

Chroniques génomiques



médecine/sciences 1994 ; 10 : 898-902

*Carte physique
du génome humain :
l'état des lieux*

Bertrand R. Jordan

Directeur de recherche au Cnrs, responsable du groupe « Structure du Génome et fonctions immunitaires ». CIML, Inserm/Cnrs, case 906, 13288 Marseille Cedex 9 France.

Septembre 1992 : un article de l'équipe de Daniel Cohen (CEPH-Généthon) dans la très cotée revue *Cell* [1] décrit une technique de *fingerprint* (empreinte) de chromosomes artificiels de levure (YAC). Appliquée à une échelle jusque-là inimaginable à la banque de « MégaYacs » du CEPH, elle permet de déterminer les positions relatives de ces clones, et donc de construire la carte physique de l'ensemble du génome humain. La méthode, mise en œuvre à grand renfort d'appareillage sophistiqué et d'informatique, devrait, selon les auteurs, produire très prochainement un taux de couverture de 90 %. L'impact médiatique est considérable : on pensait jusque-là que ces cartes n'étaient réalisables que chromosome par chromosome, et qu'elles ne seraient pas établies avant plusieurs années. Aux États-Unis – jusque-là très largement en tête dans ce domaine –, on accuse le coup ; tel patron de laboratoire rassemble son équipe et lui annonce qu'il va falloir changer de sujet puisque les Français ont tout raflé... Dans notre pays règne un certain triomphalisme ; des journalistes annoncent déjà que la carte est terminée, les gènes trouvés et que les applications médicales vont suivre à brève échéance. Début mai 1994, à Cold Spring Harbor, se tient l'annuel congrès *Genome Mapping and Sequencing*, de loin la meilleure réunion sur le sujet. Quelques mois plus tôt est paru dans *Nature* un court article présentant « Une carte physique de

première génération du génome humain », signé par Daniel Cohen, Ilya Chumakov et Jean Weissenbach [2]. Le balisage présenté couvre entre 50 % et 80 % du génome et, malgré l'extrême concision de l'exposé, il semble bien que les moyens employés n'ont plus grand-chose à voir avec la méthode vantée dans *Cell* un an plus tôt. A Cold Spring Harbor – comme déjà à Kobe lors de *Human Gene Mapping 93*, début novembre –, on constate que chaque *Genome Center* poursuit la construction de la carte physique de « son » chromosome, apparemment comme si de rien n'était. Difficile de s'y retrouver dans cet imbroglio ; c'est cette situation complexe que je vais essayer d'analyser dans cette Chronique, me fondant pour l'essentiel sur les publications des uns et des autres ainsi que sur les communications ou les « posters » présentés tout récemment à Cold Spring Harbor.

Rappelons pour commencer les éléments techniques de base. Une carte physique fiable et opérationnelle doit reposer, tout le monde s'accorde maintenant à le dire, sur l'alignement de clones le long du génome. Elle représente chaque chromosome par un ensemble de segments d'ADN clonés dont les positions relatives sont connues, de préférence avec précision et certitude. Le prototype à cet égard est la carte du génome d'*Escherichia coli* publiée en 1987 par Kohara *et al.* [3] : il avait déterminé la carte de restriction de 3 400 phages contenant chacun un

TIRÉS A PART

B.R. Jordan.

segment d'ADN bactérien de 15 à 20 kilobases et les avait positionnés le long des 4 700 kilobases du génome de la bactérie. Carte fiable puisqu'on peut reprendre n'importe lequel de ces clones et vérifier sa carte ainsi que ses recouvrements avec ses voisins, carte utile puisque tout chercheur intéressé par l'examen détaillé d'une région du génome située par exemple à 37 minutes (la carte génétique d'*Escherichia coli* emploie ces unités, liées au transfert d'information entre bactéries par conjugaison) peut demander le ou les clones correspondants et les étudier tout à loisir. Les 476 clones représentant le « jeu » minimum pour couvrir ce

génom sont d'ailleurs maintenant distribués par la firme Takara Biochemicals, sous forme d'un filtre contenant l'ADN de ces clones et appelé *the Escherichia coli gene mapping membrane*.

Quelques téméraires s'étaient lancés dès 1987-1988 dans la construction de telles cartes pour un chromosome humain. L'entreprise semblait vouée à l'échec : un chromosome moyen mesure une centaine de mégabases alors que les cosmides – vecteurs les plus performants à l'époque – véhiculent au maximum 40 kilobases. C'étaient donc, au bas mot, dix mille cosmides qu'il aurait fallu analyser en

détail, tâche apparemment surhumaine d'autant que l'on savait dès le départ certaines régions « inclonables » dans ces vecteurs. La cartographie physique du génome humain n'est devenue réellement envisageable qu'avec l'invention par Burke, Carle et Olson [4] des chromosomes artificiels de levure (YAC, *yeast artificial chromosomes*) capables de cloner des segments d'ADN humain longs de plusieurs centaines de kilobases : le nombre de « pièces » à assembler pour couvrir un chromosome est alors de l'ordre du millier, et chacun sait qu'un puzzle à mille pièces est infiniment plus facile qu'un puzzle à dix mille pièces...

Les *Genome Centers* américains, créés à partir de 1989, se sont alors donné pour objectif l'établissement de cartes, chromosome par chromosome. Parmi les stratégies employées, la plus généralement admise fut celle du *STS* (*sequence tagged site*) *content mapping* dont le principe est représenté sur la *figure 1*. En bref, il s'agit d'abord de baliser le chromosome choisi par un millier de STS, petites séquences définies chacune par un couple d'amorces PCR et uniques dans le génome. Ces STS peuvent être obtenus en prenant des sondes déjà connues pour provenir de ce chromosome, en les séquençant partiellement, en spécifiant un couple d'amorces à partir de cette séquence – et en vérifiant que ce dernier permet l'amplification de la région choisie et d'elle seule. Les STS peuvent aussi être des « microsatellites » déjà déterminés lors de la cartographie génétique du même chromosome – ce qui offre l'avantage d'intégrer *ipso facto* la carte physique et la carte génétique puisque les balises employées sont les mêmes. Les STS servent alors au criblage par PCR d'une banque de YAC : banque spécifique du chromosome étudié si elle existe, banque générale sinon, la spécificité des STS assurant en principe l'obtention d'un YAC provenant du bon chromosome. Si la position des STS le long du chromosome est connue, cela indique du même coup celle des YAC correspondants ; dans le cas contraire, il faudra ordonner les

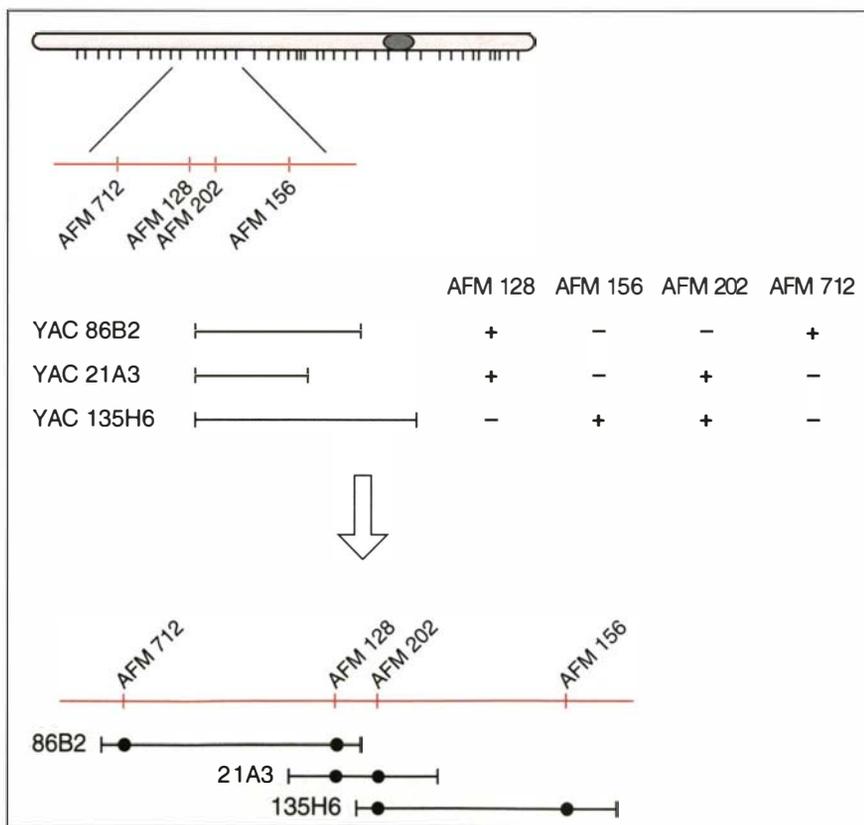


Figure 1. Construction d'une carte physique par STS content mapping. Les STS, précédemment définis et positionnés le long du chromosome par analyse génétique (en haut), sont testés (par PCR) sur les YAC de la banque (en employant divers schémas de mélange afin de limiter le nombre de réactions à effectuer). Deux YAC, positifs pour le même STS, présentent a priori un recouvrement et l'on peut ainsi progressivement assembler des « contigs » (bas de la figure).

STS, soit par cartographie génétique, soit à l'aide d'un « panel » d'hybrides somatiques contenant différentes régions du chromosome. Deux YAC contenant le même STS seront forcément chevauchants, et la détermination de leur position approximative par hybridation *in situ* autorisera quelques vérifications. Ce processus est naturellement long, complexe et coûteux. Le « tarif » qu'appliquaient les organismes comme le DOE ou le NIH était de l'ordre de dix millions de dollars (soixante millions de francs) pour un chromosome : cette somme permettait de faire fonctionner durant trois ans un groupe d'une trentaine de personnes, équipé d'un outillage performant (séquenceurs, robots), et devait aboutir à la carte physique du chromosome choisi ou, du moins, à des progrès très significatifs dans cette direction. Seuls les États-Unis semblaient s'engager dans cette course : au Royaume-Uni, l'accent était mis sur la génétique et les ADNc, le programme japonais sortait à peine des limbes ; quant à la France, son Programme génome « officiel » tardait à décoller tandis que Généthon, tout nouvellement créé par l'AFM et le CEPH, paraissait centré sur la carte génétique (programme dirigé par Jean Weissenbach), les ADNc (Genexpress et Charles Auffray) et les services – l'impressionnante salle des « MarkII » avec ses vingt automates à *Southern blot* destinés aux équipes qui souhaitaient « localiser » des maladies.

En réalité, Généthon, le CEPH et Daniel Cohen préparaient une bombe. Le CEPH, au prix d'efforts prolongés et d'une remarquable persévérance, avait construit une banque de YAC contenant de très grands segments, de l'ordre de la mégabase : les fameux « MégaYacs ». Et il se livrait à l'assemblage simultané de contigs de YAC sur l'ensemble du génome humain par la technique d'empreinte mentionnée plus haut. Le principe est simple (figure 2) : l'ADN d'un clone de levure contenant un YAC est coupé par une enzyme de restriction qui produit une centaine de fragments. Ces derniers sont séparés par électrophorèse,

puis transférés sur un filtre de nylon, lequel est ensuite hybridé avec une séquence répétée humaine (séquence Alu par exemple) qui va révéler une ou deux dizaines de fragments. L'opération est ainsi effectuée pour chaque YAC de la banque ; la comparaison des empreintes permet de repérer ceux qui contiennent des fragments de taille identique – ce qui indique qu'ils présentent sans doute un chevauchement. Méthode largement employée, en Grande-Bretagne et aux États-Unis, pour l'assemblage de « contigs » de cosmides, elle n'avait jamais été mise en pratique avec succès pour les YAC ; elle fut appliquée à très grande échelle à Généthon. Ce n'était pas une mince affaire : il fallait répéter l'analyse sur plusieurs dizaines de milliers de YAC, dans des conditions autorisant une comparaison précise entre chacune de ces empreintes et chacune des cinquante mille autres. Cela posa de redoutables problèmes de standardisation des expériences, sans parler des programmes et de la puissance informatique nécessaires pour confronter cette multitude d'objets complexes. Au printemps 1992, cet assemblage semblait en très bonne voie, ce qui suscita l'article de *Cell...* et la panique des concurrents. En un ou deux ans, l'équipe de Généthon avait plus avancé que la douzaine de *Genome Centers* américains : il y avait de quoi pavoiser, et l'on ne s'en priva pas.

La carte à 90 %, annoncée pour la fin de 1992 [5], se fit pourtant attendre. Il est fréquent dans la recherche que les derniers 10 % d'un projet demandent autant d'efforts et de temps que les premiers 90 %. Ceux qui ont déterminé des séquences d'ADN ou de protéines ne me démentiront pas... Mais la lettre publiée dans *Nature* fin 1993 [2] témoigne d'un net changement de stratégie. Co-signée par Jean Weissenbach, chef d'orchestre de l'impeccable opération « carte génétique » au Généthon, elle est en fait largement fondée sur une approche du type STS *content mapping* dans laquelle les microsatellites positionnés pour la carte génétique jouent le rôle principal. Une astucieuse tech-

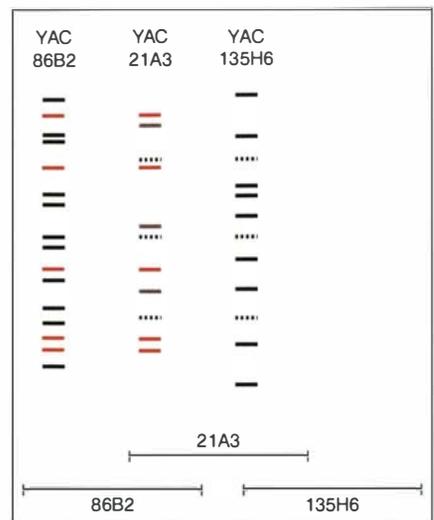


Figure 2. **Analyse des recouvrements de YAC par empreinte (fingerprint).** L'ADN de chaque clone de levure contenant un YAC est traité par une enzyme de restriction, séparé sur gel, transféré sur filtre et hybridé avec une sonde répétée humaine comme « Alu » (qui ne révèle pas les fragments d'ADN provenant des chromosomes de levure). Les positions des bandes obtenues sont comparées deux à deux ; la présence de plusieurs bandes « identiques » dans deux clones (bandes rouges entre la 1^{re} et la 2^e colonne, bandes hachurées entre la 2^e et la 3^e) donne une présomption de recouvrement entre les clones correspondants. Pour une discussion détaillée de cette méthode, voir [1].

Figure 3. **Zoom sur la carte physique.** On voit, à gauche, une partie du bras long du chromosome 17, les barres horizontales repèrent la position des STS déterminée par cartographie génétique. Les quatre jeux de bandes sur la droite indiquent le taux de couverture de cette région au niveau 1 (un seul YAC joignant deux STS), 3 (trois YAC), 5 et 7. Plus le niveau est élevé, plus la couverture est complète, mais moins elle est fiable, du fait notamment des problèmes causés par le chimérisme des YAC. (D'après [2].)

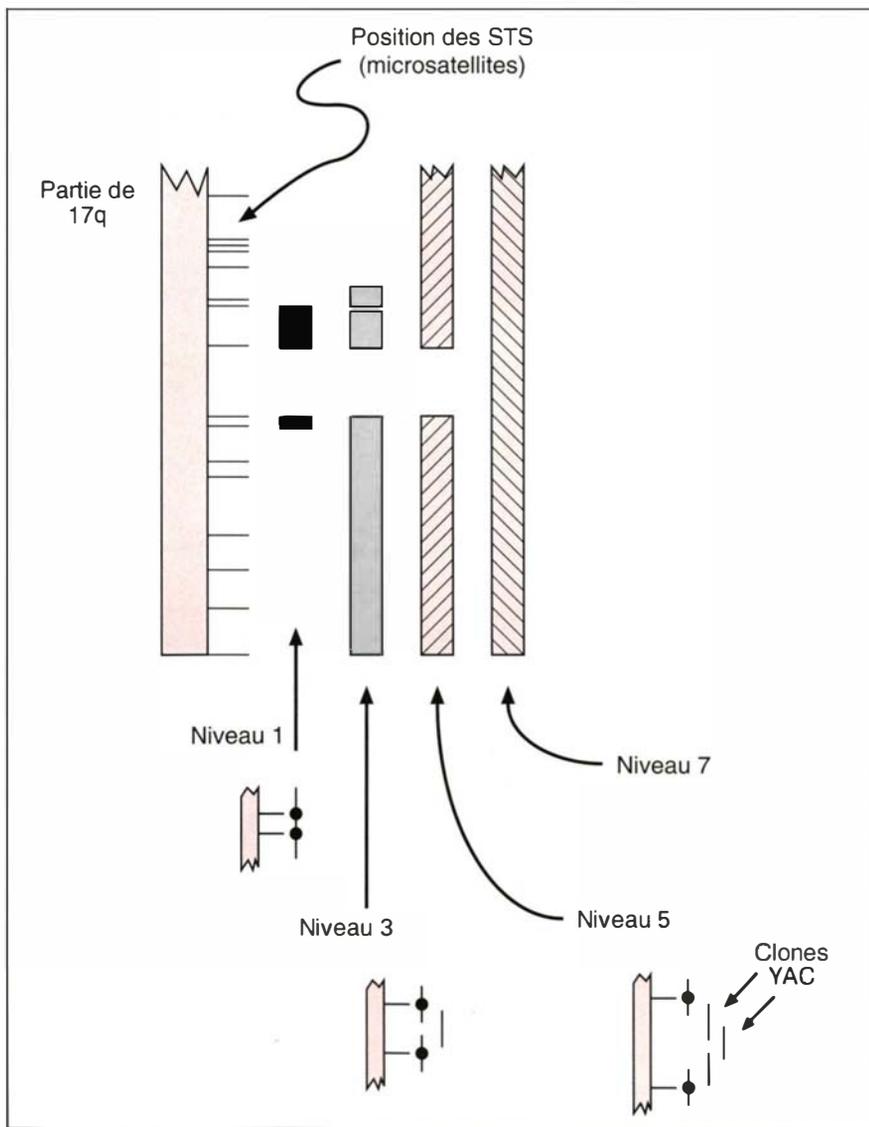
nique d'hybridation des produits Alu-PCR de YAC avec les autres YAC, ou avec les mêmes produits obtenus à partir d'hybrides somatiques, donne des informations complémentaires, et l'hybridation *in situ* effectuée sur quelques centaines de YAC confirme les positionnements. La méthode des *fingerprints*, qui devait à elle seule produire l'essentiel des données, est reléguée à un rang secondaire. Le résultat final est une carte dont le taux de couverture et la fiabilité sont inversement corrélés (figure 3) : si l'on se limite au cas où la jonction entre deux STS est assurée par un seul YAC (niveau 1), la couverture est de 10 % ; si l'on

admet entre deux STS trois YAC positionnés par Alu-PCR ou *finger-print* (niveau 3), elle monte à 30 % ; au niveau 7, elle approche 90 %, mais la probabilité d'erreurs est alors élevée.

Pourquoi tant d'incertitudes ? N'oublions pas l'échelle de l'entreprise : cinquante mille clones à étudier par diverses techniques, c'est colossal. Pensons à la facilité avec laquelle, dans nos laboratoires, on se trompe de clone, de marqueur de taille ou de boîte dans le congélateur... et l'on imaginera l'organisation quasi militaire qui s'impose pour manipuler mille fois plus d'objets en réduisant au minimum

les risques de confusion. Mais le problème est plus profond, il tient à la nature même des YAC et aux imperfections de ce système de clonage. Les chromosomes artificiels de levure ont joué un rôle primordial, et restent encore irremplaçables (quoique BAC, PAC et autres MAC* commencent à les concurrencer sérieusement) ; mais ils ont le gros défaut de présenter un chimérisme fréquent. Chimérisme, c'est-à-dire que le segment d'ADN inséré dans un YAC est en fait formé de deux (ou plus) fragments provenant d'autant de régions du génome. Cet inconvénient a été repéré assez vite, mais sa fréquence est longtemps restée sous-estimée. On annonçait 10 %, 15 %, peut-être 20 % de chimérisme. En réalité, les mesures les plus récentes, rapportées dans un article de Buddy Brownstein (soumis pour publication), indiquent un chimérisme moyen de 50 % – un clone sur deux contient deux segments d'origine différente. Cela est vrai pour plusieurs banques, et même pour celle de Rakesh Anand (dite banque ICI) que l'on a cru longtemps épargnée par ce fléau. Fléau, c'est bien le mot : car un clone chimérique dans un contig le fait « sauter » d'un chromosome à un autre, ou – ce qui n'est pas mieux – d'un point à un autre très éloigné sur le même chromosome. Cela d'autant plus que l'on cherche naturellement à employer les YAC les plus longs possibles pour la carte physique (moins de pièces) et que le taux de chimérisme augmente fortement avec la taille... Ainsi, selon Eric Lander, lorsqu'un YAC contient deux STS, ces derniers appartiennent à deux chromosomes différents dans 60 % des cas ! On voit le problème posé aux bâtisseurs de contigs, et les risques d'erreurs qui en résultent pour les programmes d'assemblage les mieux conçus.

* BAC, bacterial artificial chromosomes ; PAC, pombe artificial chromosomes ; MAC, mammalian artificial chromosomes. Ces différents systèmes de clonage promettent de propager de grands segments d'ADN exogène dans des bactéries, dans *Schizosaccharomyces pombe* ou même dans des cellules de mammifères – mais sont encore loin d'être pleinement opérationnels.



C'est ainsi qu'une bonne douzaine de *Genome Centers* américains poursuivent des programmes de cartographie physique. Le plus important, celui d'Eric Lander, créé fin 1992 un peu à l'image de Généthon, semble bien reprendre tout à zéro puisque – sur les YAC du CEPH – il applique une stratégie pure et dure de *STS content mapping* appuyée sur une automatisation poussée (une imposante machine permet d'effectuer 147 456 (1 536 fois 96) réactions de PCR en parallèle). Les autres centres, travaillant chacun sur un ou deux chromosomes, ont pris des options moins extrêmes. Ils emploient très largement les YAC du CEPH : la banque de Méga-Yac, malgré ses imperfections (qu'elle partage avec les autres banques existantes), reste la meilleure à ce jour grâce à la taille des fragments insérés et à la couverture d'ensemble. Les laboratoires des États-Unis ont d'ailleurs fait preuve d'une grande efficacité pour la dupliquer et la distribuer dès qu'elle leur a été fournie par le CEPH. Les *Genome Centers* s'appuient également sur la carte publiée fin 1993. Peu se hasardent à lui faire confiance au-delà du niveau 3, mais elle constitue un premier « bâti » (au sens des couturières) que l'on vérifie, corrige, enrichit, et qui constitue le point de départ de cartes plus restreintes mais plus solides. Chacun s'accorde à rendre hommage au CEPH non seulement pour l'œuvre accomplie, mais aussi pour la rapidité avec laquelle les réactifs (les YAC) ainsi que toutes les informations sur la carte (contigs, chemins d'un YAC au suivant, résultats du criblage par les STS...) ont été mis à la disposition de la communauté. Paradoxalement, en raison du sous-développement informatique de la plupart des laboratoires français, ces données sont souvent plus accessibles outre-Atlantique que chez nous : leur abord (libre) implique des systèmes informatiques (de préférence une station de travail) et, en tout cas, des compétences dans l'emploi des réseaux qui sont, hélas, encore peu répandues en France.

Les paradoxes apparents de la situation deviennent donc compréhensibles.

A l'échelle d'un génome aussi complexe que celui de l'homme, la différence entre carte génétique et carte physique tend à s'estomper : aucune des deux ne sera jamais complètement terminée, chacune demandera toujours plus de précision, de détails, et restera sujette à correction. Le placement d'un YAC devient probabiliste : il y a 90,99 ou 99,9 chances pour cent que la position donnée soit correcte, mais nous n'aurons jamais de certitude absolue sur 50 000 YAC répartis sur trois milliards de nucléotides. Ainsi, la carte physique CEPH-Généthon n'est ni « la fin de l'histoire » (ce que n'a d'ailleurs jamais prétendu Daniel Cohen) ni un « coup » raté. C'est une entreprise dont les résultats tangibles mais imparfaits servent de support aux avancées suivantes – ce qui est, au fond, la finalité de toute recherche ■

RÉFÉRENCES

1. Bellanne-Chantelot C, Lacroix B, Ougen P, Billault A, Beaufile S, *et al.* Mapping the whole human genome by fingerprinting yeast artificial chromosomes. *Cell* 1992 ; 70 : 1059-68.
2. Cohen D, Chumakov I, Weissenbach J. A first-generation physical map of the human genome. *Nature* 1993 ; 366 : 698-701.
3. Kohara Y, Akiyama K, Isono K. The physical map of the whole *E. coli* chromosome : application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell* 1987 ; 50 : 495-508.
4. Burke DT, Carle GF, Olson MV. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* 1987 ; 236 : 806-12.
5. Dorozynski A. Gene mapping the industrial way. *Science* 1992 (Science in Europe section) ; 256 : 463.