

第 3 章

物理图绘制

前面一章我们已经学习了如何进行遗传分析将基因及 DNA 标记定位在染色体上绘制遗传图的方法。本章要接触另一种基因组作图，直接检测 DNA 标记在染色体上的实际位置，即绘制基因组物理图。我们将看到遗传图与物理图如何相互结合提供基因组测序的工作框架。

为何要绘制物理图？遗传分析绘制的基因组图为何不能指导基因组计划的测序？主要原因有以下几点：

(1) 遗传图的分辨率有限。这一点对微生物来说不成问题，因为微生物基因组较小，记录成百上千次交换试验就能获得十分详尽的遗传图，标记分布密度仅为数千核苷酸。人类及大多数高等真核生物由于不可能获得大量的子代，只有少数的减数分裂事件可供研究，连锁分析的分辨力受到很大限制。20 世纪 80 年代中期，当讨论人类基因组计划的设想时，人们就已提出人类遗传图理想的标记密度应为 1 Mb (1 000 kb)，但当时估计可能达到的最好结果是每 2~5 Mb 一个标记。基因组测序和组装所要求的标记密度最好在 100 kb 左右，显然仅靠遗传作图是无法达到这一目标的，必须借助于其他的作图方法。

(2) 遗传图的覆盖面较低。前面一章中提到过 Sturtevant 的关于交换是随机发生的猜想。现在已经知道，这一想法只是部分正确，因为重组热点的存在使染色体某一区段的交换频率高于其他区段。由于许多未知的原因，有些染色体区段很少发生重组事件，因而在这些区段难以绘制高密度连锁图。特别是倒位区段，由于同源姐妹染色体的交换受到限制，无法绘制精细遗传图。

(3) 遗传图分子标记的排列有时会出现差错。遗传作图的依据主要是子代个体的基因型重组及其分离比。由于环境因素和取样误差，可能存在非随机的群体组成。在这种条件下，采用不同的杂交组合有时会得出不相同的结果，相同的分子标记在连锁图上的位置不同。

由于遗传图存在以上这些局限，在进行大规模的 DNA 测序之前，对大多数真核生物的

遗传图必须进行验证并利用其他作图技术予以校正和补充。这些非遗传学的作图技术统称为物理作图，它们都是以 DNA 分子的序列排列为基础。物理作图的方法和策略多种多样，常用的有以下 4 类：

(1) 限制性作图 (restriction mapping)。它是将限制性酶切位点标定在 DNA 分子的相对位置。

(2) 基于克隆的基因组作图 (clone-based mapping)。根据克隆的 DNA 片段之间的重叠序列构建重叠群 (contig)，绘制物理连锁图。

(3) 荧光原位杂交 (fluorescent *in situ* hybridization, FISH)。将荧光标记的探针与染色体杂交确定分子标记的所在位置。

(4) 序列标签位点 (sequence tagged site, STS) 作图。通过 PCR 或分子杂交将小段 DNA 序列定位在基因组的 DNA 区段中。

3.1 限制性作图

前面我们曾提到 RFLP 标记，这是由限制性内切核酸酶 (简称限制性内切酶，或限制酶) 酶切产生的一段 DNA 片段。不同限制酶识别的序列各异，大多数限制酶的识别序列为 4~6 个碱基对，但也有一些限制酶可识别的序列为 7、8 甚至更多的碱基对 (图 3.1)。

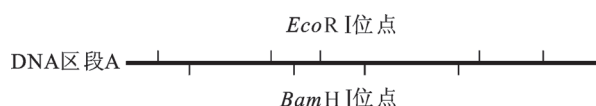


图 3.1 同一段 DNA 分子上有不同的限制性酶切位点

已知有 3 种限制性内切酶：

(1) I 型限制性内切酶。这是一类兼有限制性内切酶和修饰酶活性的蛋白复合体，可催化宿主 DNA 的甲基化。它们在识别位点约 1 000 bp 以外的任何位置切割 DNA 链，不产生确定的限制片段和明确的条带。

(2) II 型限制性内切酶。在识别位点之中或临近的确定碱基位置特异地切开 DNA 链。它们产生确定的限制片段和条带，是 3 类限制性内切酶中唯一可用于 DNA 限制性作图的酶。

(3) III 型限制性内切酶。兼有限制性酶切和修饰两种功能的酶，可催化宿主 DNA 的甲基化。它们在识别位点之外 25~27 bp 的位置切开 DNA 链，并且要求识别位点是反向重复序列，如 *Bcg* I，CGANNNNTGCG；它们很少能产生完全酶切的片段，因而不具备实用价值。

3.1.1 限制性作图的基本方法

最简单的构建限制图的方法是比较不同限制酶产生的 DNA 片段的大小。首先用一种限制酶处理样品，经琼脂糖凝胶电泳分离染色后可见大小确定的 DNA 片段。然后用第二种限制酶处理获得第二组片段。最后用 2 种酶混合处理，获得第三组片段。收集所有上述资料进行对比组装，对于 2 种酶切位点交替出现的区段，利用加减法即可确定酶切位点的相对位置。在连续出现 2 个或多个相同酶切位点的区段，其排列序列可有多种选择，此时采用部分酶切的

办法使该区段只发生一次酶切，然后计算产生片段的长度，选择其中正确的排序。后者称为部分限制作图（partial restriction mapping）。

部分限制作图可提供完整物理图绘制所需的信息，但是当涉及的限制位点过多时，这种方法就显得力不从心。特别是内部含有大小相同的片段时，位置重叠的片段无法区分，使排序变得非常复杂。此时可采用末端同位素标记（或其他标记）结合部分酶解的方法进行物理图绘制。由于只在 DNA 分子的 5′ 端或 3′ 端进行标记，在部分酶解处理后，样品经琼脂糖凝胶电泳分离，只有含末端标记的片段才显示信号，形成梯形分布带。

3.1.2 限制性作图的局限

如果样品中仅存在较少的限制性位点，用常规的限制酶即可绘制 DNA 物理图。随着 DNA 分子长度的增加及酶切位点的增多，单酶切、双酶切及部分酶切产生的条带数成比例扩大，需要对大量的片段进行比对与组装。虽可借助计算机帮助完成一些可明显区分的 DNA 条带的排列顺序，但仍不可避免地会产生许多问题。因为大量片段经琼脂糖凝胶电泳分离时总会有些片段相互贴近，增加了分辨单个条带的困难，很可能遗漏一些片段。因此限制性作图只能应用于较小的 DNA 分子，其上限取决于作图分子中限制酶位点的频率。通常采用的 6 碱基对识别序列的限制酶仅适合 50 kb 以下 DNA 分子的精确作图，但 50 kb 远低于细菌及真核生物染色体的长度，只可覆盖少数病毒和细胞器基因组。限制图能否用于大于 50 kb 基因组的作图呢？下面我们对这一问题给予定性的解答。

稀有切点限制酶

现已发现 2 500 多种限制性内切酶，其中有 300 多种常用于实验研究。许多限制性内切酶的识别序列为 6 碱基对，其他一些识别序列或长或短，有一些限制性内切酶可识别兼并的碱基序列，如来自流感嗜血杆菌（*Haemophilus influenzae*）的限制酶 *Hinf* I 可识别 5′-GATNC-3′，其中 N 代表 4 种碱基中任何一种。

获得理想大小的 DNA 片段对基因组的限制性作图至关重要，其关键是选择合适的限制酶。限制酶识别位点的核苷酸数目与组成决定了该酶产生的 DNA 片段的大小。如果基因组中 4 种碱基的分布是随机的，预期产生的限制性 DNA 片段大小为 4^n ，这里 n 代表限制酶识别位点的碱基数目。据此可以推算得表 3.1：

表 3.1 限制酶识别位点碱基数及其在 DNA 序列中分布间隔长度

识别位点碱基数目	预计 DNA 片段平均长度 /kb	识别位点碱基数目	预计 DNA 片段平均长度 /kb
4	1	8	64
6	4	10	1 000

常用的 4 或 6 碱基限制酶只能产生较小的 DNA 片段，要获得大于 50 kb 的 DNA 片段必须采用稀有切点限制酶。所谓稀有切点限制酶系指该酶识别的碱基序列在基因组中只有很少数量，因此可产生较大的 DNA 片段。

稀有切点限制图绘制 选用稀有切点限制酶时有几点必须注意：①一般而言，识别序列越长产生的片段越大。但是，当识别位点中含有非特异序列时，由于随机选择的干扰，片段

的长度比预期的小。②识别位点的碱基序列组成影响限制性片段的大小。例如,人类基因组像其他哺乳类一样,其 DNA 的 A/T 比例接近 60%。如选用高比例 A/T 位点限制酶,产生的片段数多于高比例 G/C 识别位点酶。③有些限制酶识别位点较长,如酵母编码的 *I-Sce* I 识别序列为 18 bp,即使高等真核生物基因组亦无此序列。此时可采用变通的方法,将这类识别位点引入待测的基因组中。由于导入的 *I-Sce* I 序列可随机插入到基因组中,因而可以获得大小不同的 DNA 区段用于物理图绘制。④基因组 DNA 的甲基化状态。基因组 DNA 序列并非完全随机分布,某些特异限制酶识别序列只分布在一定的区段。例如,人类基因组 DNA 中 5'-CG-3' 的序列较少,因为人细胞中具有将 5'-CG-3' 中胞嘧啶 5- 碳原子甲基化的酶,产生的 5- 甲基胞嘧啶在生理条件下极不稳定,脱氨基后转变为胸腺嘧啶。在漫长的进化历史中,人类基因组中原有的许多 5'-CG-3' 序列由于甲基化突变逐渐转变为 5'-TG-3' 序列。此外,活细胞中大量 5'-CG-3' 序列的胞嘧啶被甲基化 (m^5C),许多切点含有 5'-CG-3' 序列的限制酶在人类基因组 DNA 中只能找到很少可切位点。如限制酶 *Sma* I (5'-CCCGGG-3') 处理人类基因组 DNA 产生的片段平均长度为 78 kb,而 *Bss*H II (5'-GCGCGC-3') 平均为 390 kb。*Not* I 是一个 8 核苷酸限制酶,靶序列 (5'-GCGCGCGC-3') 含 3 对 5'-CG-3',在人类基因组 DNA 中平均每 1 Mb 含一个 *Not* I 切点。果蝇和酵母基因组 DNA 缺少甲基化修饰,因此甲基化限制性内切酶不适合这种场合。限制性作图的潜力随使用的稀有切点限制性内切酶种类而增加,在低等生物基因组以及大分子 DNA 克隆片段的物理图绘制中被广泛采用 (表 3.2, 图 3.2)。

表 3.2 限制性内切酶及其识别位点

限制性内切酶	识别位点 (5' → 3')
	识别长序列位点
<i>I-Sce</i> I	TAGGGATAA/CAGGGTAAT
<i>VDE</i>	TATSYATGYYGCTGY/GGRGAARKMGA/AWGAAWG
<i>I-Ceu</i> I	TAACTATAACAATAATA/AGGTAGCGA
<i>I-Tli</i> I	GGTTCTTTATGCGGACAC/TGACGGCTTTATG
<i>I-Ppo</i> I	CTCTCTTAA/GGTAGC
	识别序列大于 6 碱基
<i>Pac</i> I	TTAAT/TAA
<i>Pme</i> I	GTTT/AAAC
<i>Swa</i> I	ATTT/AAAT
<i>Sse</i> 8388t1	CCTGCA/GC
	识别序列大于 6 碱基,可针对 CpG 岛序列
<i>Not</i> I	GC/GGCCGC
<i>Srf</i> I	GCCC/GGGC
<i>Sfi</i> I	GGCCNNNN/NGGCC

续表

限制性内切酶	识别位点 (5' → 3')
	识别 CpG 岛序列, 产生大于 200 kb 片段
<i>Mlu</i> I	A/CGCGT
<i>Sal</i> I	G/TCGAC
<i>Nru</i> I	TCG/CGA
<i>Eag</i> I	C/GGCCG
<i>Sac</i> II	CCGC/GG

注: N, 指任意碱基; R, A 或 G; Y, C 或 T; M, A 或 C; K, G 或 T; S, G 或 C; W, A 或 T。

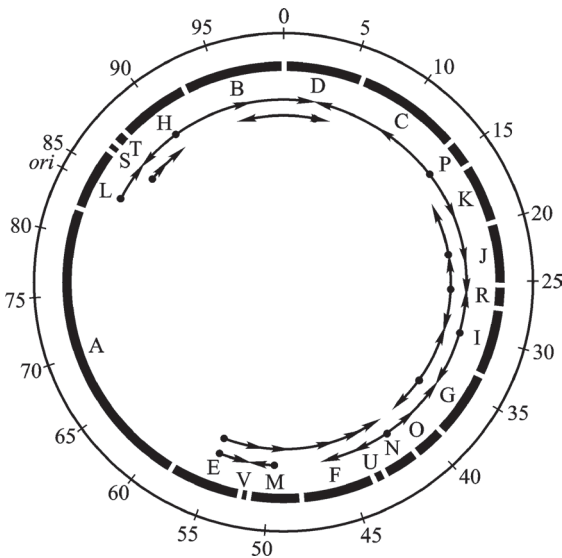


图 3.2 大肠杆菌基因组限制性酶切位点物理图

大肠杆菌基因组全长 4.5×10^6 kbp, 用 *Not* I 酶切产生了 21 个片段, 长度在 20 ~ 1 000 kb 之间。经脉冲凝胶电泳分离后, 将酶切片段转移到杂交膜上。选用已经遗传定位的基因为探针与之杂交, 可确定各酶切片段排列位置, 带箭头连线为部分酶解确定的物理图 (Smith CL 等, 1987)

大分子 DNA 片段的分离 稀有切点限制性内切酶产生的 DNA 片段往往超过 50 kb, 如 *Not* I 酶切片段可达 1 000 kb, 必须用特殊的凝胶电泳技术才能分离与观测。DNA 分子的长度与其在琼脂糖凝胶电泳中的迁移率仅在很小的范围内呈线性关系, 随着 DNA 分子长度的增加, 凝胶电泳的分辨力急剧下降。常规琼脂糖凝胶电泳即使在低浓度的凝胶条件下, 也只能分离约 30 kb 以下的线性分子。大于 30 kb 以上的 DNA 在常规电泳凝胶中彼此挤压, 形成单一 DNA 迁移带。为了解决这一问题, Schwartz DC 和 Cantor CR (1984) 首次采用脉冲凝胶电泳 (pulsed-field gel electrophoresis, PFGE), 突破了琼脂糖凝胶分离大分子 DNA 的限制。在不改变其他如凝胶浓度、电场强度和缓冲液的条件, 脉冲凝胶电泳使大分子 DNA 的分辨力提高了 2 个数量级, 达到 10 Mb。PFGE 的基本原理是, 将一个方向不断变换的电场取代简单的单一电场 (单向电场), 使电泳中受阻的 DNA 分子在电场改变时扭转迁移方向, 达到分离

的目的。正交交变电场凝胶电泳 (orthogonal field alternation gel electrophoresis, OFAGE) 是这种复杂电场的一个例子。OFAGE 有两对电极, 每对电极与凝胶的长轴成 45° 角, 两个电场方向彼此成一直角 (图 3.3A)。垂直交变电场的两个电流方向是固定的, 但两个电极交替打开与关闭。凝胶中的 DNA 分子在电场方向不断变换的作用下, 随时改变泳动的方向。小分子 DNA 比大分子 DNA 更易在凝胶中重新定向, 因而迁移速度更快。根据同一原理设计的电泳技术还包括: 夹角等值均一电场 (contour-clamped homogeneous electric field, CHEF) 和电场逆转凝胶电泳 (field inversion gel electrophoresis, FIGE)。交变电场分辨大分子 DNA 的范围远远超出常规电泳, 可将酵母染色体彼此分开 (图 3.3B)。

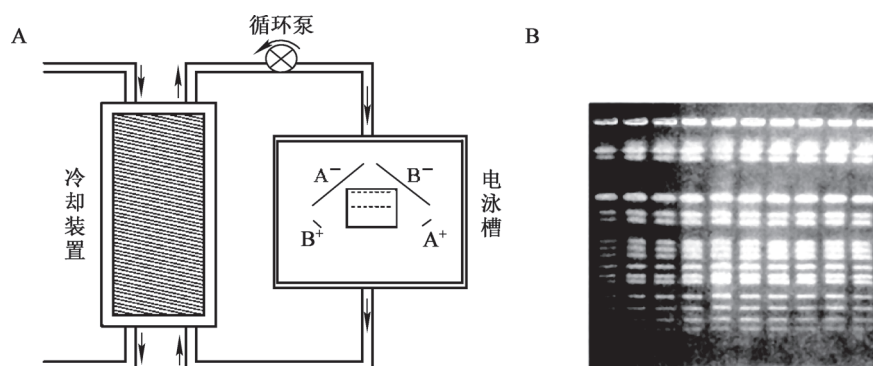


图 3.3 常规或非常规琼脂糖凝胶电泳

A. 正交交变电场凝胶电泳 (OFAGE), 有 2 对电极分别位于凝胶对角线的两端, 2 对电极轮流接通电场产生一个交替的方向不同的电场, DNA 分子在凝胶中不断改变迁移方向。交变电场电泳主要用于相对分子质量较大 DNA 的分离。B. 酵母染色体交变电场电泳图, 染色体按相对分子质量大小分离

3.2 基于克隆的基因组作图

20 世纪 70 年代发展起来的细菌质粒载体通常只适合克隆小分子 DNA。随着插入的 DNA 片段增大, 质粒的稳定性下降, 因此常规的质粒载体不适用于大分子 DNA 的克隆。为了解决这一问题, 人们对 λ 噬菌体基因组进行改造, 构建了可调控的 λ 克隆载体以及 λ 与质粒相结合的黏粒 (cosmid) 载体。这类载体所能克隆 DNA 片段的大小决定于 λ 噬菌体颗粒头部的容量, 因为重组的载体 DNA 必须包装进 λ 噬菌体才可感染大肠杆菌。噬菌体头部的大小是确定的, 所以 λ 载体克隆的 DNA 片段有一上限。尽管有这一限制, 一些较小的基因组, 如啤酒酵母 (*Saccharomyces cerevisiae*) 基因组主要是先构建黏粒文库, 然后建立克隆重叠群完成测序的 (Olive SG 等, 1992)。黏粒载体克隆的 DNA 片段比质粒载体提高了一个数量级, 但对于高等生物基因组而言仍然太小。高等植物中, 最小的拟南芥基因组为 1.0×10^8 bp。如果克隆的 DNA 平均长度为 40 kb, 覆盖一个单倍体拟南芥基因组必需 2 500 个克隆。要达到 99.9% 的基因组覆盖率, 需要 25 000 个克隆。而人类基因组是拟南芥基因组的 30 倍, 必须上百万个黏粒克隆才能满足基因组物理图绘制的要求。当人们打算实施人类基因组计划时, 寻找大分子 DNA 的克隆载体便成为当务之急。

3.2.1 大分子 DNA 的克隆载体

克隆大于 100 kb DNA 的突破首先是从酵母人工染色体 (yeast artificial chromosome, YAC) 的发明开始的 (Burke DT 等, 1987)。与质粒和噬菌体载体不同的是, 重组 YAC 分子只能在酵母细胞中扩增。YAC 载体的设计原理不同于质粒与噬菌体, 它是根据真核染色体的结构设计的。YAC 载体主要由 3 个基本部件构成 (图 3.4):

- (1) 着丝粒。在细胞分裂时负责染色体均等分配。
- (2) 端粒。位于染色体端部的特异 DNA 序列, 保持人工染色体的稳定性。
- (3) 自主复制序列 (autonomous replication sequence, ARS)。在细胞中启动染色体的复制。

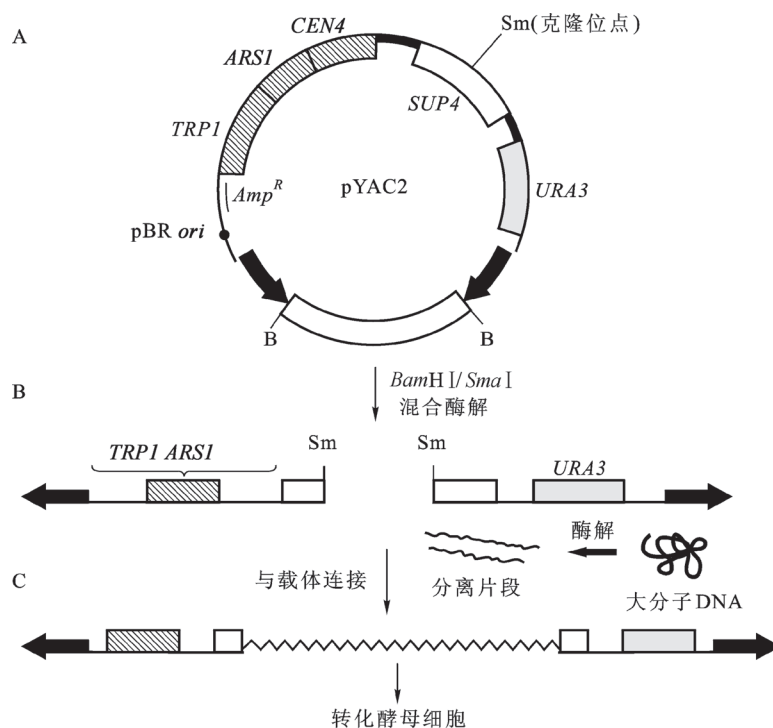


图 3.4 酵母人工染色体 (YAC) 操作示意图

A. 酵母人工染色体载体 pYAC2 物理图。B. 外源 DNA 的克隆。用 *Bam*H I 消化 pYAC2 环化载体, 将端粒之间的一段填充序列除去。*Sma* I 处理在 *SUP4* 基因内部产生插入位点, 克隆的大分子 DNA 片段与两段载体臂连接。C. 重组的酵母人工染色体。pYAC2 载体携带选择标记 *TRP1* 和 *URA3* 基因, 受体菌基因型为 *trp1*⁻ 和 *ura3*⁻。在缺少野生型 *TRP1* 和 *URA3* 基因时, 受体菌只能在添加色氨酸和尿苷酸培养基上生长。转化之后, 凡是含有 pYAC2 载体的受体菌均可在基本培养基上生长。酵母人工染色体由 2 个臂组成, 左臂与右臂各含一个筛选标记基因, 受体菌只有同时含有 2 个臂才能生长在基本培养基上。载体是否插入了外源 DNA 可由 *SUP4* 酶活性检测。在插入片段存在时, *SUP4* 失活, 克隆显示红色。无插入片段时, 克隆显示白色。*CEN4*, 酵母 4 号染色体着丝粒; *ARS1*, 酵母自主复制序列 1; pBR ori, 大肠杆菌质粒复制起始点; *Amp*^R, 质粒扩增筛选标记基因

上述元件组合在一个长度为 10~15 kb 的 DNA 分子中, 内含一个插入位点及数个用于筛选的标记基因。酵母染色体的可载容量为 230 kb 至 1 700 kb 之间, 相当于天然的酵母染色体。标准的 YAC 操作程序克隆的 DNA 片段平均约 600 kb, 改进的程序可获得 1 400 kb 插

入子。YAC 克隆系统在 20 世纪 80 年代末与 90 年代初被广泛采用,但后来发现 YAC 克隆存在插入子的稳定性问题以及同一酵母细胞多个 YAC 引起交换产生嵌合序列。此外还发现 YAC 载体对人类基因组富含 GC 区的克隆效率比较低 (Chumakov I, 1992) (Chumakov I 等, 1992), 因而迫使人们又去寻找更好的大分子克隆载体。这些进展包括:

噬菌体 P1 载体 (Sternberg N, 1990) 该载体与 λ 载体非常相似,也是将天然噬菌体基因组中的一段区域缺失,其容量也取决于缺失区段的大小以及噬菌体颗粒能容纳的空间。P1 基因组比 λ 基因组更大,克隆的片段可达 125 kb。

细菌人工染色体 (bacterial artificial chromosome, BAC) (Shizuya H 等, 1992) BAC 载体源于大肠杆菌中天然的 F 质粒。F 质粒与一般质粒有 2 点不同: ①单拷贝复制; ②相对分子质量大。因此 F 质粒衍生的 BAC 载体具有较大的克隆容量,可达 300 kb 以上,而且比较稳定。因每个细胞只有一份 BAC 拷贝,克隆的 DNA 片段不会因重组发生嵌合问题。BAC 的另一优点是,可采取类似制备质粒的方法直接提取克隆的 DNA,在大规模测序时便于机械化操作 (图 3.5)。

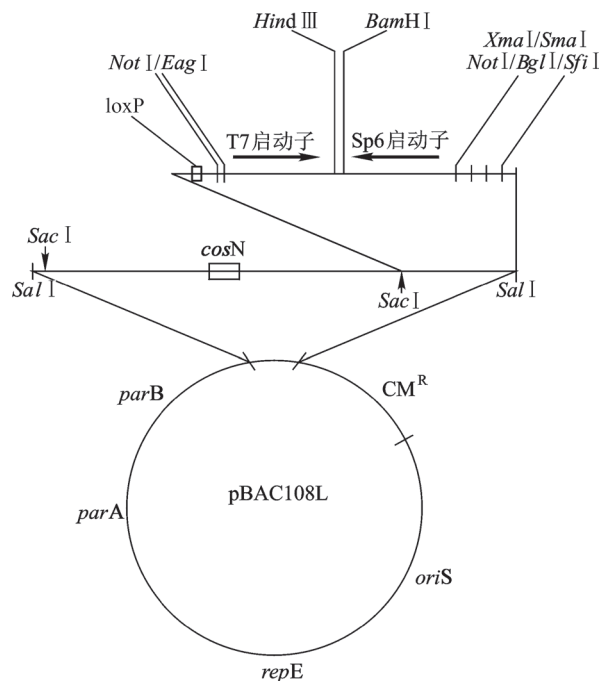


图 3.5 BAC 载体物理图

pBAC 载体为 mini-F 的衍生质粒。*oriS* 和 *repE* 基因介导 F 因子的单向复制, *parA* 和 *parB* 维持低水平质粒拷贝数, 每个基因组约 1~2 个拷贝。*cosN* 为 λ 噬菌体末端酶专一性切割位点, *loxP* 为 P1 Cre 蛋白作用位点, 均可将环状 BAC DNA 转变为线型分子, 便于物理图绘制。*Hind* III 和 *Bam* H I 为外源 DNA 插入位置, 两侧的 *Not* I 位点可用于直接检测克隆的大分子 DNA

P1 人工染色体 (P1-derived artificial chromosome) 又称 PAC (Ioannous PA 等, 1994), 该系统结合了 P1 载体与 BAC 载体的优点, 可载片段达 300 kb (图 3.6)。

F 黏粒 (fosmid) 该载体含有 F 质粒的复制起始点以及 λ 的 *cos* 位点。其功能与黏粒

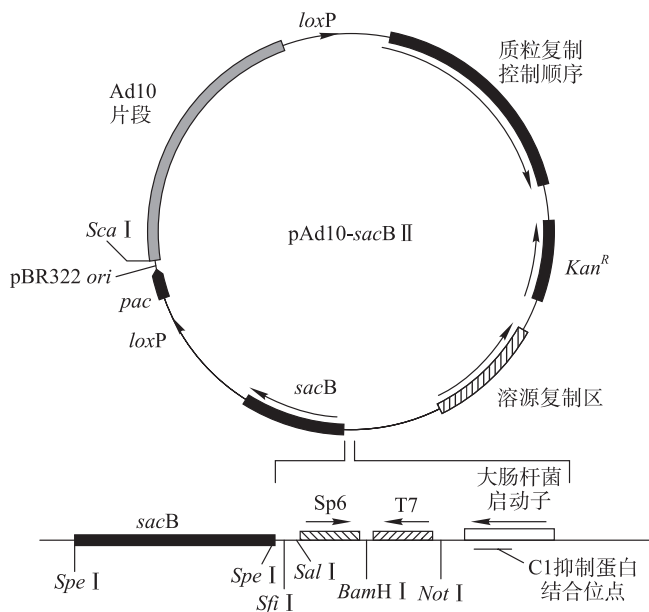


图 3.6 PAC 载体

图示 P1 正选择克隆载体 pAd10-sacB II，*loxP* 位点将该载体分为两个结构区。一个结构区含有 P1 *pac* 序列，pBR322 复制起始点和一段约 10 kb 来自腺病毒的填充区段。另一结构区含有 P1 质粒复制子，卡那霉素抗性筛选标记基因 (*Kan^R*)，IPTG 诱导裂解复制子和正筛选 *sacB* II 卡盒。*sacB* II 卡盒（载体图下横线标示）含一个人工合成的近似的大肠杆菌启动子一致序列，一个可以调控大肠杆菌启动子表达的 P1 C1 抑制因子结合位点，一个将 Sp6 和 T7 分开的 *BamH* I 位点和数个稀有酶切位点 *Not* I，*Sfi* I 以及 *Sal* I。*sac* 基因编码将蔗糖转变为果聚糖 (levan) 的果聚-蔗糖酶。果聚糖的积累可使大肠杆菌致死，在 2% 蔗糖存在时细菌不能生长。当外源 DNA 插入到克隆位点后可阻止果聚-蔗糖酶的表达，因而只有重组子才能存活

(cosmid) 类似，只是拷贝数较低，但避免了不稳定性。

3.2.2 重叠群组建

相互重叠的 DNA 片段组成的物理图称为重叠群 (contig)，克隆重叠群的组建最早采用的是染色体步移 (chromosome walking) 法。染色体步移最初用于 λ 或黏粒载体的克隆重叠群构建，涉及的范围比较短。首先从基因组文库中挑取一个指定的或随机的克隆，将该起始克隆的末端序列分离纯化作为探针，然后在基因组文库中寻找与之重叠的第二个克隆。在第二个克隆的基础上再按同样的操作程序寻找第三个重叠的克隆，依次延伸，直到完成所需要的重叠群 (图 3.7)。

染色体步移的速度缓慢，仅适合于小基因组及小区段染色体的物理图绘制。对大基因组的物理图绘制，必须寻找更加快捷而准确的方法。

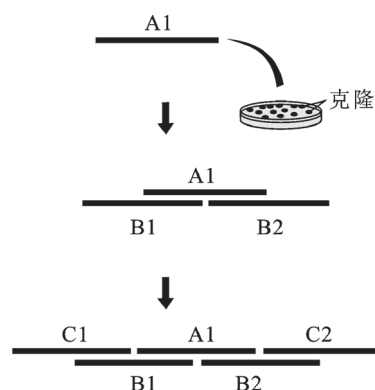


图 3.7 染色体步移

本例中起始克隆为 A1，将其插入片段作为探针与文库所有克隆杂交，筛选到 2 个不同的阳性克隆 B1 和 B2。再分别以 B1 和 B2 插入子作为探针重复杂交，筛选到下一轮阳性克隆 C1 和 C2。重复这一过程，可建立一个彼此连接的重叠群

3.2.3 指纹作图

在基因组范围内查找重叠的克隆, 最好的方法首推克隆指纹 (clone fingerprinting) 排序。这里所指的指纹系指克隆的 DNA 序列所具有的特定的 DNA 片段组成。这些 DNA 片段可以通过不同的方法产生, 经凝胶电泳显示后组成特定排列模式的 DNA 条带, 即指纹。每个克隆的 DNA 序列都有其特征性的指纹, 这些指纹称为克隆指纹。如果两个 DNA 克隆含有部分重叠的序列, 那么由这些重叠的 DNA 序列产生的 DNA 指纹应该是相同的。反之, 如果两个 DNA 克隆含有部分相同的指纹, 那么这两个克隆就含有部分重叠的 DNA 序列, 它们在基因组中应该相邻排列 (图 3.8)。

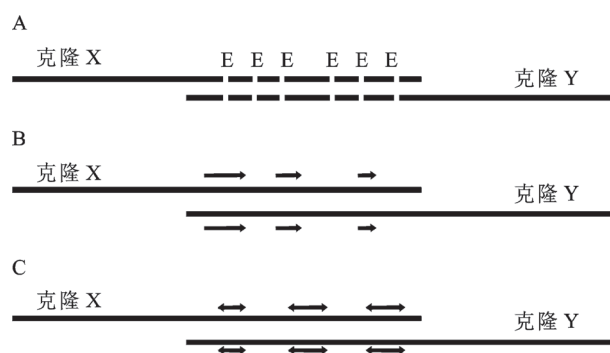


图 3.8 3 种克隆指纹分析方法

A. 限制性带型指纹; B. 重复序列指纹; C. STS 作图指纹。

克隆指纹主要用于重叠群构建, 其原理是, 如果两个克隆彼此重叠, 它们一定含有相同的序列。

这些序列的限制性物理图组成一致 (A), 含有相同的重复序列 (B) 以及单一序列 (STS)。

E, 限制酶酶切位点

限制性带型 (restriction pattern) 指纹 用不同限制性酶处理样品, 经凝胶分离产生 DNA 条带。如果 2 个克隆产生的 DNA 条带有部分是相同的, 说明它们含有重叠的序列。

重复序列 DNA (repetitive DNA) 指纹 将不同克隆的 DNA 限制性片段电泳后转移到杂交膜中, 用一种或几种基因组范围的重复序列 (genome-wide repeat) 作为探针与之杂交, 如出现相同的杂交带型, 可判断这 2 个克隆是重叠的。

重复序列 DNA PCR (repetitive DNA PCR) 或分散重复序列 PCR (interspersed repeat element PCR, IRE-PCR) 指纹 分散重复序列在基因组中的密度很高, 它们与单一序列间隔排列, 很少串联。2 个相邻重复序列之间的单一序列长度是不一致的, 设计一对与重复序列互补的引物, 对检测的克隆进行 PCR 扩增可产生一系列单一序列 DNA 条带, 这是一种特殊的指纹。如果 2 个克隆具有相同的 PCR 产物, 说明它们含有重叠的序列。人类基因组中广泛分布的重复序列 Alu 家族, 平均密度为 4 kb。一个 150 kb 的 BAC 克隆中可预期获得 38 个 PCR 产物, 有足够的指纹可供分辨。

指纹重叠群构建 BAC 克隆的重叠群构建在水稻和小鼠等模式生物基因组物理图的绘制中起着重要作用。由于 BAC 克隆 DNA 可以大规模机械化制备, 酶切和电泳凝胶分离, 通过计算机对酶切片段的扫描识别, 可对 BAC 克隆进行大范围的指纹比对和排列, 由

此快速构建指纹重叠群 (fingerprinted contig, FPC) (图 3.8)。如水稻基因组 BAC 库总容量为 23 X (400 Mb), 完成组建的指纹重叠群总计为 438 个; 小鼠基因组 BAC 库总容量为 16 X (2.8 Gb), 完成组建的指纹重叠群 296 个 (<http://mips.helmholtz-muenchen.de/plant/maize/about/background.jsp>)。

为了提高指纹作图的质量与精度, Ding Y 等 (1999) 开发了一种将荧光标记与指纹相结合的技术——高信息量荧光指纹 (fluorescent-based high-information-content fingerprinting, HICF)。Nelson WM 等 (2005) 在 DNA 酶切片段末端荧光标记基础上, 采用 4 碱基和 6 碱基限制酶组合处理方法使酶切片段尽可能缩短, 使之可在自动化测序仪上检测, 提高了灵敏度和工作效率 (图 3.9)。由于指纹的末端携带荧光标记碱基序列信息, 经毛细管电泳可快速准确识别 BAC 酶切片段指纹组成, 有效地降低了重叠群组建的误差。

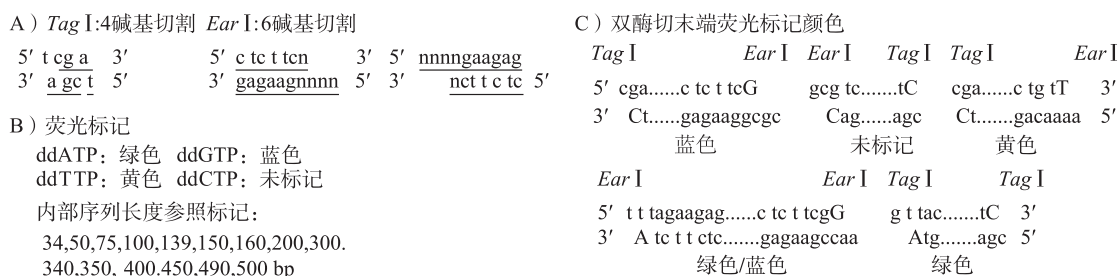


图 3.9 DNA 酶切片段指纹荧光标记法

A) 4 碱基限制酶 *Tag*I 和 6 碱基限制酶 *Ear*I 的酶切模式; B) 酶切片段的单链突出末端经 DNA 聚合酶复制补平, 有针对性地掺入 3 种携带不同荧光颜色的双脱氧核苷酸; C) *Tag*I 和 *Ear*I 组合酶切产生的双链 DNA 片段末端突出单链序列, 在末端复制补平时可掺入的脱氧核苷携带的标记荧光颜色不同, 提供了判别的依据

STS 作图 (STS mapping) 指纹 当需将某一克隆重叠群锚定到现有的已具有 STS 标记的物理图上时, 这一方法特别有效。因为 STS 是单一序列, 并且序列已知。根据选定的某个 STS 序列设计专一性引物, 可对大量的单个克隆进行 PCR 检测, 凡是能扩增出条带的克隆均含有序列重叠的插入子 (Goold RD 等, 1993)。

3.3 原位染色体连锁图

在某些生物种属中, 通过原位杂交 (*in situ* hybridization) 的方法可将基因或 DNA 分子标记定位在染色体的某一区段, 由此绘制基因或 DNA 分子标记所在的染色体位置图称为原位染色体连锁图。原位染色体连锁图可以和染色体分带技术结合, 将基因或 DNA 分子标记定位在特定的染色体区带。所谓染色体显带 (chromosome banding) 系指染色体经一定的物理、化学因素处理并用特定的染料染色后, 在显微镜下显示的特征性的条纹 (见 6.2.1)。最常用的细胞图绘制技术为荧光原位杂交 (fluorescent *in situ* hybridization, FISH)。

荧光原位杂交

荧光原位杂交与同位素杂交的原理相同, 只是标记物不同。FISH 的荧光标记信号可在显微镜下观测, 直接显示探针与染色体杂交的位置 (Heiskanen M 等, 1996)。

3.3.1 同位素或荧光标记探针的原位杂交

原位杂交是杂交技术的一种延伸,其杂交的靶子是完整的染色体,由杂交信号提供作图信息。在原位杂交中,必须先使染色体 DNA 成为单链(这一过程称为 DNA 变性),只有单链 DNA 才可与同样是单链的互补探针杂交。温和的染色体 DNA 变性方法不破坏染色体的自然形态,其程序是首先将样品涂抹在载玻片上自然干燥,然后用甲酰胺处理使其变性。

早期的原位杂交大都采用放射性同位素标记探针,但效果不尽如人意。原因在于同位素标记很难兼顾敏感性与分辨率,这二者对于成功原位杂交缺一不可。敏感性要求放射性标记具有较高的辐射能量(如 ^{32}P),但辐射能量过高会使信号扩散从而降低分辨力。用能量较低的同位素(如 ^3H)可提高分辨率,但因其敏感性低,需延长曝光时间,从而增加了背景干扰,降低了信号分辨的可信度。

20 世纪 80 年代末由于非同位素标记化合物即 DNA 荧光染料的发现使上述问题迎刃而解。这些标记化合物将高敏感性与高分辨力集于一身,从根本上改进了原位杂交的效果。为了使敏感性尽可能提高,可选用能产生极强信号的荧光物质。以往为了达到这一目的,常选择长链 DNA 分子作为探针,一般不少于 40 kb 的 DNA 片段。由于强信号荧光标记物的发现, DNA 探针的长度可以减少。

原位杂交中,如果探针是较长的 DNA 分子,特别是高等真核生物的 DNA,其中可能含有重复序列。探针中的重复序列可与染色体上的许多位点杂交,而不仅仅是与特异的位点杂交。为了减少非专一性干扰,可预先将探针与样品 DNA 混合,后者可以是总 DNA,但最好是富含重复序列的 DNA。探针中的重复序列与非标记 DNA 互补序列复性后,单一序列仍然暴露。在随后的原位杂交时,暴露的探针序列可与样品中的专一性序列杂交,从而降低了非特异性序列的干扰。

3.3.2 染色体原位杂交

染色体原位杂交最初利用的是细胞分裂中期染色体。这些染色体处于高度浓缩状态,有可分辨的染色体带型(chromosome banding pattern)及明显的着丝粒位置。各条染色体的带纹有各自的形态,有可识别的外形。对特定探针而言,原位杂交所显示的中期染色体荧光信号所处的位置与染色体短臂末端的相对距离是不变的,经过比例换算(Flpter 值, Flpter value)可将其标定在连锁图上。高度压缩的中期染色体的不利之处在于分辨力较低,要求 2 个分子标记之间至少需隔开 1 000 kb (Trask BJ 等, 1991)。中期染色体的原位杂交不适合构建可利用的染色体图,主要用于确定新发现的分子标记属于哪条染色体,给出十分粗略的染色体定位。

原位杂交的分辨力取决于杂交的技术以及染色体在制备时所处的状态。中期染色体过于收缩不适于精细作图,为改进作图精度,必须采用处于更为伸展的染色体。有 2 种方法可达此目的(Heiskanen M 等, 1996):

机械伸展的染色体 将分离与制备中期细胞核的方法稍作修改即可获得伸展的染色体,离心可产生剪切力使染色体的伸展长度增加 20 倍。用这种方法制备的单个染色体形态仍可识别,其原位杂交信号的作图与正常的中期染色体作图相同。

非中期相染色体 只有在细胞分裂中期染色体才高度收缩，其他时期处于松弛状态。从提高分辨率考虑，间期染色体更为适用。已有一些改进的染色体制备方法可获得既松散又保持分辨特征的染色体。在分辨率达到 25 kb 时，染色体已无可区分的外观形态，对杂交探针的定位将缺少整体坐标的参考。如果已经获得基本的位置信息，极度松弛的染色体可用于小区段分子标记排序研究。

3.4 辐射杂种作图

限制性作图虽然快速，并可提供详细的信息，但不适合大基因组。克隆重叠群构建程序复杂，费时费力。指纹作图对大基因组仍存在不少问题，如选用的限制酶不当或重复序列干扰，会出现误排。至于原位杂交因其操作困难，资料积累慢，一次实验定位的分子标记不超过 3~4 个。因此绘制详细准确的物理图还必须寻找更为有效的技术。

用于构建详细的大基因组物理图的主流技术为 STS 作图 (STS mapping) (Olson M 等, 1989)，主要为辐射杂种 (radiation hybrid) 作图。这一技术的主要原理是：①基因组中单一序列标签位点 (sequence tagged site, STS) 都有独一无二的序列组成，在染色体上的位置是确定的。②位于染色体上近邻排列的 STS 是机械连接的，在外力作用下断裂 DNA 或部分酶切降解 DNA 时，2 个不同的 STS 出现在同一片段的机会取决于它们在基因组中的相对位置。凡是位置靠近的 STS 有最大的机会同时出现在同一个 DNA 片段中。如果它们相距甚远，有时会在同一片段，有时在不同片段 (图 3.10)。相距越远，出现在不同 DNA 片段中的概率越大。当两个片段含有同一 STS 序列时，可以确认这两个片段彼此重叠。

用于作图的 STS 通常是一小段长度在 100~500 bp 的 DNA 序列，由于每个基因组仅一

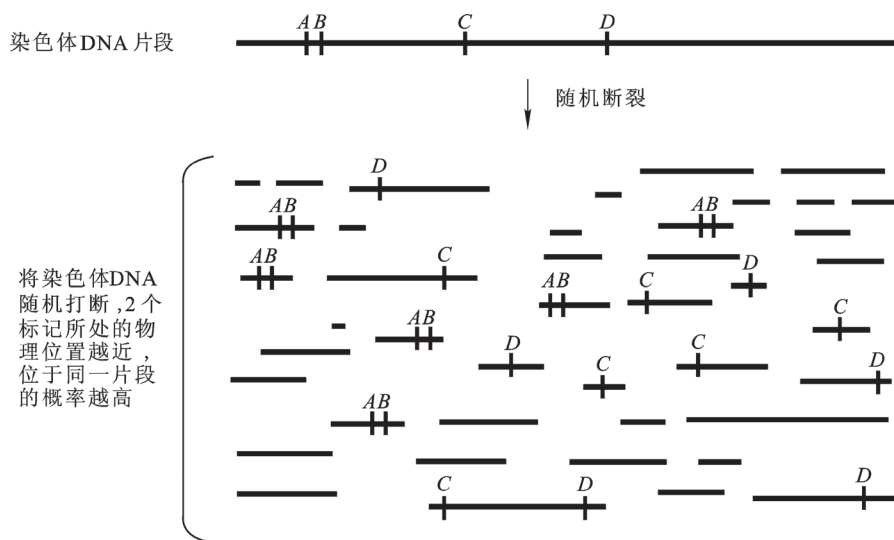


图 3.10 STS 作图原理

随机收集的 STS 片段分布在整条染色体上，图中标示了 4 个 STS 位置，其中 A 与 B 紧密连锁。随机打断的染色体 DNA 长度不一，彼此靠近的两个标记有很大的概率位于同一片段，相距较远的标记分开的概率很高

份拷贝, 很易分辨。在进行 STS 作图时, 必须收集来自同一染色体或整个基因组随机断裂的 DNA 片段。不同 DNA 片段之间有各种可能的重叠, 可以覆盖整个作图区段。依次采用单个 STS 检测它们所在的 DNA 片段, 然后根据它们彼此的重叠关系逐段绘制 DNA 物理图。一般可用杂交但更多的是用 PCR 来分析 STS 所在位置, 这一程序已经机械化, 可自动操作。STS 的分析资料可用来估算 2 个分子标记之间的相对距离, 其原理与连锁分析大同小异 (Cox DR 等, 1990)。连锁分析中, 位点间的图距与 2 个标记之间的交换频率有关。STS 作图中, 标记之间的图距依赖于两点之间断裂的频率。

3.4.1 序列标签位点

作为合格的 STS 要满足 2 个条件: 首先, 它应是一段序列已知的片段, 可据此设计 PCR 反应来检测不同的 DNA 片段中是否存在这一序列。其次, STS 必须在染色体上有独一无二的位置, 即每个 STS 都是单一序列。如果某一 STS 在基因组中多个位点出现, 那么由此得出的作图数据将是含混不清的。

上述标准不难达到, 现已有许多寻找 STS 的办法, 最常用的有以下几种:

(1) 表达序列标签 (expressed sequence tag, EST)。这是一些从 cDNA (complementary DNA) 克隆中找到的小段序列 (Marra MA 等, 1998)。因为细胞中的 mRNA 来自编码蛋白质基因, 因此 cDNA 代表了 mRNA 所在细胞中表达的基因。EST 被视为一种获取重要基因的捷径, 即使其序列不全, 仍具有重要价值。EST 可转变为 STS, 条件是这个 EST 来自单拷贝基因而非基因家族成员, 或者位于基因的非翻译区。在基因的非翻译区, 同一基因家族的不同成员其序列常常具有专一性。

(2) SSLP。我们曾讨论过微卫星序列和其他的 SSLP 在遗传作图中的应用。SSLP 像 STS 一样也可用于物理作图。具有多态性并已在连锁分析中定位的 SSLP 具有特别的价值, 因其可直接建立遗传图与物理图之间的联系。

(3) 随机基因组序列。从克隆的 DNA 中随机测序可获得有用的序列, 也可从数据库中寻找所感兴趣的某些序列。

3.4.2 辐射杂种作图的程序与方法

STS 作图程序的第二个内容是将已知的 STS 定位在染色体或基因组中的 DNA 区段, 这些 STS 标记和作图用的染色体区段以及 DNA 克隆有时又称为作图试剂 (mapping reagent)。

辐射杂种作图 辐射杂种 (radiation hybrid) 是含有另一种生物染色体片段的啮齿类细胞 (McCarthy L, 1996)。这一技术最早见于 20 世纪 70 年代, 由 Goss SJ 和 Harris H 首创 (1975)。他们发现将人体细胞暴露在 $3\,000 \sim 8\,000$ 拉德 (rad)* 剂量的 X 射线中可引起染色体随机断裂, 辐射的 X 射线剂量越大, 产生的染色体片段越小。经强辐射处理的人体细胞很快死亡, 但若在辐射后立即将处理过的细胞与未辐射的仓鼠或其他鼠类细胞融合, 有些人体细胞染色体片段将会整合到鼠类染色体中进行扩增, 因此又称为辐射与融合基因转移 (irradiation and fusion gene transfer, IFGT)。化学试剂 PEG (polyethylene glycol) 或仙台病毒

* $1\text{ rad}=1 \times 10^{-2}\text{ Gy}$ 。

(sendai virus) 可诱导细胞融合 (图 3.11)。并非所有参与融合的鼠类细胞都会保留来自人体细胞的染色体, 因此要选择一种方法来鉴别杂种细胞。常规的办法是选用一种不能合成胸苷激酶 (thymidine kinase, TK) 或次黄嘌呤磷酸转移酶 (hypoxanthine phosphoribosyl transferase, HPRT) 的缺陷细胞, 这种细胞在添加次黄嘌呤、氨基蝶呤 (aminopterin) 和胸苷的培养基 (HAT) 中是致死的。将融合的细胞置于 HAT 培养基中, 凡是从人体细胞中获得编码 TK 和 HPRT 基因的杂种细胞都可在选择性培养基中生长, 由此获得一系列随机插入人类 DNA 片段的杂种细胞。通常选择到的每个阳性杂种细胞系可保留 5 ~ 10 Mb 大小的人细胞染色体片段, 约占人类基因组的 15% ~ 35%。人体染色体片段保留在杂种细胞中的比率可由滞留率表示, 根据特定的公式计算 (Boehnke M 等, 1991; Lunetta KL 等, 1994)。这些杂种细胞的集合体称为辐射杂种群 (radiation hybrid panel), 可作为一种作图试剂用于 STS 定位。每次从杂种细胞分离 DNA, 用 PCR 检测杂种细胞中含有的 STS 标记, 根据成对 STS 出现的频率, 可判断这些标记是否连锁以及连锁程度。

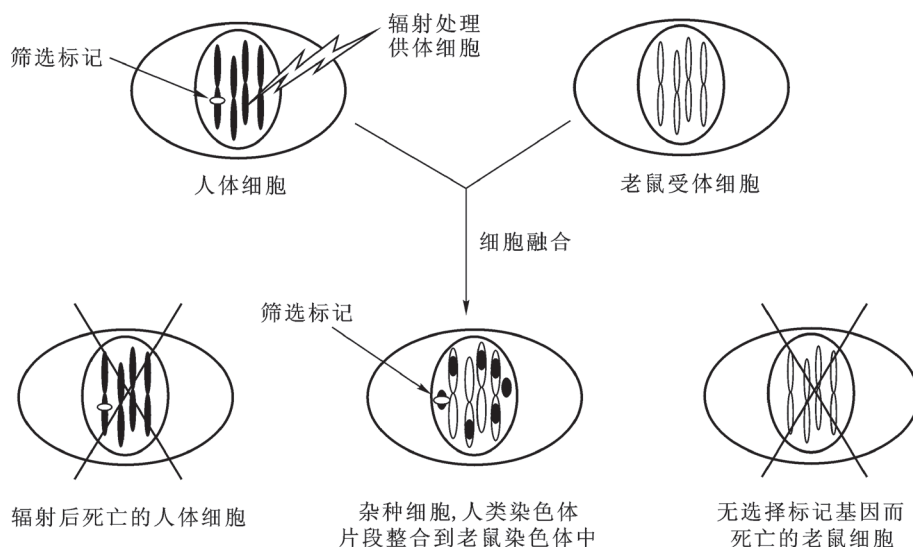


图 3.11 辐射杂种

作为供体的人体细胞含有标记基因, 经辐射处理后染色体被打断成长度不一的片段。X 射线辐射剂量越高, 产生的片段越小。辐射处理的人体细胞与仓鼠细胞融合后产生的辐射杂种可在筛选培养基上存活

第二种辐射杂种群是仅含特定的单个人体染色体, 这些单个染色体可在啮齿类细胞中稳定传代。将这些杂种细胞经辐射处理后, 再与仓鼠细胞融合, 经选择后可得到一批杂种细胞, 它们或者含有人体染色体片段, 或者仓鼠染色体片段, 或者兼而有之。用人类基因组中广泛分布的重复序列, 如短散在元件 (short interspersed nuclear element, SINE) Alu 作为探针可从上述杂种细胞中挑选出只含人体染色体片段的杂种细胞, 其他杂种细胞弃之不用。Alu 序列在人基因组中的重复拷贝数约为 7.5×10^5 , 平均每 4 kb 即含 1 个 Alu, 用这一方法可筛选绝大多数含人体 DNA 片段的杂种细胞。一旦确证含有来自已知单个人体染色体片段的杂种细胞后, 即可采用 PCR 方法检测其中所具有的 STS 标记以及连锁情况。

辐射杂种的作图单位为厘镭 (centiRay, cR), 其定义为 DNA 分子暴露在 N rad X 射线剂

量下两个分子标记之间发生 1% 断裂的频率。显然两个标记在染色体上的位置离开越远, 其间发生断裂的可能性越高。由于断裂的染色体片段在杂种细胞中随机整合, 因而两个标记滞留在同一杂种细胞中的比例与其之间的连锁关系成正比。图 3.12 为人类 21 号染色体部分区段辐射杂种图, 表 3.3 为人类基因组辐射杂种图与遗传图图距比较。

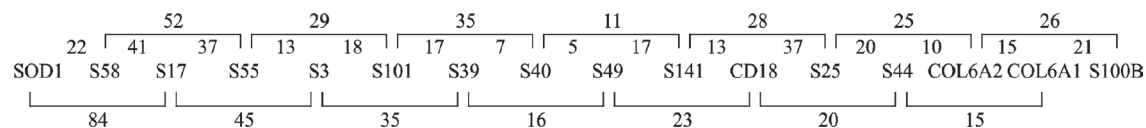


图 3.12 人类 21 号染色体辐射杂种物理图

图中表示探针的排列次序以及相邻每一对探针之间的图距 (上下框线距离), 图距单位为厘镭 (cR)。端粒靠近右边 S100B 所在位置 (引自 Burmeister 等, 1991)

表 3.3 人类辐射杂种图的若干特征

染色体	物理长度 /Mb	cM/cR	kb/cR	平均滞留率 /% (范围)
1	263	0.20	197	21.1 (14.3 ~ 50.6)
2	255	0.21	225	23.8 (14.3 ~ 37.5)
3	214	0.27	233	26.9 (14.3 ~ 41.1)
4	203	0.28	256	24.3 (11.3 ~ 44.0)
5	194	0.29	272	26.9 (17.3 ~ 49.4)
6	183	0.24	243	30.6 (21.4 ~ 49.4)
7	171	0.23	229	27.7 (19.0 ~ 44.6)
8	155	0.26	271	31.0 (31.0 ~ 45.8)
9	145	0.38	305	22.1 (14.3 ~ 27.4)
10	144	0.26	253	28.5 (16.2 ~ 56.3)
11	144	0.30	270	25.7 (20.2 ~ 38.7)
12	143	0.21	234	32.1 (23.2 ~ 50.6)
13q	98	0.22	179	22.9 (16.7 ~ 36.3)
14q	93	0.34	208	32.0 (21.1 ~ 57.1)
15q	89	0.36	203	31.7 (24.4 ~ 45.2)
16	98	0.43	201	30.5 (21.6 ~ 39.9)
17	92	0.23	147	61.6 (33.9 ~ 93.5)
18	85	0.30	172	34.3 (20.7 ~ 43.5)
19	67	0.20	110	32.2 (21.4 ~ 45.8)
20	72	0.30	191	31.0 (16.1 ~ 41.1)
21q	39	0.31	151	50.1 (39.3 ~ 71.3)
22q	43	0.95	185	36.6 (30.4 ~ 50.6)

续表

染色体	物理长度 /Mb	cM/cR	kb/cR	平均滞留率 /% (范围)
X	164	0.31	231	18.4 (11.9 ~ 30.4)
基因组	3 154	0.28	208	29.2 (11.3 ~ 93.5)

人类基因组辐射杂种作图最初是采用特定染色体而非整个基因组,因为单个染色体作图所用的杂种数量要比全基因组少得多。一个高分辨力的单个染色体的作图要求约 100 ~ 200 个杂种群,很适合常规的 PCR 检测程序。全基因组和单个染色体的辐射杂种群的构建方法不同,前者只涉及人体细胞辐射再与老鼠细胞融合,后者需先辐射仅含单个人类染色体的老鼠细胞然后再与老鼠细胞融合,因而含较多老鼠 DNA 及较少人 DNA。因此在单个染色体杂种群中,人体细胞 DNA 的含量要远低于全基因组杂种群细胞。所幸的是,由于方法的改进,现在已能用不到 100 个全基因组辐射杂种系进行详细作图。采用不同的辐射剂量,可以获得分辨率不同的辐射杂种图,有很好的灵活性。全基因组辐射杂种作图 (whole genome radiation hybrid mapping) 在人类基因组计划 (HGP) 的物理图绘制中占有十分重要的地位,是人类基因组计划的核心内容之一。

人类全基因组辐射杂种群 Genebridge 4 采用了 3 000 rad 辐射剂量,共有 93 个人体 DNA 杂种系,片段大小约 10 Mb (Gyapay G 等, 1996)。G3 辐射杂种群采用了 10 000 rad 辐射剂量,获得 83 个杂种系,片段大小平均 4Mb (Stewart EA 等, 1997)。TNG4 辐射杂种群采用了 50 000 rad 辐射剂量,获得 90 个杂种系,片段大小平均 800 kb (Lunetta KL 等, 1996)。Schwartz DC 等 (1993) 采用 G3 辐射杂种群,利用包括 10 487 个 STS 在内的共计 30 000 分子标记,绘制了一份平均密度达到 80 kb 的人类辐射杂种图,为全基因组测序和序列组装奠定了扎实的工作基础。

辐射杂种作图对于一些缺少遗传作图资源的模式生物基因组图绘制具有重要意义,特别是牛和马这类妊娠期较长的大生物,它们的单仔生殖只能提供极有限的作图数据。其他如河豚鱼,因其繁殖困难不适宜经典遗传学分析,辐射杂种作图是一种理想的方法。斑马鱼 (zebrafish) 的辐射杂种作图已于 1999 年完成 (Geisler R 等)。Senger F 等 (2006) 绘制的海鲷 (*Sparus aurata*) 全基因组 RH 图,含有 440 个分子标记,分辨率达到 2 Mb。

此外,辐射杂种作图原理还可用于一些遗传图绘制存在很大难度的作物。面包小麦是由 2 倍体山羊草 (*Aegilops tauschii*) (染色体组型 DD) 与硬粒小麦 (*Triticum durum*) (染色体组型 AABB) 杂交而来的异源 6 倍体 (AABBDD),单套染色体数目 $n=7$ 。小麦基因组 80% 以上为重复序列,很多区段缺少遗传重组,无法构建高分辨遗传图。Kumar A 等 (2012) 通过伽马射线处理人工 6 倍体小麦 (AABBDD) 种子,然后将辐射处理种子再生的植株与 4 倍体硬粒小麦杂交,获得 5 倍体辐射杂种 (RH) 株系 (AABBD)。这些含有山羊草单套染色体组 D 的辐射杂种植株中,可检测到大量的不同 D 组染色体成员的区段缺失。由于缺失区段存在紧密连锁的分子标记,通过分子标记逐个检测,构建了高密度山羊草 D 组染色体辐射杂种物理图,分辨率小于 140 kb。

辐射杂种作图的分辨率最高可达到 50 kb,远胜于荧光原位杂交的分辨率 (1 Mb)。由于

辐射杂种作图直接以染色体区段为作图试剂, 它的可信度高于克隆重叠群绘制的物理图。

3.5 基因组整合图

基因组遗传图和物理图的绘制是根据不同的作图原理, 采用不同的作图方法进行的, 它们之间虽然有共同的基础, 但绘制的图谱仍有很大的差别。由于特定生物基因组的结构都是唯一的, 因此原则上来自同一基因组的遗传图和物理图在组成上是一致的。但是在作图过程中, 由于各种因素的干扰以及实验误差, 往往在绘制的遗传图和物理图之间在染色体的某些区段会存在同一分子标记所处位置不相同的现象。这些差异如果不加以澄清和校正, 在进入基因组序列组装阶段将会带来很大的麻烦。因此在分别完成基因组的遗传图和物理图之后, 必须将它们进行整合。

遗传图和物理图的整合必须根据多种作图方法获得的结果进行综合分析, 然后加以取舍确定。由于每个分子标记的序列是单一的, 在染色体上都只有一个特定的对应位置, 因此可以通过分子标记使克隆重叠群物理图与遗传图彼此衔接。这些用于衔接不同基因组图的分子标记又称为锚定标记 (anchored marker)。如果同一连锁群上所有的分子标记在遗传图和物理图上的排列次序都是相同的, 表明遗传图和物理图之间是吻合的。如果分子标记在遗传图和物理图上的排列次序出现差异, 说明其中一定存在错误, 必须重新验证。此外, 用于绘制遗传图和物理图的分子标记由于各种原因有些是相同的, 有些是不同的。通过遗传图和物理图的整合, 可将来源不同的分子标记归并在一张整合图上, 提高了基因组整合图 (integrated genome map) 的分子标记密度, 有利于下一步基因组的测序和序列组装。

3.5.1 人类基因组整合图

人类基因组物理图绘制稍后于遗传图。20 世纪 90 年代初, 经过相当的努力, 采用 STS 筛选法及其他指纹技术产生了一份基于克隆的重叠群物理图 (clone contig map)。这一时期物理图绘制计划的最大进展是发表了由 33 000 个 YAC (yeast artificial chromosome) 组成的全基因组重叠群物理图 (Cohen D 等, 1993)。虽然 YAC 克隆的插入片段平均为 0.9 Mb, 但很快人们就对 YAC 重叠群的价值产生了怀疑, 因为有不少 YAC 克隆含有 2 个或多个嵌合的 DNA 片段。嵌合 YAC 克隆使基因组中一些原来分散的 DNA 片段在作图时错误地连接到彼此相邻的位置, 引起大范围的错位。这些问题促使研究者转而采取辐射杂种进行 STS 标记作图, 并对 YAC 物理图进行校正。这一工作主要归功于美国麻省理工大学基因组研究中心的 White Head 研究所 (The White Head Institute/MIT Genome Center in Massachusetts)。他们采用辐射杂种作图方法, 将 6 193 个基因组位点和 5 264 个遗传位点整合, 绘制了一份含有 15 086 个 STS 标签的人类基因组物理图 (Hudson TJ, 1995)。次年在这份物理图的基础上又添加了 2 010 个 STS, 其中大多数是 EST, 从而将许多蛋白质编码基因定位在物理图上 (Schuler GD 等, 1996)。这份物理图的密度为每 100 kb 一个标记, 达到人类基因组最初确立的物理图作图目标。1998 年在已有的人类基因组物理图基础上, Deloukas P 等 (1998) 进一步扩大分子标记的范围, 采用 30 181 个人类基因 cDNA, 结合其他来源的分子标记共组成 41 466 个 STS 标签, 包括 39 886 个 cDNA, 1 641 个 SSR, 13 个端粒标记, 以及源自 G3RH 的 5 013 个 cDNA,

2 091 个 SSR, 绘制了一份密度更高的辐射杂种物理图, 标记的平均密度约 75 kb。

除了上述的辐射杂种作图外, 另一份以 BAC 克隆为基础的物理图也于 2001 年发表 (The International Human Genome Mapping Consortium, 2001)。这份物理图采用指纹作图法将覆盖人类基因组约 15 个单倍体的 283 287 个 BAC 克隆组建成 7 133 个簇群 (cluster), 同时利用 13 695 个来自遗传图和辐射杂种图的分子标记将 96 283 个不同的 BAC 克隆锚定到基因组连锁图上。进一步再利用 3 412 个 BAC 作为探针, 采用荧光原位杂交 (fluorescence *in situ* hybridization, FISH) 将这些 BAC 克隆定位到染色体的细胞遗传图上。这些作图法的交叉运用, 以共同的分子标记为媒介使 7 133 个 BAC 簇群与遗传图、辐射杂种物理图和细胞图衔接, 构成了一幅可用于指导基因组测序与序列组装的基因组整合图。

3.5.2 水稻基因组整合图

早期水稻基因组物理图的绘制采用 YAC 载体。1997 年, Umehara Y 等构建了水稻第一张物理图, 含 5 701 个 YAC 克隆, 其中 2 117 个单一的 YAC 克隆分布在 12 条染色体上, 跨度 216 Mb (Megabase, 百万碱基对), 覆盖水稻基因组 50%。2001 年, 日本水稻基因组计划组织 (RGP) 又构建了一个跨度 270 Mb 的 YAC 克隆物理图, 由 6 934 个 YAC 组成, 覆盖 63% 水稻基因组。

鉴于 YAC 克隆作图的许多难以克服的缺点, 2001 年 RGP 为了克服 YAC 克隆的局限性, 又以 PAC 为载体构建了水稻 (*Oryza sativa japonica*) 日本晴 (Nipponbare) 基因组文库。此文库由 72 000 个 *Sau*3A I 酶切克隆组成, 平均插入片段长 120 kb, 覆盖水稻基因组 16 倍。RGP 并对 PAC 克隆进行了排序, 采用已定位的分子标记作为路标, 将鉴定的 PAC 克隆及其重叠群锚定到染色体物理图上 (Chen M 等, 2002)。与此同时, 美国康纳尔大学基因组研究所也构建了两个 BAC 文库, 一个为 *Hind*III 部分酶切的 BAC 文库, 含 37 000 个 BAC 克隆, 插入片段平均长度为 128.5 kb。另一个为 *Eco*R I 部分酶切的 BAC 文库, 含 56 000 个 BAC 克隆, 插入片段平均大小为 120 kb, 两者覆盖水稻基因组约 26 倍。这些 BAC 克隆也根据分子标记所在的位置逐一锚定到染色体物理图 (Chen M 等, 2002)。

比较水稻遗传图与物理图上相邻分子标记的相对距离会发现, 染色体遗传图和物理图的相对位置有很大差别, 造成这一状况的原因是因为同源染色体不同区段的交换值不同。为了获得一份可用于指导全基因组测序的基因组图, 必须将遗传图和物理图加以整合: 检测同一分子标记位于遗传图和物理图上的次序是否相同。由此产生的整合图可作为水稻基因组测序与组装的基本骨架图 (图 3.13) (Chen M 等, 2002)。

3.5.3 玉米基因组最短重叠路径整合图

玉米是谷类作物中基因组最大的粮食作物, 染色体组数目为 $n=10$ 。玉米基因组大小在 2 300 ~ 2 700 Mb 之间 (Rayburn A 等, 1993), 约为水稻基因组 (390 Mb) 的 7 倍 (Yu J 等, 2002)。由于玉米基因组在进化过程中曾经发生过逆转座子的极度扩张 (见 13.3.1), 因而基因组中约 85% 的序列为转座子成分 (transposable elements, TE)。玉米基因组中大量的逆转座子成分分散在不同的染色体区域, 它们之间存在许多序列相同的拷贝。这一组成特点给玉米基因组的物理图谱绘制带来极大的困难, 因为分散在不同区域的相同序列的逆转座子在进

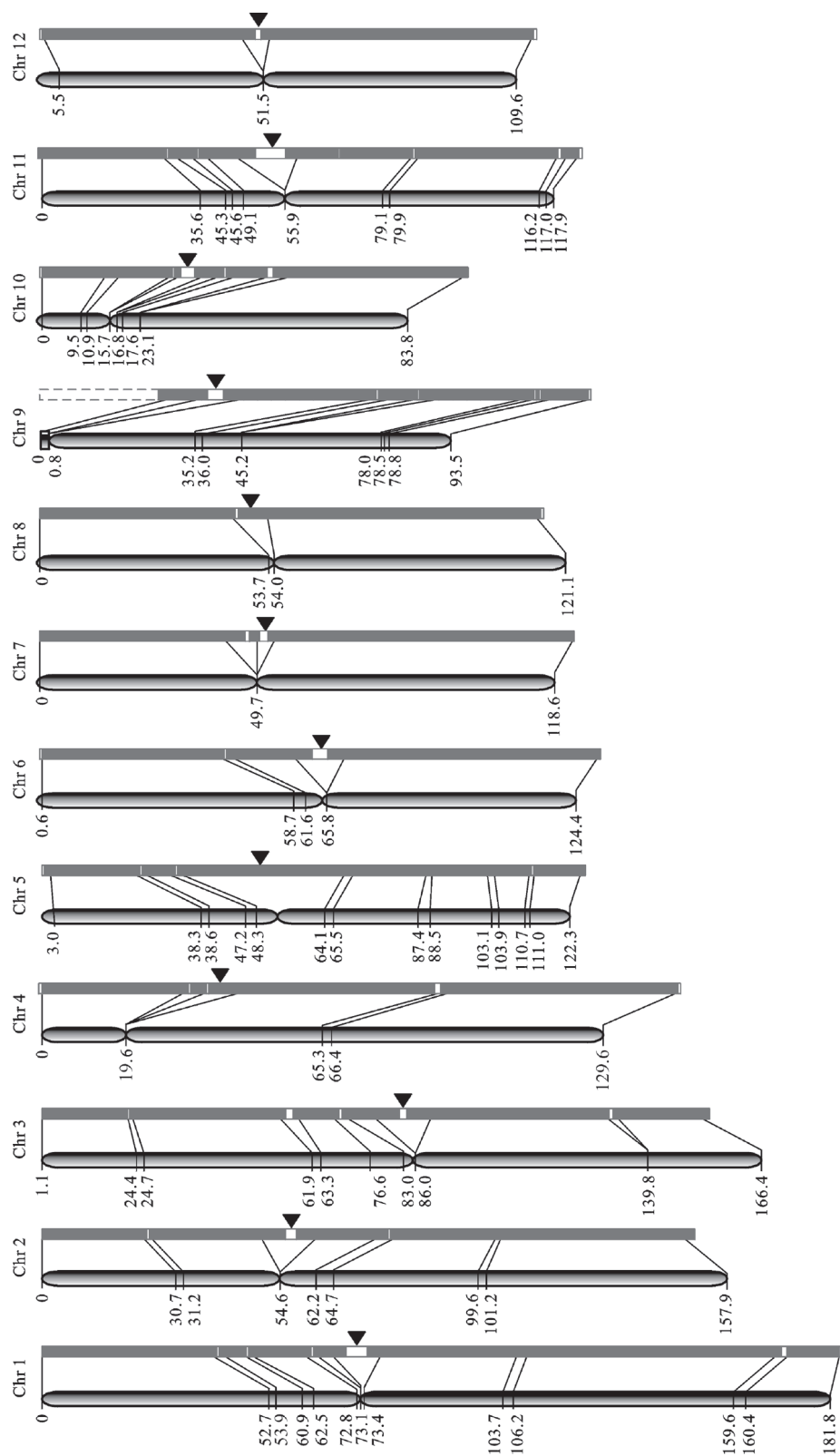


图 3.13 水稻染色体遗传图与物理图的整合

由于染色体不同区段交换值的差别，遗传图与物理图的相对位置有很大差别。本图为水稻 12 条染色体遗传连锁图（每条染色体左侧）与 BAC 克隆重叠群物理图（每条染色体右侧）的比较。◀ 示着丝粒位置

行指纹作图时常常会发生错误拼接,使重叠群的组装发生偏差。为了克服这一困难,玉米基因组整合图采取了不同的构建路线。

Davis GL 等(1999)利用玉米基因组 DNA 和 cDNA 克隆,采用同工酶和 EST 作为分子标记,构建了含有 1 736 个分子标记的玉米遗传连锁图。Sharopova N 等(2002)采用 978 个 SSR 标记,构建了另一份玉米遗传连锁图。Young-Sun Yim 等(2002)以玉米自交系 B73 为材料,构建了 3 个不同的基因组 DNA BAC 文库,经过滤筛选获得 331 776 个 BAC 克隆,插入片段平均长度为 154 kb。其中 *Hind* III BAC 文库覆盖单倍体基因组 6 个当量(6X),*Eco*R I 和 *Mbo* I BAC 文库覆盖单倍体基因组分别为 7 个当量(7X)。

最短重叠路径作图法 在上述已有的遗传连锁图和 BAC 文库的基础上,Wei F 等(2009)采取了一种称为最短重叠路径(minimum tiling path, MTP)的作图法构建了一份覆盖玉米基因组 93%,物理长度为 2 120 Mb 的整合图。其操作过程如下:①首先从 BAC 文库中挑选了 3 200 个高质量的内部含有分子标记的种子 BAC 克隆,每个 BAC 克隆的两个末端均已完成测序。②收集了玉米 *Eco*R I/*Mbo* I BAC 克隆末端序列 BES(BAC end sequences)815 473 个(70% 成对)以及 827 571 个 fosmid 克隆 BES(72% 成对),用于指纹重叠群构建。③为了使连接 BAC 克隆的重叠区长度保持最短最佳长度,参考了基于荧光标记的玉米高质量 BAC 克隆指纹图数据(high information content fingerprint, HICF)(Nelson WM 等,2005),使拼接 BAC 克隆之间重叠区长度维持在 39 kb 距离。这一跨度超出了玉米逆转座子平均 15 kb 的长度,从而有效地避免了含逆转座子的拼接重叠区段在 BAC 步移时产生的差错。最短重叠路径作图法使玉米基因组物理重叠群数目减少到 435 个,并将 405 个重叠群(跨度 2 103.4 Mb)锚定到遗传连锁图上。此外,利用 8 315 个遗传标记确定了覆盖基因组物理图 94%(1 993 Mb)的 336 个重叠群的排列顺序和方向,为全基因组的测序奠定了可靠的基础(Schnable PS 等,2009)。

复习思考题

1. 什么是基因组物理图?物理图与遗传图有何不同?
2. 什么是 BAC 载体?它有什么优缺点?请说明 BAC 载体各组成部件的功能?
3. 如何将重叠群锚定到遗传图上?
4. STS 作图依据的原理是什么?
5. 叙述辐射杂种作图的过程。辐射杂种作图的图距单位是如何确定的。
6. 什么是 DNA 指纹?有哪些 DNA 指纹?
7. 如何进行定位克隆?

网上更多学习资源……

※ 教学课件

※ 教学视频

※ 自测题

※ 参考文献