座子的整体活性在过去5000万年的时间里显著下降

BGI华大

与其他物种的比较

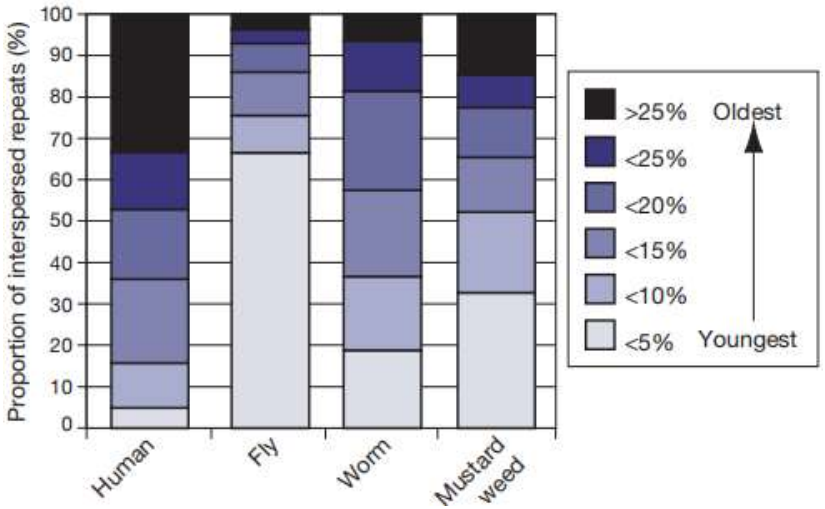


图3. 各物种重复片段年龄分布

结论：1.人类基因组散在重复片段的密度比其他三种生物体的更高；2. 人类具有更多旧重复片段的积累；3. 各转座子比例存在差异，人类相对缺乏水平传播元素可能是由于免疫系统发达；4. 人类的碱基替换速率比小鼠要低，且持久性比小鼠转座子弱

啮齿动物的种群数量较大，原始种群往往较小，可能经历频繁的瓶颈。受这些因素影响的进化力量包括**近亲繁殖**和**遗传漂变**，影响活性转座因子的持久性

Table 12 Number and nature of interspersed repeats in eukaryotic genomes

	Human		Fly		Worm		Mustard weed	
	Percentage of bases	Approximate number of families	Percentage of bases	Approximate number of families	Percentage of bases	Approximate number of families	Percentage of bases	Approximate number of families
LINE/SINE	33.40%	6	0.70%	20	0.40%	10	0.50%	10
LTR	8.10%	100	1.50%	50	0.00%	4	4.80%	70
DNA	2.80%	60	0.70%	20	5.30%	80	5.10%	80
Total	44.40%	170	3.10%	90	6.50%	90	10.50%	160

The complete genomes of fly, worm, and chromosomes 2 and 4 of mustard weed (as deposited at ncbi.nlm.nih.gov/genbank/genomes) were screened against the repeats in RepBase Update 5.02 (September 2000) with RepeatMasker at sensitive settings.

在基因组上的分布特征

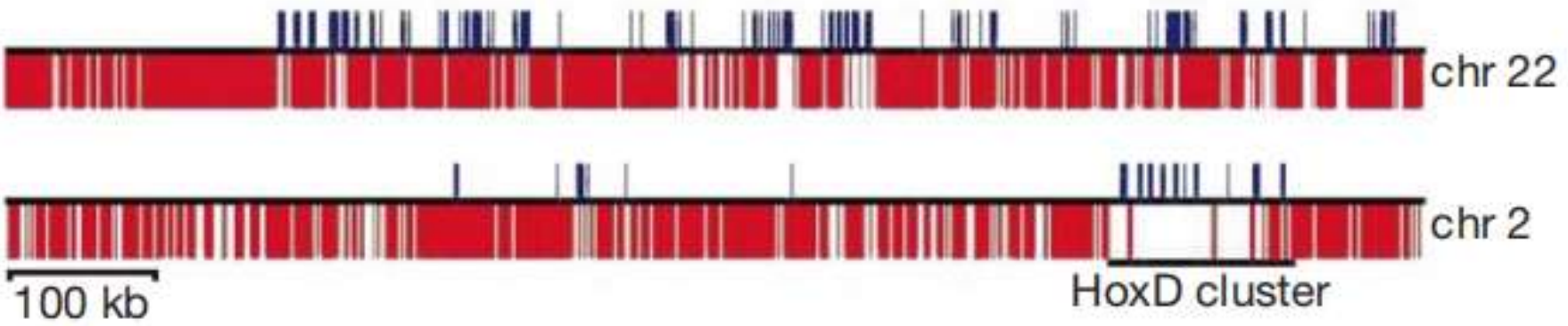


图4. 染色体上转座子密度（红色为转座子富集区域）

一些区域在重复序列中非常密集，Xp11染色体上的一个525kb的区域，整体转座子密度高达89%。同时一些区域很少转座子，如HoxA\B\C\D基因簇（图4）。密度低的区域可能有顺式调控的作用。可以重点研究进化过程中低密度区域是否能抵抗转座子插入。

BGI华大

转座子性质与GC含量相关性

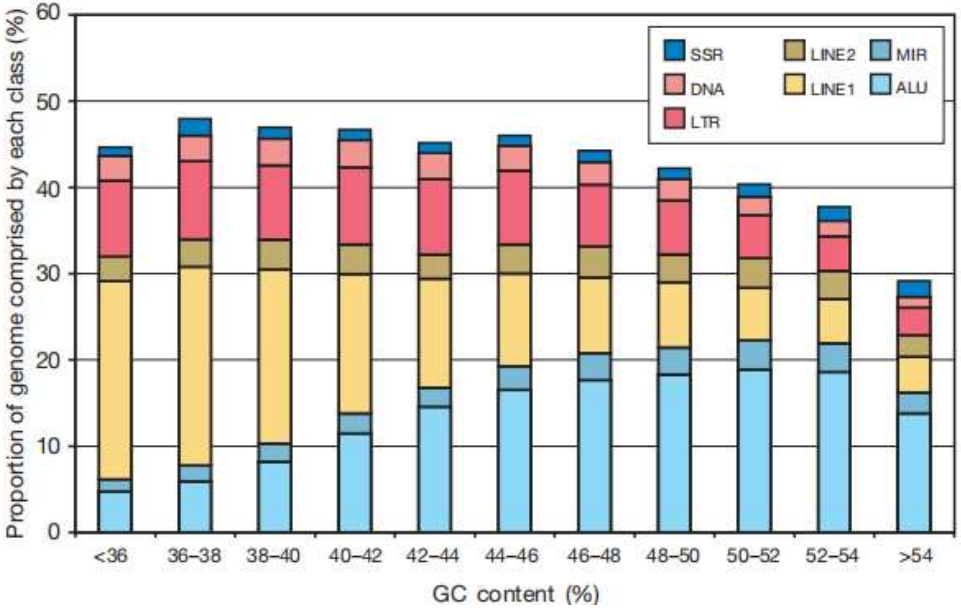


图5. 各重复片段分布与GC值之间的关系

LINE存在“寄生”的可能。对AT的偏好可能是由于富含AT的区域基因数少，减少宿主的突变负担。从机制上讲，LINE内切酶的首选切割位点是TTTT/A（斜线表示切割点），可以更高效地启动逆转录。而**SINE的分布可能存在某种进化力量的重塑**

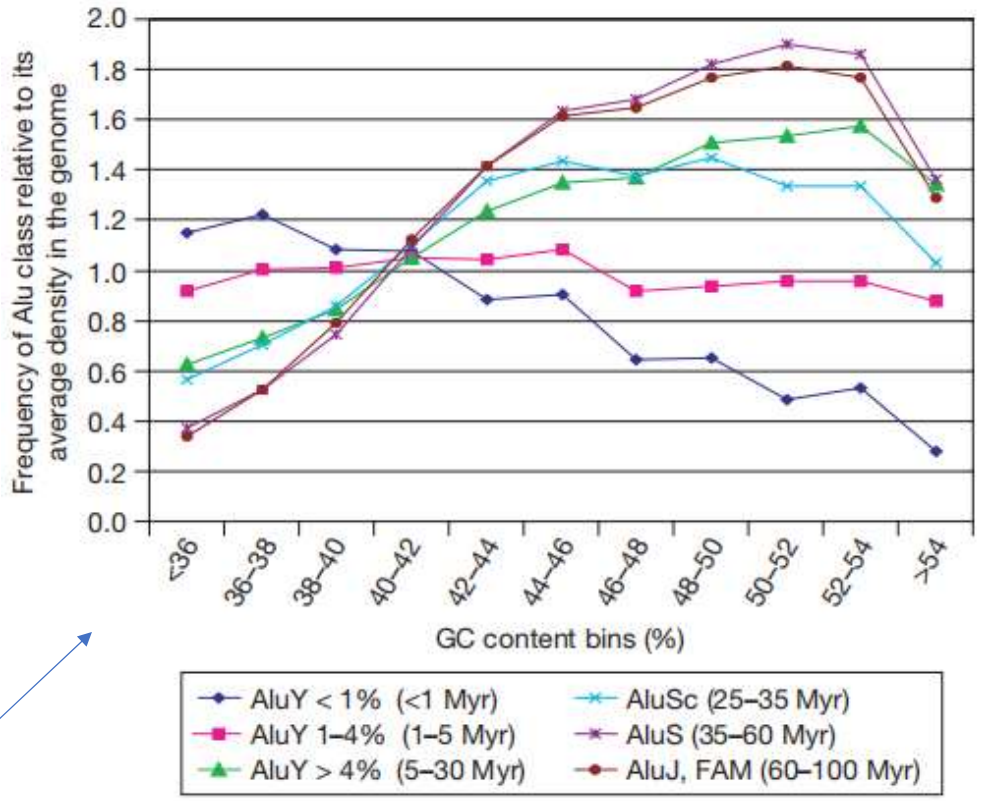


图6. 不同年龄段Alu转座子GC偏好之间的关系

BGI华大

Alu与GC含量相关性的原因

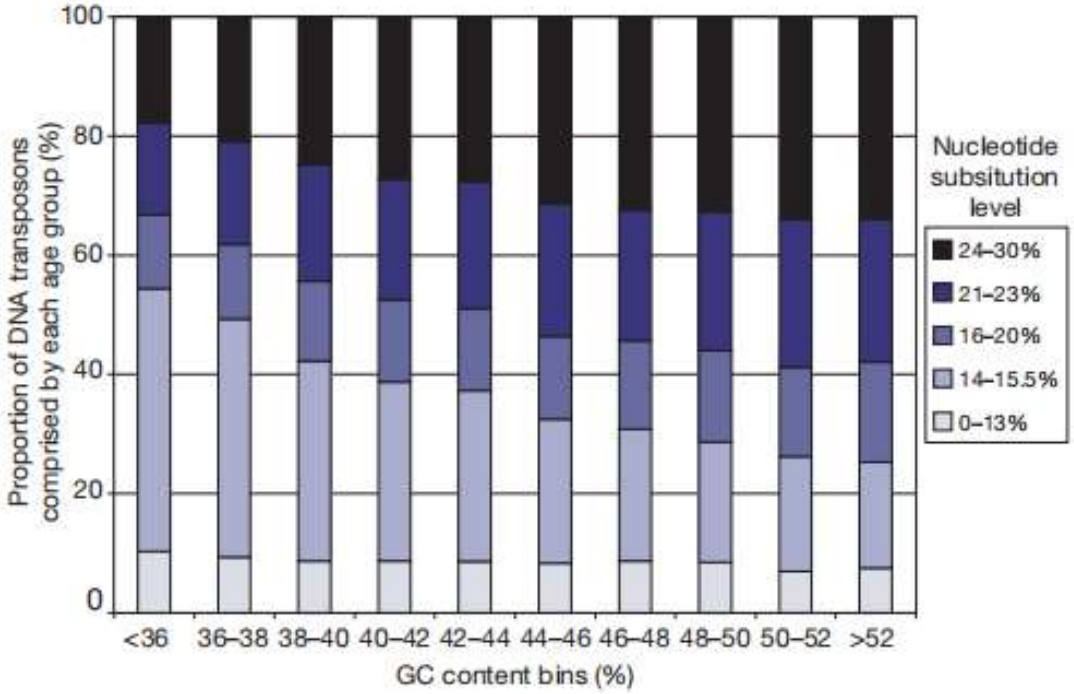


图7. 不同年龄段DNA转座子与GC偏好间的关系

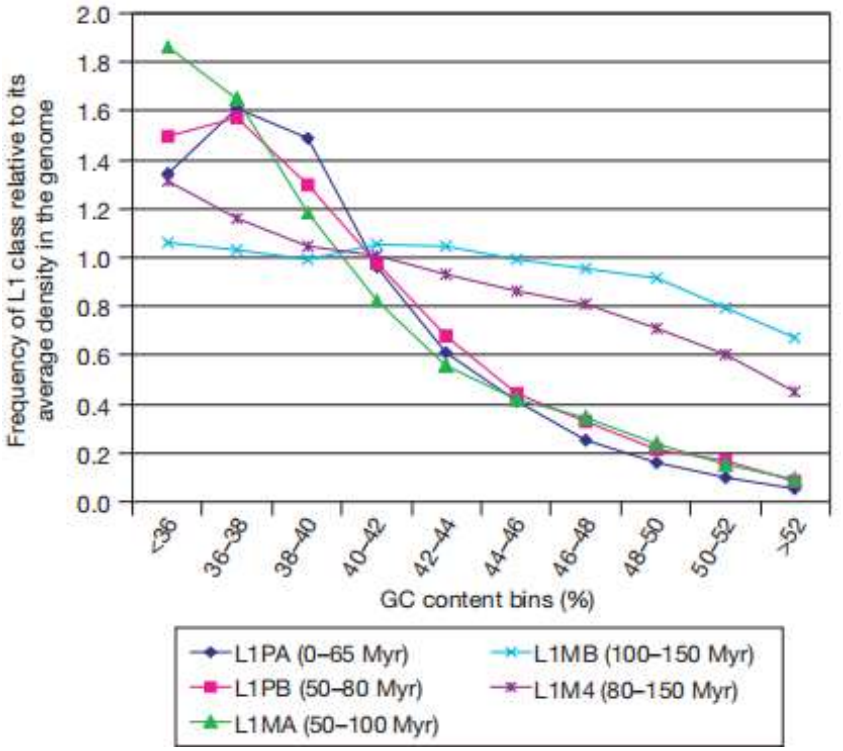


图7. 不同年龄LINE与GC偏好间的关系

解释1：DNA转座子中发现缺失在基因较少的AT富集区比在基因较多的GC富集区更容易耐受，导致年龄较大的Alu在GC富集区富集。

问题：与Alu的快速变化相比，LINE分布变化不明显。

BGI华大

Alu与GC含量相关性的原因

~~解释1：富含AT中的对Alus有阴性选择~~

~~解释2：富含AT中的Alus随机丢失率高~~

解释3：富含GC中的Alus阳性选择

假说：据观察，多个物种中SINE是在压力胁迫下转录的，在应激作用辅助蛋白翻译。因此在富含GC、容易开放转录的染色质中Alus可能有正向选择

现象与结论：19号染色体的中GC偏好更强。标准化GC含量后相对于同等GC含量来说，19号染色体上基因密度异常高，而Y染色体上异常低的体细胞活性基因密度(均相对于GC含量)有关。这并非是由GC含量本身引起的。

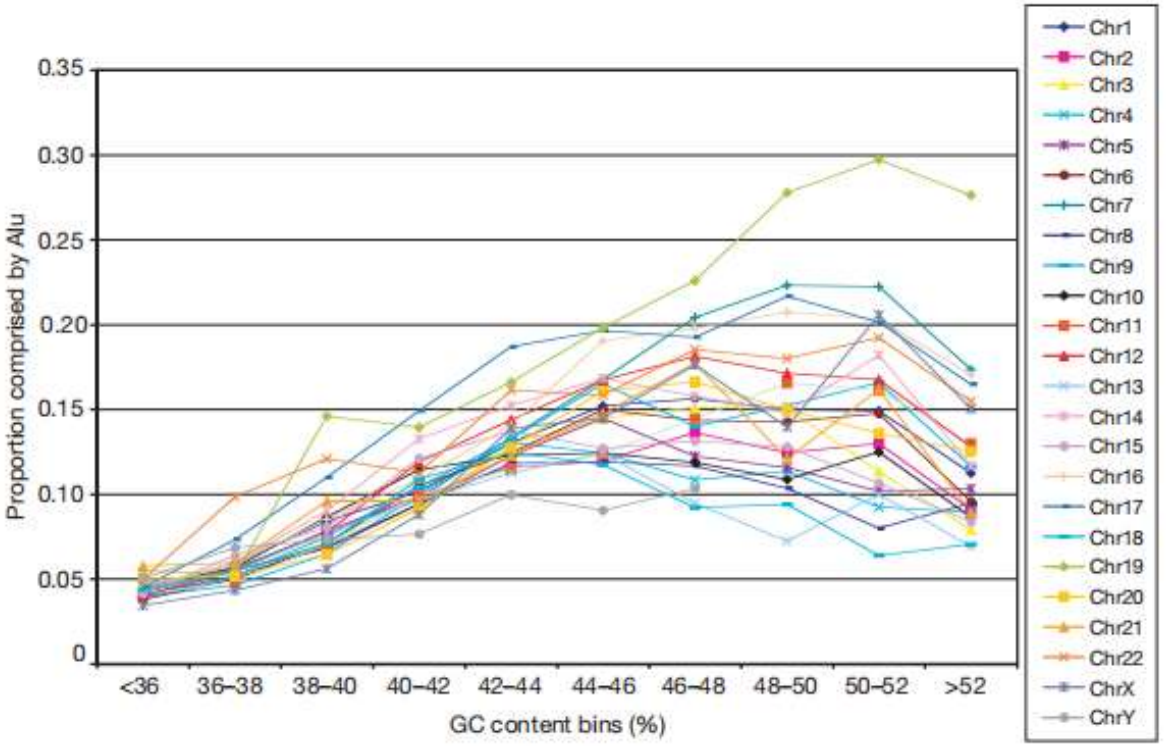


图8. 不同染色体上SINE与GC偏好间的关系

基因突变的偏差

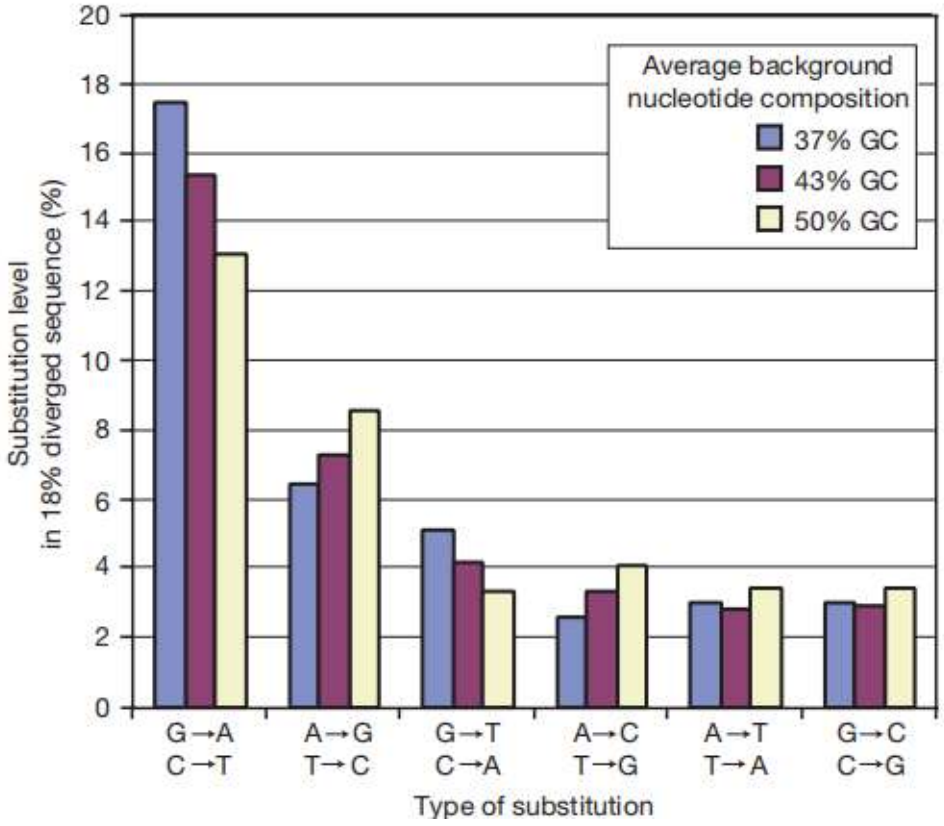


图9. DNA转座子替代模式随GC含量变化而产生的中性替代水平差异。注：转座子与一致序列的平均差异为18%，平均GC含量为43%;

1. 在替代模式中存在区域偏差，GC富集的区域更易发生GC对到AT碱基对的突变。这可能是因为GC富集区域更早复制，限制复制后期的鸟嘌呤池引起突变。也有可能跟是DNA修复机制与转录活性有关，从而与GC相关。
2. 替代模式在基因组中的数量比例也不平衡，GC含量超过了预计的平衡状态。这一方面可能是因为对编码区和CpG岛本身维持高含量GC；另一方面，新的重复片段引入导致更高的GC含量

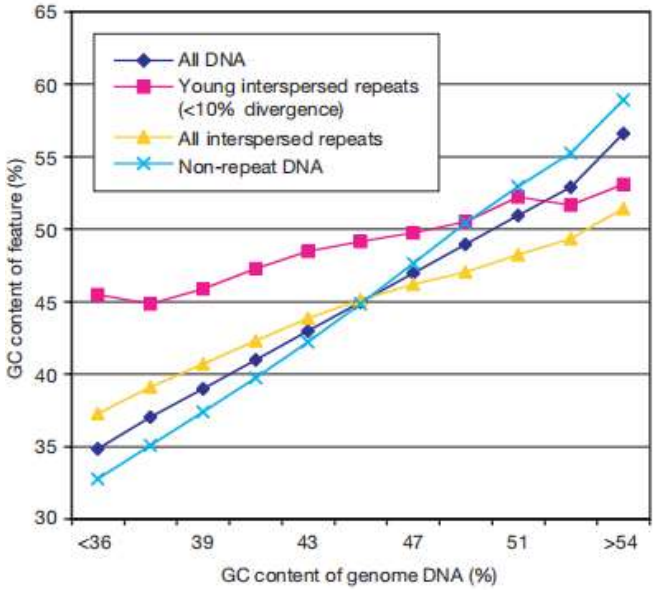


图10. 各片段GC含量的“邻居效应”

BGI华大

Y染色体的维持

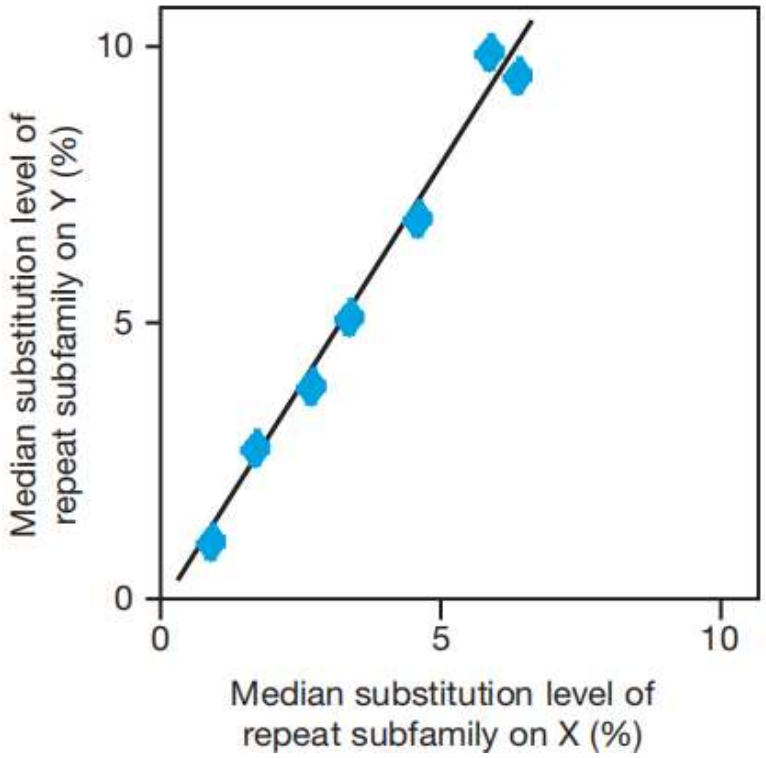


图11. Y染色体与X染色体替换水平比较

利用散在重复片段数据库，可以算出染色体上各重复片段亚家族的替代

Y染色体上较为“年轻”的散在重复片段的比例明显高于其它染色体，表明Y染色体通过快速的片段周转来维持此状态。

男性生殖系突变率较高可能是精子形成过程中细胞分裂多于卵子，以及精子和卵子中修复机制的差异。

BGI华大

转座子产生新基因

大多数转座子起源和扩展的新产物而并非对宿主的选择优势。然而，这些DNA片段会产生新元件乃新基因

Table 13 Human genes derived from transposable elements

GenBank ID*	Gene name	Related transposon family†	Possible fusion gene§	Newly recognized derivation‖
nID 3150436	BC200	FLAM Alu‡		
pID 2330017	Telomerase	non-LTR retrotransposon		
pID 1196425	HERV-3 env	Retroviridae/HERV-R‡		
pID 4773880	Syncytin	Retroviridae/HERV-W‡		
pID 131827	RAG1 and 2	Tc1-like		
pID 29863	CENP-B	Tc1/Pogo		
EST 2529718		Tc1/Pogo		+
PID 10047247		Tc1/Pogo/Pogo		+
EST 4524463		Tc1/Pogo/Pogo		+
pID 4504807	Jerky	Tc1/Pogo/Tigger		
pID 7513096	JRKL	Tc1/Pogo/Tigger		
EST 5112721		Tc1/Pogo/Tigger		+
EST 11097233		Tc1/Pogo/Tigger		+
EST 6986275	Sancho	Tc1/Pogo/Tigger		
EST 8616450		Tc1/Pogo/Tigger		+
EST 8750408		Tc1/Pogo/Tigger		+
EST 5177004		Tc1/Pogo/Tigger		+
PID 3413884	KIAA0461	Tc1/Pogo/Tc2	+	
PID 7959287	KIAA1513	Tc1/Pogo/Tc2		+
PID 2231380		Tc1/Mariner/Hsmar1‡	+	
EST 10219887		hAT/Hobo	+	+
PID 6581095	Buster1	hAT/Charlie	+	
PID 7243087	Buster2	hAT/Charlie	+	
PID 6581097	Buster3	hAT/Charlie		
PID 7662294	KIAA0766	hAT/Charlie	+	
PID 10439678		hAT/Charlie		+
PID 7243087	KIAA1353	hAT/Charlie		+
PID 7021900		hAT/Charlie/Charlie3‡		+
PID 4263748		hAT/Charlie/Charlie8‡	+	
EST 8161741		hAT/Charlie/Charlie9‡		+
pID 4758872	DAP4,pP52[rPK]	hAT/Tip100/Zaphod		
EST 10990063		hAT/Tip100/Zaphod		+
EST 10101591		hAT/Tip100/Zaphod		+
pID 7513011	KIAA0543	hAT/Tip100/Tip100	+	
pID 10439744		hAT/Tip100/Tip100		+
pID 10047247	KIAA1586	hAT/Tip100/Tip100		+
pID 10439762		hAT/Tip100	+	+
EST 10459804		hAT/Tip100		+

图12. 起源于转座子的基因

已发现47个人类基因来自于转座子。这些基因可能早期是转座子组件的一部分，然后慢慢衍生成为新基因。

LINE1机制也可以导致基因mRNA的逆转录

简单序列重复

Table 14 SSR content of the human genome

Length of repeat unit	Average bases per Mb	Average number of SSR elements per Mb
1	1,660	36.7
2	5,046	43.1
3	1,013	11.8
4	3,383	32.5
5	2,686	17.6
6	1,376	15.2
7	906	8.4
8	1,139	11.1
9	900	8.6
10	1,576	8.6
11	770	8.7

SSRs were identified by using the computer program Tandem Repeat Finder with the following parameters: match score 2, mismatch score 3, indel 5, minimum alignment 50, maximum repeat length 500, minimum repeat length 1.

图13. 人类基因组中SSR含量

Table 15 SSRs by repeat unit

Repeat unit	Number of SSRs per Mb
AC	27.7
AT	19.4
AG	8.2
GC	0.1
AAT	4.1
AAC	2.6
AGG	1.5
AAG	1.4
ATG	0.7
CGG	0.6
ACC	0.4
AGC	0.3
ACT	0.2
ACG	0.0

SSRs were identified as in Table 14.

图14 不同重复单元的SSR

SSR可以作为基因marker，提供一个全面SSR目录对人类遗传学研究有重大意义

BGI华大

片段重复

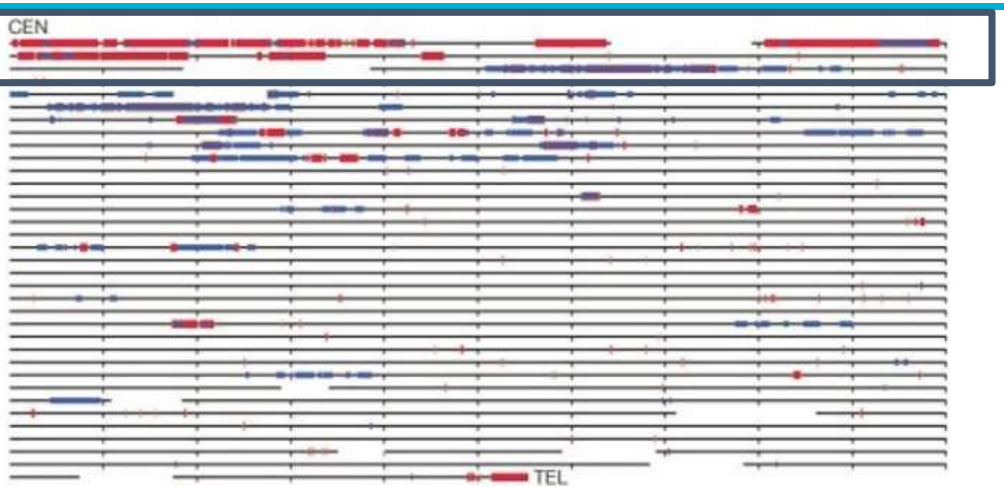


图15 21号染色体的片段重复图谱

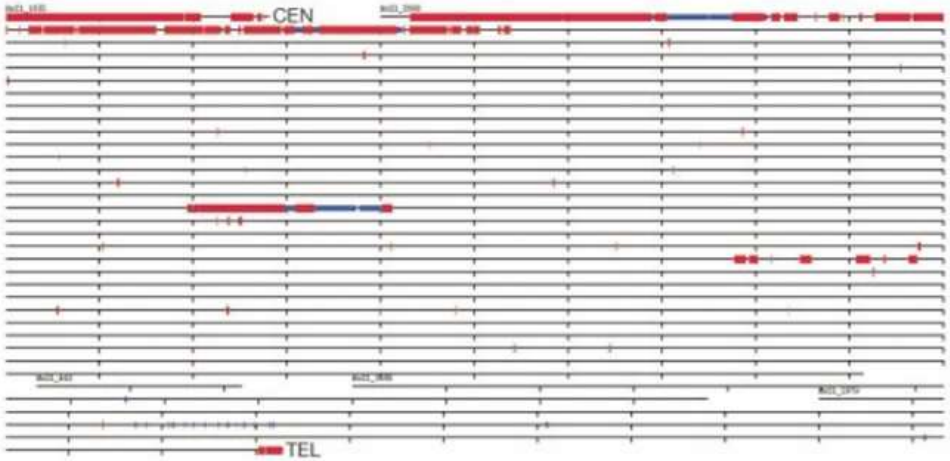


图16 22号染色体的片段重复图谱

Table 17 Fraction of the draft genome sequence in inter- and intrachromosomal duplications

Chromosome	Intrachromosomal (%)	Interchromosomal (%)	All (%)
1	2.1	1.7	3.4
2	1.6	1.6	2.6
3	1.8	1.4	2.7
4	1.5	2.2	3.0
5	1.0	0.9	1.8
6	1.5	1.4	2.7
7	3.6	1.8	4.5
8	1.2	1.5	2.1
9	2.1	2.3	3.8
10	3.3	2.0	4.7
11	2.7	1.4	3.7
12	2.1	1.2	2.8
13	1.7	1.6	3.0
14	0.6	0.6	1.2
15	4.1	4.4	6.7
16	3.4	3.4	5.5
17	4.4	1.7	5.7
18	0.9	1.0	1.9
19	5.4	1.6	6.3
20	0.8	1.4	2.0
21	1.9	4.0	4.8
22	6.8	7.7	11.9
X	1.2	1.1	2.2
Y	10.9	13.1	20.8
NA	2.3	7.8	8.3
UL	11.6	20.8	22.2
Total	2.3	2.0	3.6

图17 已完成序列染色体上片段重复的统计。注：设置上界以排除由于装配过程中不完全合并而造成的人工复制。

BGI华大

片段重复

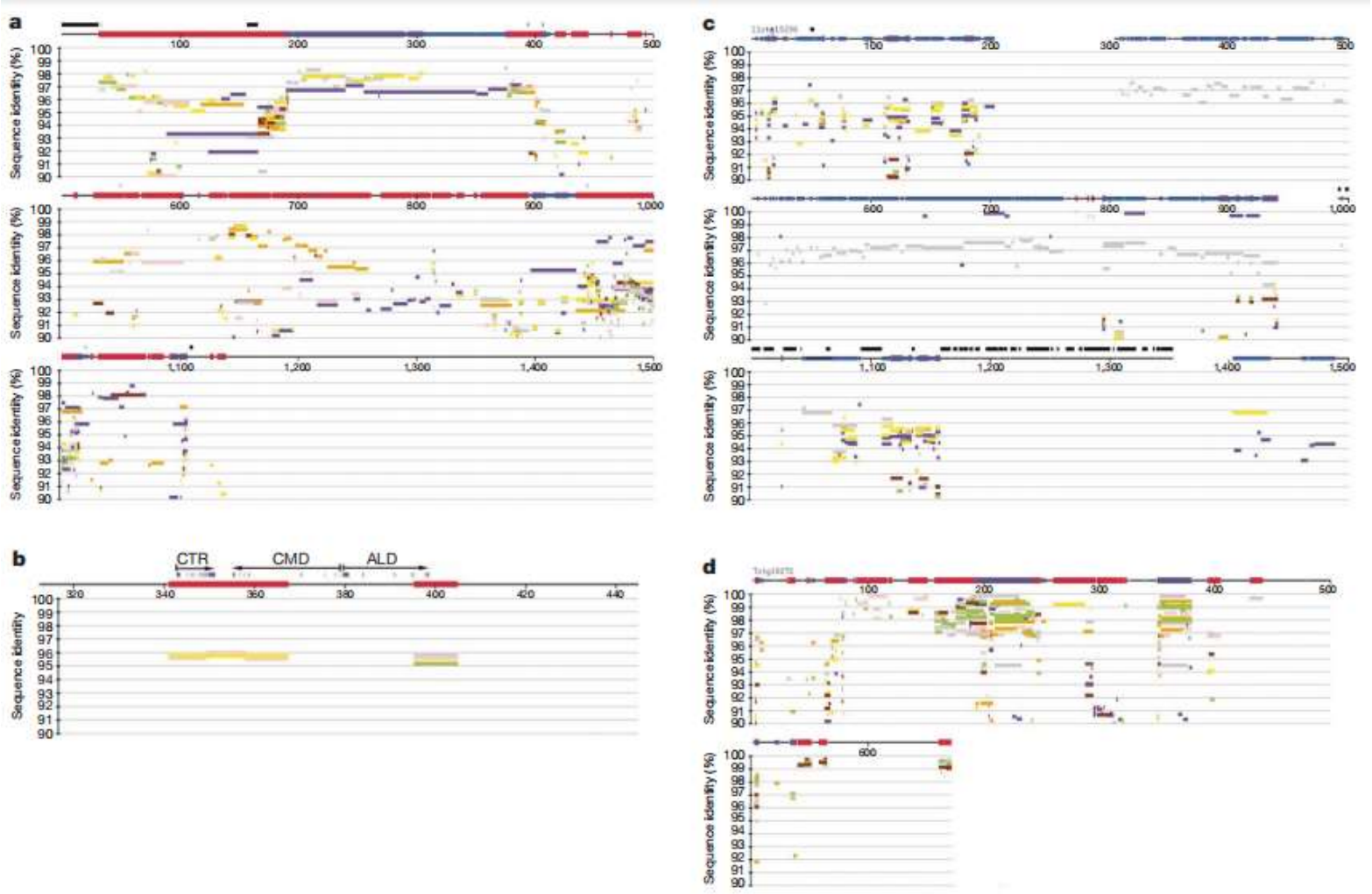


图18 22号染色体的片段重复图谱

a. 21号染色体的着丝粒区域，片段重复的保守程度较高。中间间隔一些简单重复，对靶向这些区域的复制有引导作用

b. 起源于Xq28的某个位点。该位点在一条染色体的着丝粒周围区域发生初始片段复制，然后作为一个更大的盒分配到其它区域

c. 11号染色体着丝粒的例子

d. 7号染色体的端粒区

解释：可能是因为染色体断裂的损伤修复机制会优先插入特点区域，而并非耐受性（有一些基因较少的区域也没有发现片段重复）

片段重复

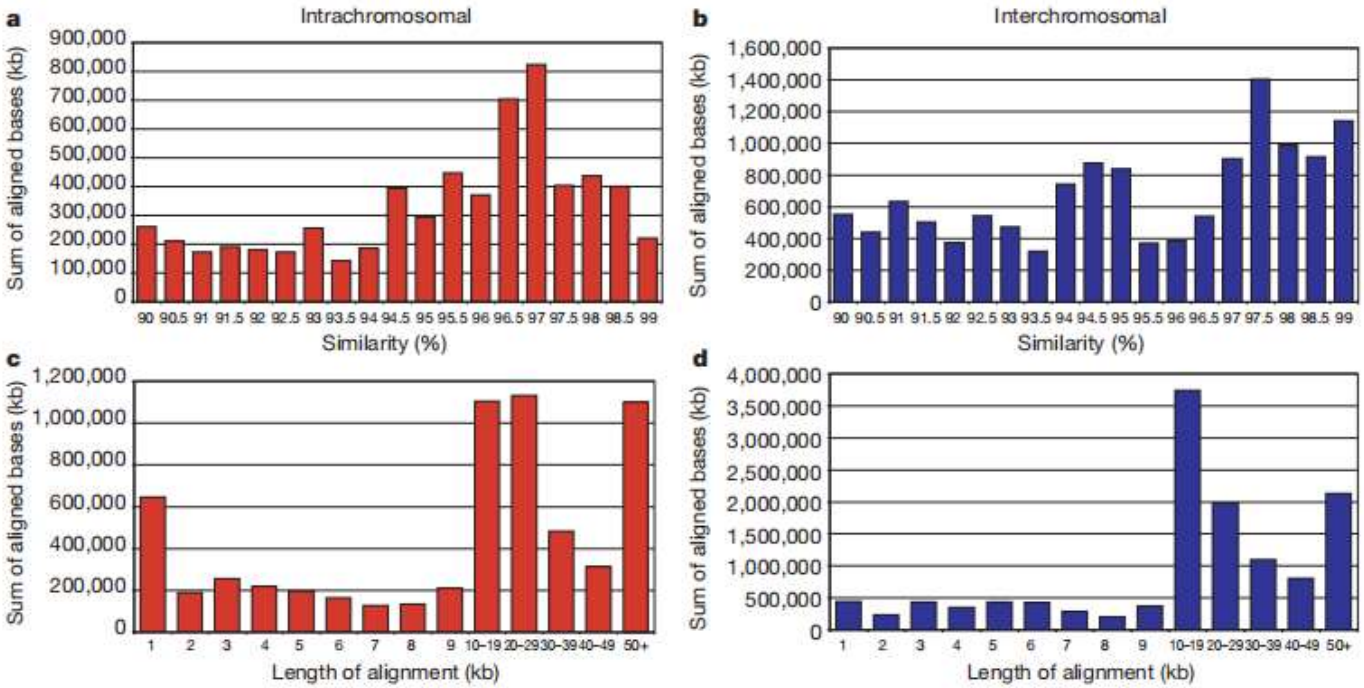


图19. 重复片段的序列特性

关键问题：序列高度同源性

1. **缺少STS和指纹的情况下如何区分同源序列和等位基因呢？**
2. **区域包含独特的外显子+内含子结构，又如何构建全基因组SNP图谱呢？**
3. **如何将全基因组复制事件与脊椎动物基因组进化动力联系起来？**

BGI华大

Gene content of the human genome

- 基因（或者说编码区）只占人类DNA的一小部分，但这些区段承担起了大部分生物功能
- 本项目的计划是列出所有人类的基因，但是实际操作上非常困难因为人类基因有有小的外显子（150bp左右）和巨大的内含子（有的甚至有10kb

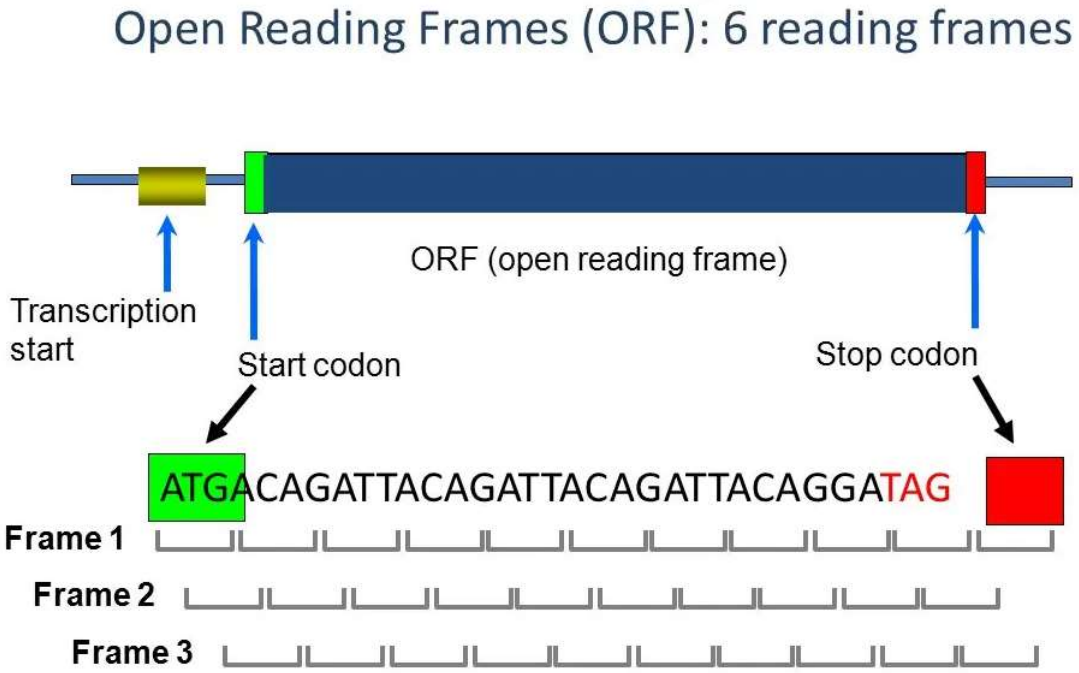
BGI华大

Gene content of the human genome

- Noncoding RNAs (ncRNAs)主要包含:
  - tRNAs
  - rRNAs
  - small nucleolar RNAs（snoRNAs），参与rRNA的生成和核仁内的碱基修饰
  - small nuclear RNAs（snRNAs），剪接体的重要构成部分，在核仁中负责剪切mRNA前体的内含子，分U2和U12两种
  - 端粒RNA、7SL信号识别RNA等
  - 功能不明的其余RNA

BGI华大

Gene content of the human genome



ncRNA没有被明确的ORF，因此ncRNAs无法通过设计算法找到，即使整个人类基因组都已经被解析，找到ncRNA依然很难。

方法是将已有的ncRNA与序列进行比对（blastn）

BGI华大

Gene content of the human genome

Table 19 Number of tRNA genes in various organisms

Organism	Number of canonical tRNAs	SeCys tRNA
Human	497	1
Worm	584	1
Fly	284	1
Yeast	273	0
Methanococcus jannaschii	36	1
Escherichia coli	86	1

- tRNA：
- 已有的实验方法预测人类tRNA基因有1310个，但在草稿基因组里只找到497个tRNA基因。
- 草稿基因组中找到了38个已知的tRNA中的37个
- tRNA的数量也许不直接和生物复杂度相关联，而是和各自物种的需求有关

BGI华大

Gene content of the human genome

- rRNA:
- 一般来说分为几类:
  - 28S rRNA （大亚基)
  - 5.8S rRNA（大亚基)
  - 5S rRNA （大亚基)
  - 18S rRNA（小亚基)

  - 在基因组中以44kb长度的一大段重复序列出现，大约有150-200个拷贝分布在染色体13、14、15、21和22上。

BGI华大

Gene content of the human genome

- small nucleolar RNAs（snoRNAs）：
- 主要负责对rRNA添加碱基修饰
- 分为两类：
  - C/D box snoRNAs，负责2 ‘-O-核糖甲基化
  - H/ACA snoRNAs，负责引导位置特异性的假酸化（pseudouridylations）
  - 草稿基因组中找到了84/97个snoRNA，或许还有更多的snoRNA因为blast算法的原因没有被找到

BGI华大

Gene content of the human genome

Table 20 Known non-coding RNA genes in the draft genome sequence

RNA gene*	Number expected†	Number found‡	Number of related genes§	Function
tRNA	1,310	497	324	Protein synthesis
SSU (18S) rRNA	150–200	0	40	Protein synthesis
5.8S rRNA	150–200	1	11	Protein synthesis
LSU (28S) rRNA	150–200	0	181	Protein synthesis
5S rRNA	200–300	4	520	Protein synthesis
U1	~30	16	134	Spliceosome component
U2	10–20	6	94	Spliceosome component
U4	??	4	87	Spliceosome component
U4atac	??	1	20	Component of minor (U11/U12) spliceosome
U5	??	1	31	Spliceosome component
U6	??	44	1,135	Spliceosome component
U6atac	??	4	32	Component of minor (U11/U12) spliceosome
U7	1	1	3	Histone mRNA 3′ processing
U11	1	0	6	Component of minor (U11/U12) spliceosome
U12	1	1	0	Component of minor (U11/U12) spliceosome
SRP (7SL) RNA	4	3	773	Component of signal recognition particle (protein secretion)
RNAse P	1	1	2	tRNA 5′ end processing
RNAse MRP	1	1	6	rRNA processing
Telomerase RNA	1	1	4	Template for addition of telomeres
hY1	1	1	353	Component of Ro RNP, function unknown
hY3	1	25	414	Component of Ro RNP, function unknown
hY4	1	3	115	Component of Ro RNP, function unknown
hY5 (4.5S RNA)	1	1	9	Component of Ro RNP, function unknown
Vault RNAs	3	3	1	Component of 13-MDa vault RNP, function unknown
7SK	1	1	330	Unknown
H19	1	1	2	Unknown
Xist	1	1	0	Initiation of X chromosome inactivation (dosage compensation)
Known C/D snoRNAs	81	69	558	Pre-rRNA processing or site-specific ribose methylation of rRNA
Known H/ACA snoRNAs	16	15	87	Pre-rRNA processing or site-specific pseudouridylation of rRNA

BGI华大

Gene content of the human genome

- Protein-coding genes:
- 最重要的一部分工作就是对蛋白编码基因进行搜索和鉴定
- 通过将已知基因的cDNA比对到草稿基因组上可以实现，但实际上只有1/4的基因可以通过这种方法鉴定
- 已知基因数据都放在RefSeq数据库上，作者们将refseq里的mRNA比对到草稿基因组上进行比对，定位了5364/10272个基因

BGI华大

Gene content of the human genome

Table 21 Characteristics of human genes

	Median	Mean	Sample (size)
Internal exon	122 bp	145 bp	RefSeq alignments to draft genome sequence, with confirmed intron boundaries (43,317 exons)
Exon number	7	8.8	RefSeq alignments to finished sequence (3,501 genes)
Introns	1,023 bp	3,365 bp	RefSeq alignments to finished sequence (27,238 introns)
3′ UTR	400 bp	770 bp	Confirmed by mRNA or EST on chromosome 22 (689)
5′ UTR	240 bp	300 bp	Confirmed by mRNA or EST on chromosome 22 (463)
Coding sequence	1,100 bp	1,340 bp	Selected RefSeq entries (1,804)
(CDS)	367 aa	447 aa	
Genomic extent	14 kb	27 kb	Selected RefSeq entries (1,804)

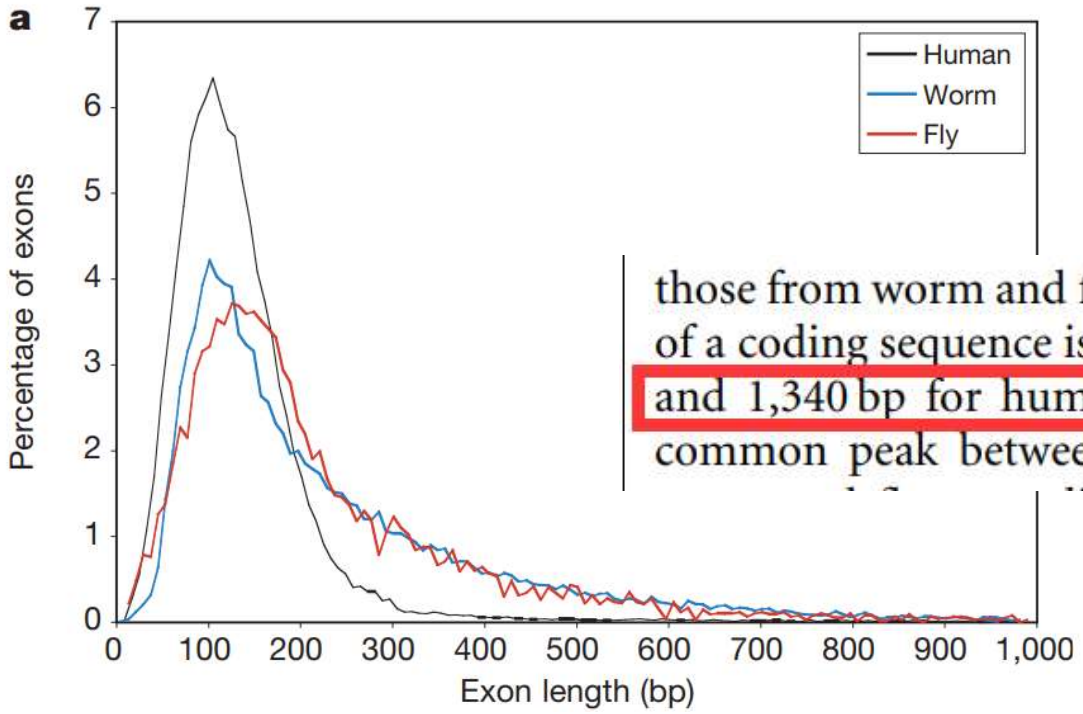
Median and mean values for a number of properties of human protein-coding genes. The 1,804 selected RefSeq entries were those that could be unambiguously aligned to finished sequence over their entire length.

草稿基因组的数据对基因结构也有新的展示

BGI华大

Gene content of the human genome

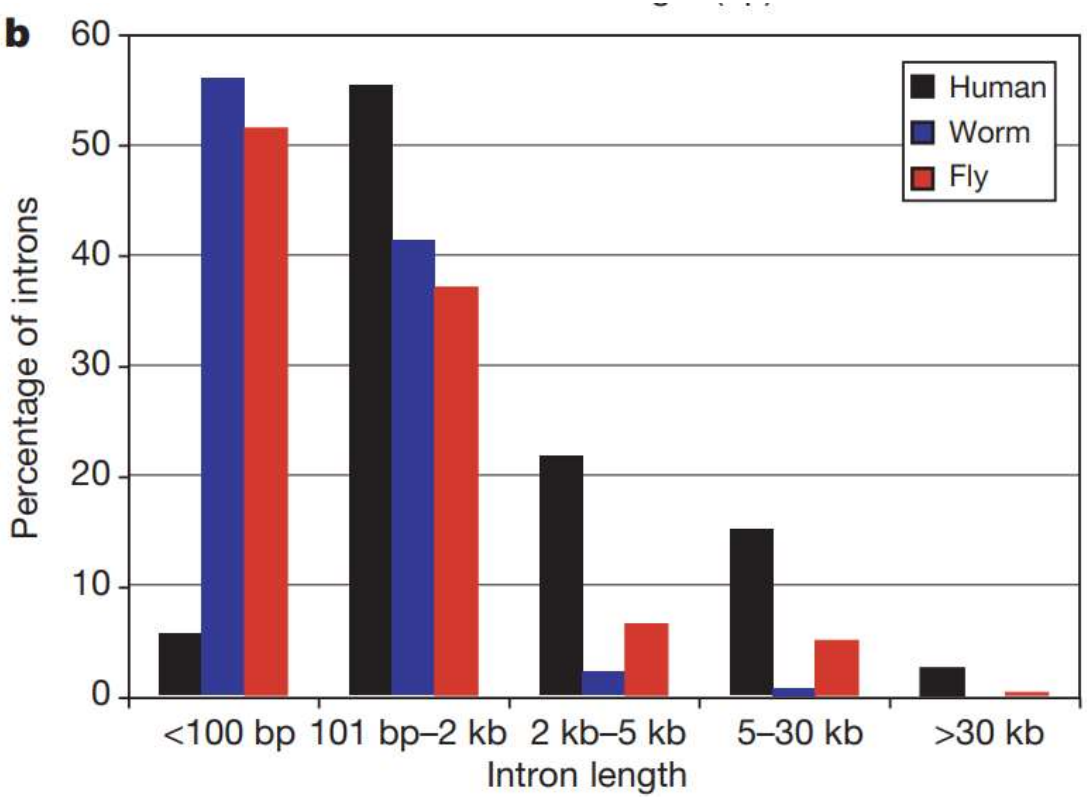
与其他基因组（蠕虫、苍蝇）的比较

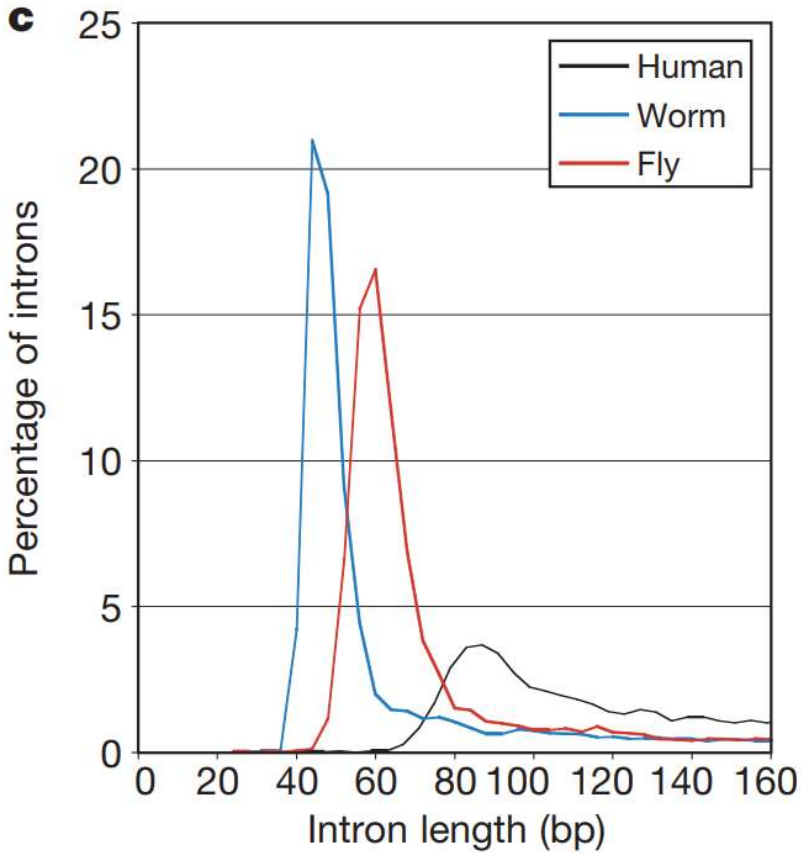


those from worm and fly. For all three organisms, the typical length of a coding sequence is simila (1,311 bp for worm, 1,497 bp for fly and 1,340 bp for human) and most internal exons fall within a common peak between 50 and 200 bp (Fig. 35a). However, the

BGI华大

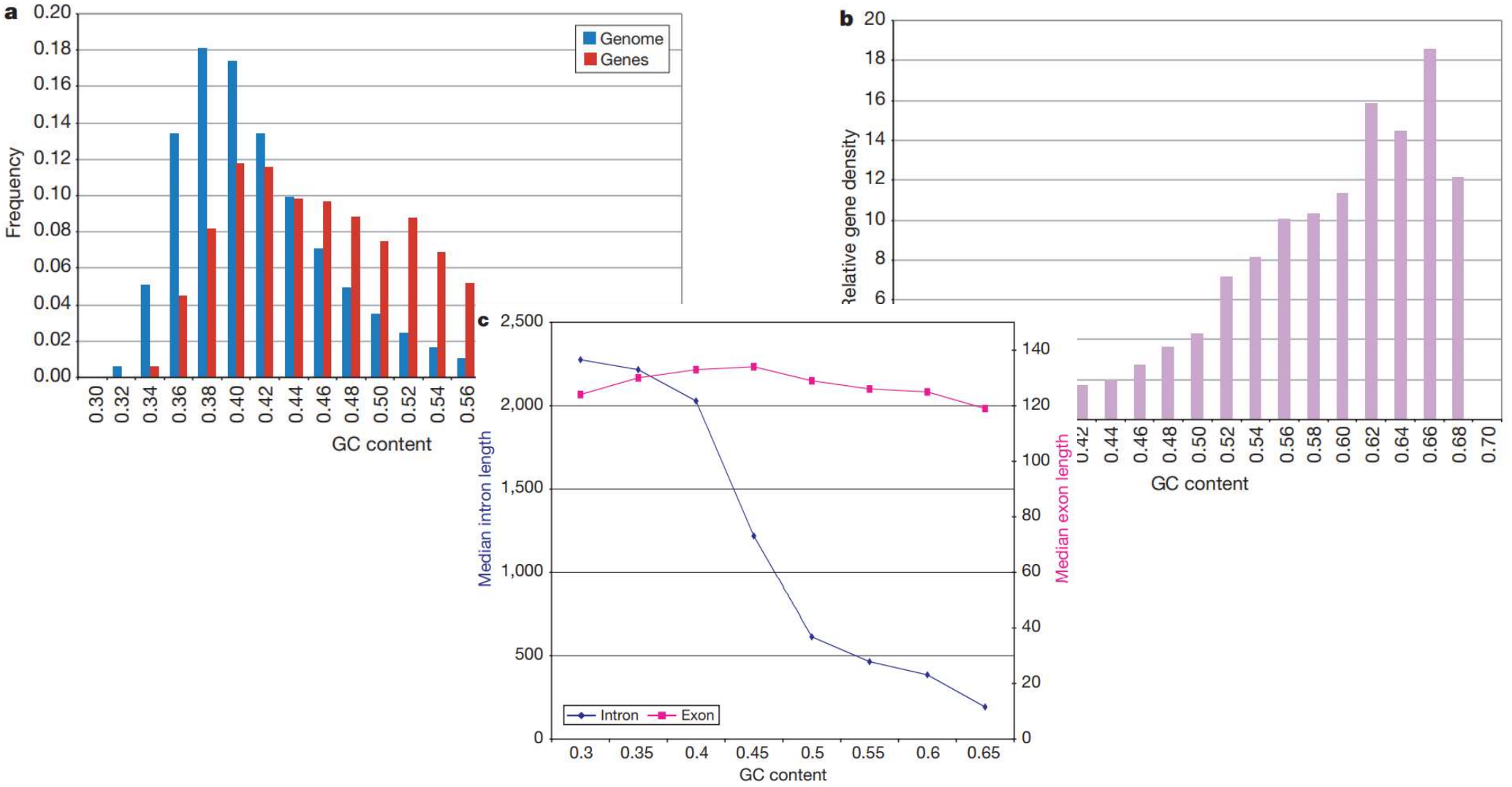
Gene content of the human genome





BGI华大

Gene content of the human genome



BGI华大

Gene content of the human genome

Table 22 Properties of the IGI/IPI human protein set

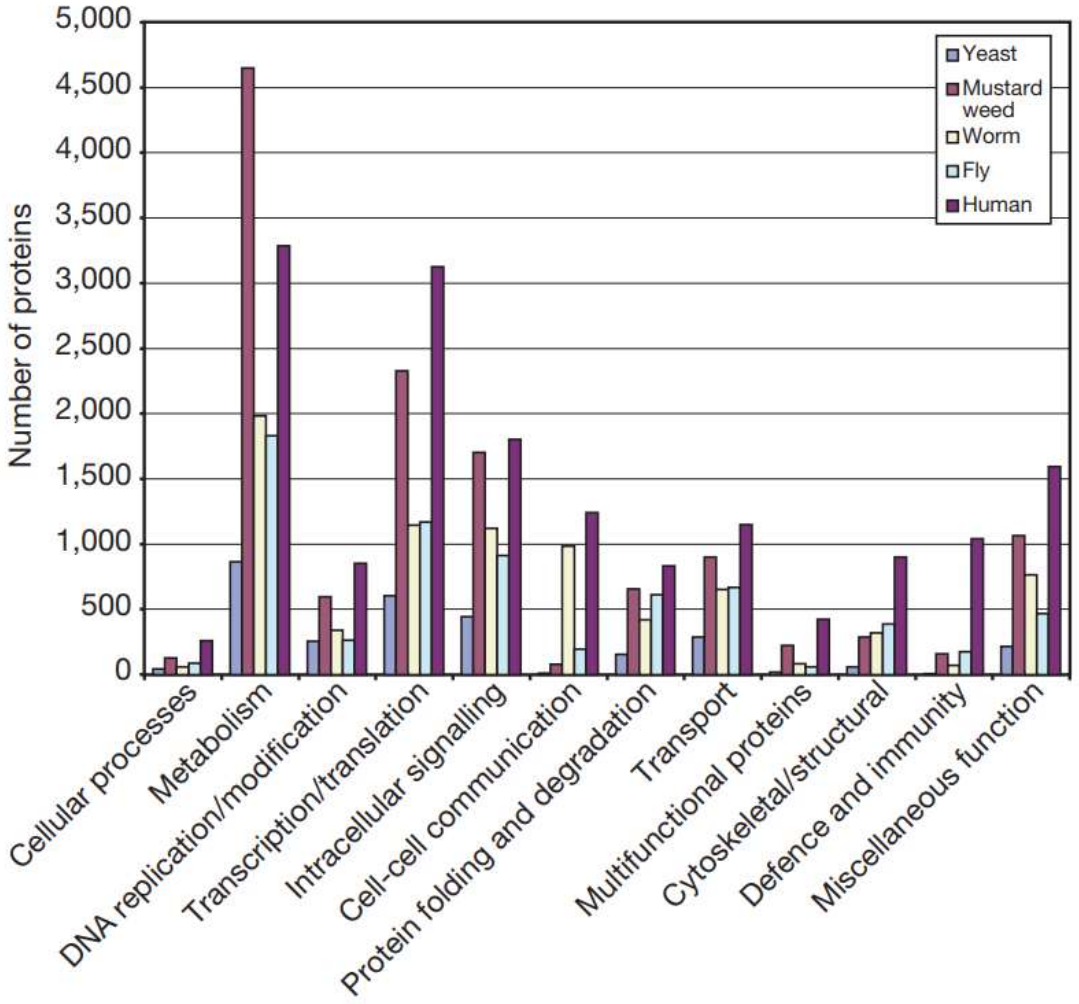
Source	Number	Average length (amino acids)	Matches to nonhuman proteins	Matches to RIKEN mouse cDNA set	Matches to RIKEN mouse cDNA set but not to nonhuman proteins
RefSeq/SwissProt/TrEMBL	14,882	469	12,708 (85%)	11,599 (78%)	776 (36%)
Ensembl–Genie	4,057	443	2,989 (74%)	3,016 (74%)	498 (47%)
Ensembl	12,839	187	81,126 (63%)	7,372 (57%)	1,449 (31%)
Total	31,778	352	23,813 (75%)	219,873 (69%)	2,723 (34%)

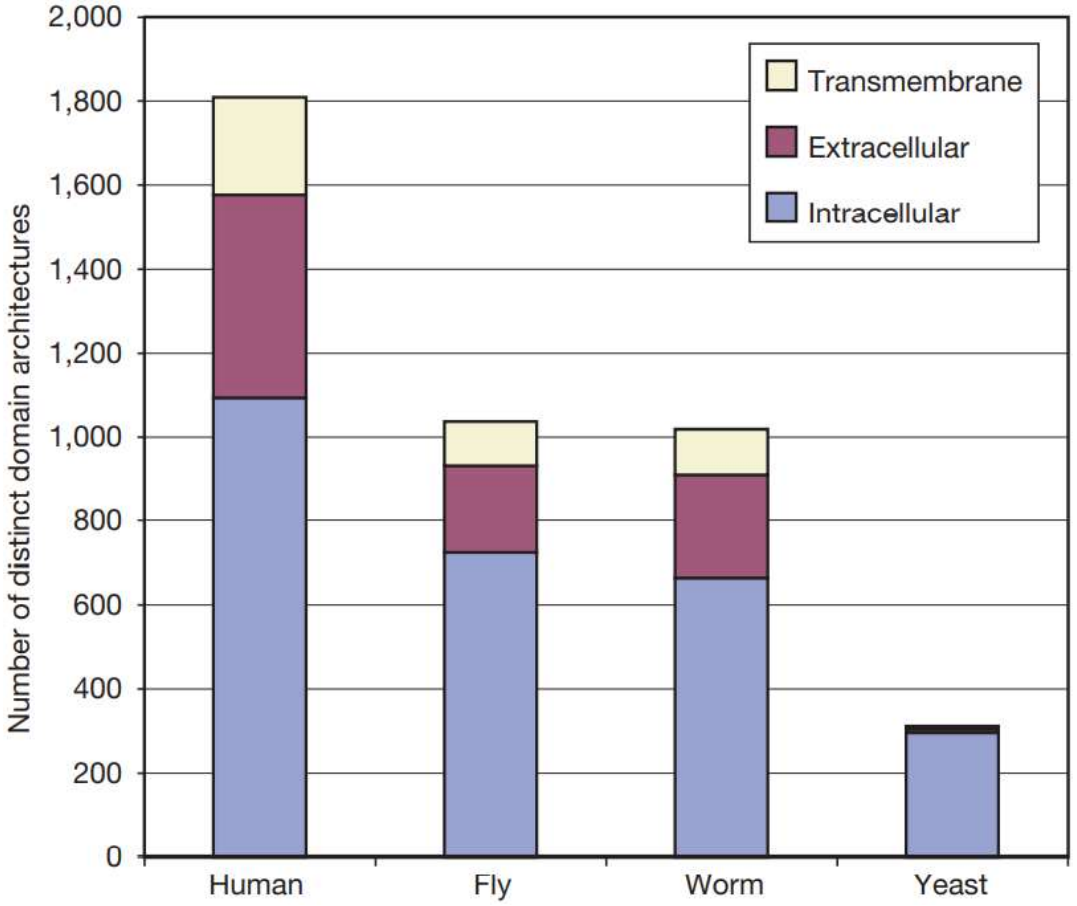
The matches to nonhuman proteins were obtained by using Smith-Waterman sequence alignment with an *E*-value threshold of 10^{-3} and the matches to the RIKEN mouse cDNAs by using TBLASTN with an *E*-value threshold of 10^{-6}. The last column shows that a significant number of the IGI members that do not have nonhuman protein matches do match sequences in the RIKEN mouse cDNA set, suggesting that both the IGI and the RIKEN sets contain a significant number of novel proteins.

随后工作人员进行了基因组的注释工作，产生了第一版本的基因和蛋白组注释文件IGI和IPI

BGI华大

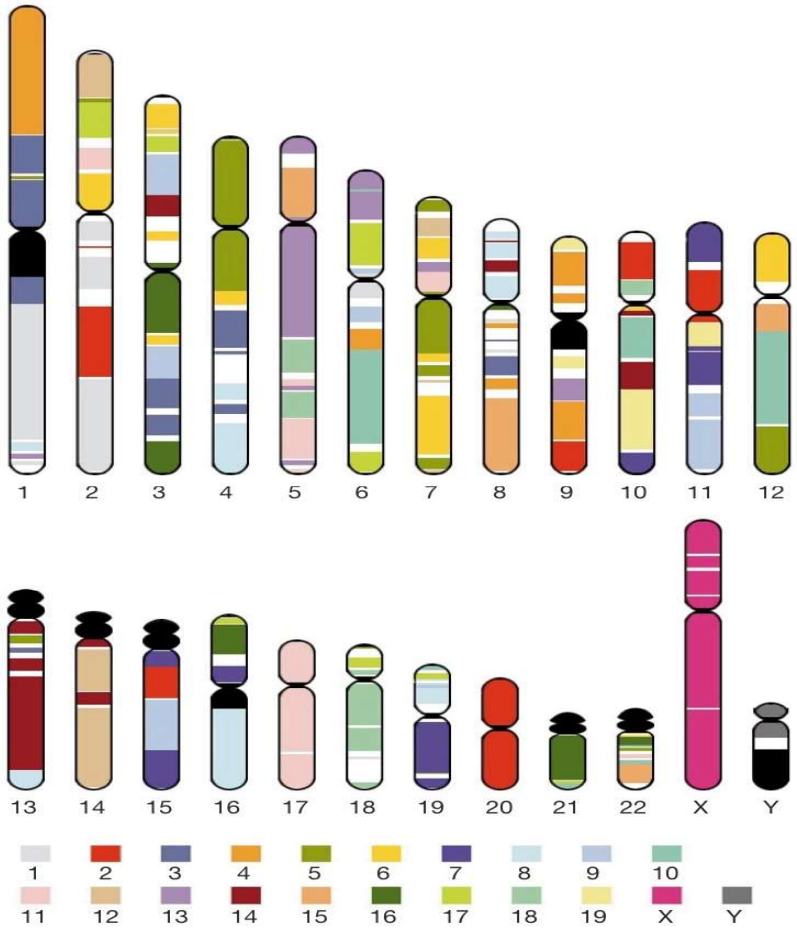
Gene content of the human genome





BGI华大

Segmental history of the human genome



- 人和小鼠之间的保守片段
- 人类基因组中古老的重复片段
- SNP

- 人和小鼠基因组中的保守片段。人类染色体，其片段包含至少两个基因，其顺序在小鼠基因组中作为色块保守。每种颜色对应一个特定的小鼠染色体。

BGI华大

Applications to medicine and biology

- 基础医学
  疾病基因
  药物靶点
- 基础生物学

Locus	Disorder	Reference(s)
BRCA2	Breast cancer susceptibility	55
AIRE	Autoimmune polyglandular syndrome type 1 (APS1 or APECED)	389
PEX1	Peroxisome biogenesis disorder	390, 391
PDS	Pendred syndrome	392
XLP	X-linked lymphoproliferative disease	393
DFNA5	Nonsyndromic deafness	394
ATP2A2	Darier's disease	395
SEDL	X-linked spondyloepiphyseal dysplasia tarda	396
WISP3	Progressive pseudorheumatoid dysplasia	397
CCM1	Cerebral cavernous malformations	398, 399
COL11A2/DFNA13	Nonsyndromic deafness	400
LGMD 2G	Limb-girdle muscular dystrophy	401
EVC	Ellis-Van Creveld syndrome, Weyer's acrodental dysostosis	402
ACTN4	Familial focal segmental glomerulosclerosis	403
SCN1A	Generalized epilepsy with febrile seizures plus type 2	404
AASS	Familial hyperlysinaemia	405
NDRG1	Hereditary motor and sensory neuropathy-Lom	406
CNGB3	Total colour-blindness	407, 408
MUL	Mulibrey nanism	409
USH1C	Usher type 1C	410, 411
MYH9	May-Hegglin anomaly	412, 413
PRKAR1A	Carney's complex	414
MYH9	Nonsyndromic hereditary deafness DFNA17	415
SCA10	Spinocerebellar ataxia type 10	416
OPA1	Optic atrophy	417
XLCSNB	X-linked congenital stationary night blindness	418
FGF23	Hypophosphataemic rickets	419
GAN	Giant axonal neuropathy	420
AAAS	Triple-A syndrome	421
HSPG2	Schwartz-Jampel syndrome	422

使用基因组序列草图定位克隆的疾病基因

BGI华大

The next steps

- 完成人类序列
- 开发 IGI 和 IPI
- 在全基因组范围内研究表观遗传修饰
- 其他大基因组的测序
- 完成人类变异目录

BGI华大

THANKS

OMICS FOR ALL

基 因 科 技 造 福 人 类