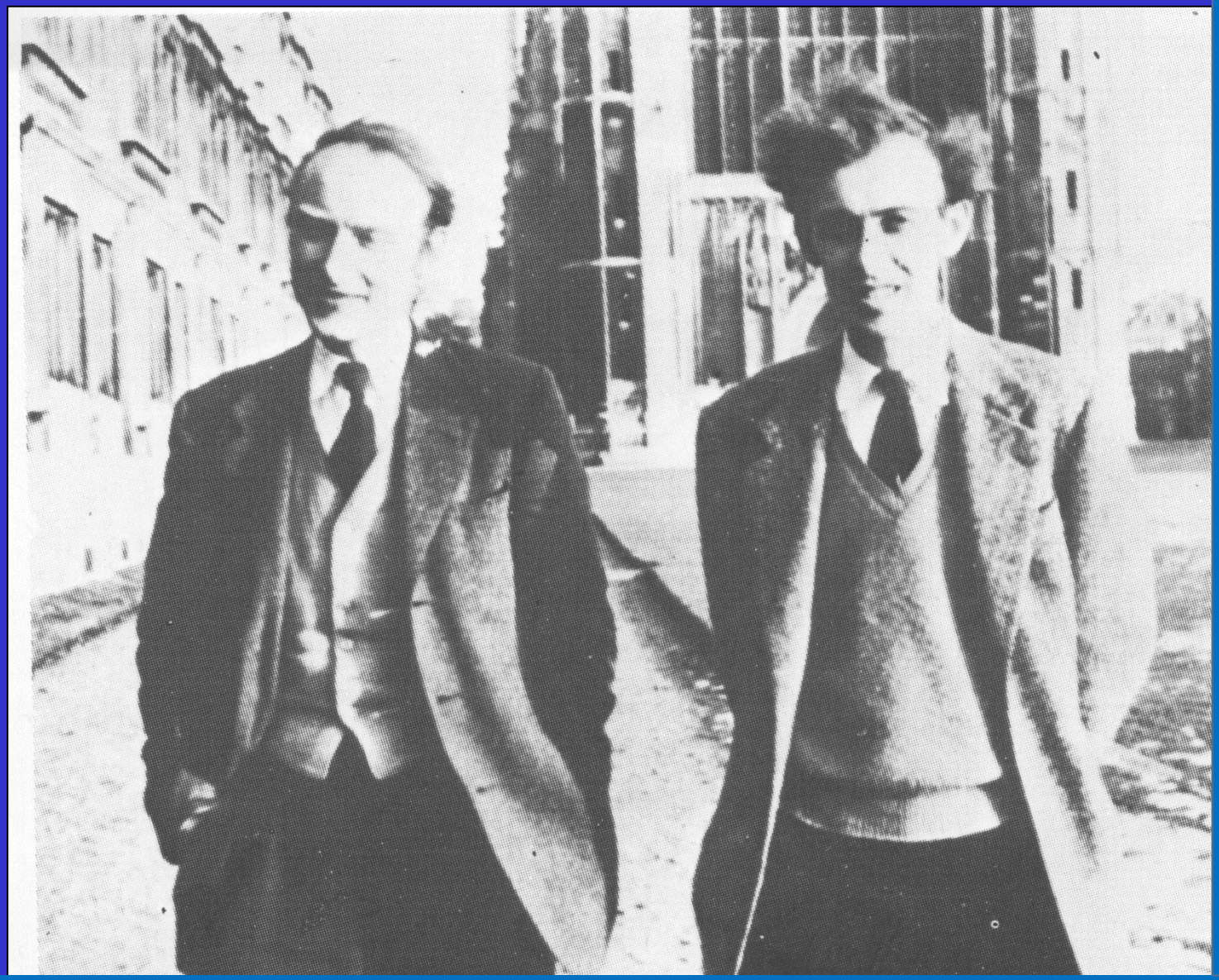
La grande aventure du séquençage du génome humain.

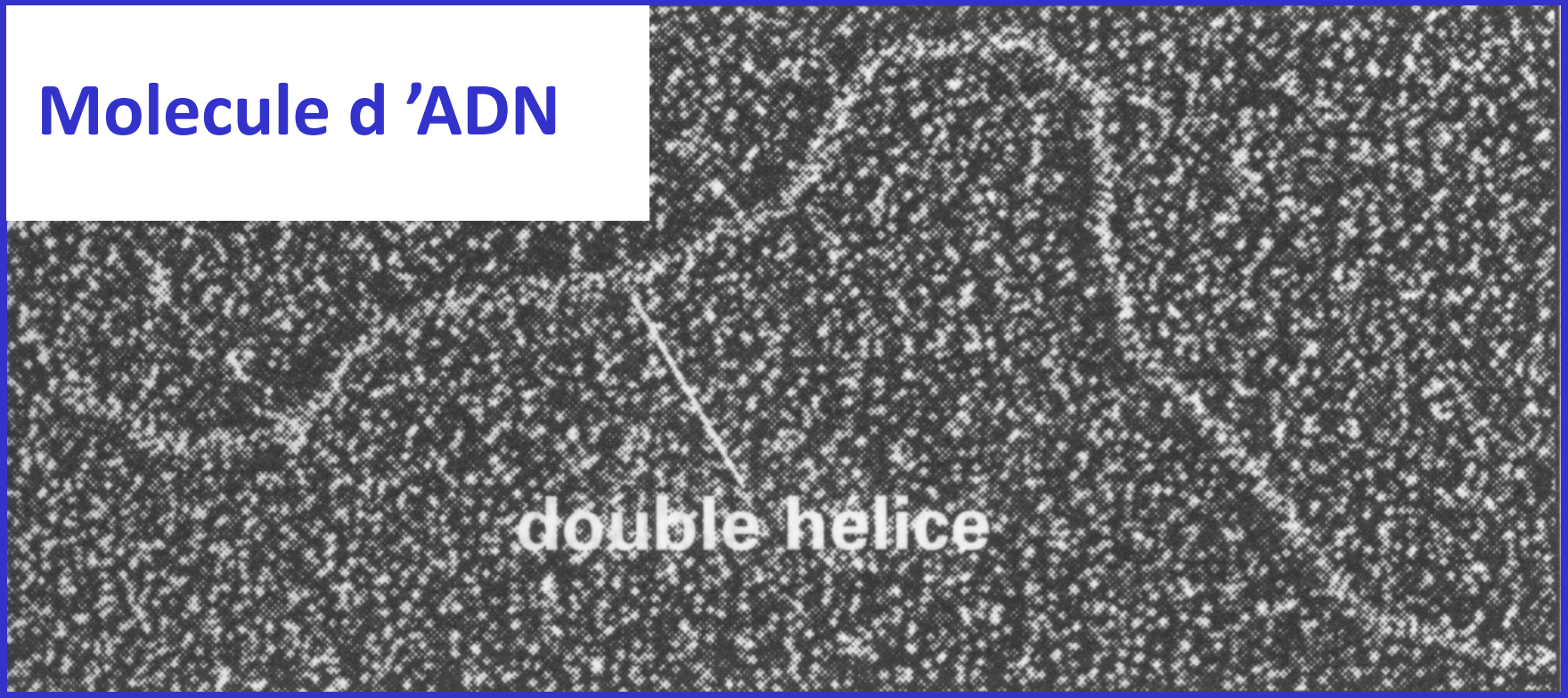
Cette présentation powerpoint peut être obtenue par simple demande sur francis.quetier @gmail.com

La pré-histoire: Rappels sur l'ADN

Francis CRICK & James WATSON



Molecule d ’ADN



Observation à très fort grossissement en microscopie électronique

La structure de l ’ADN

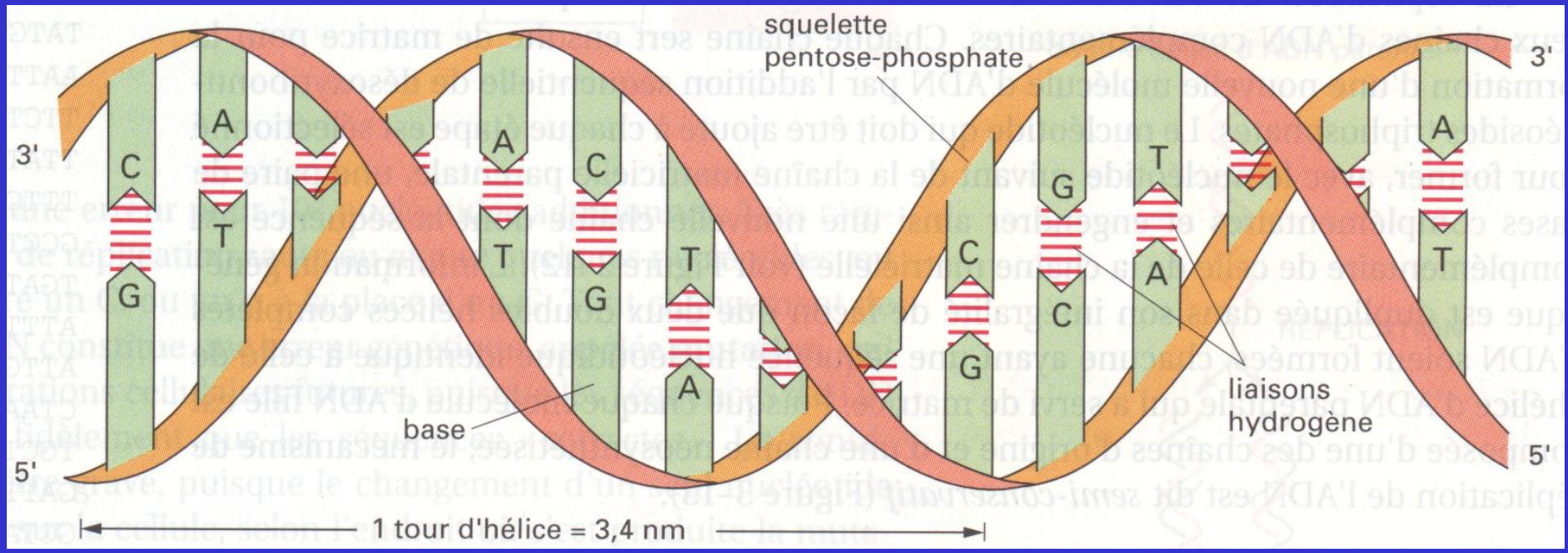
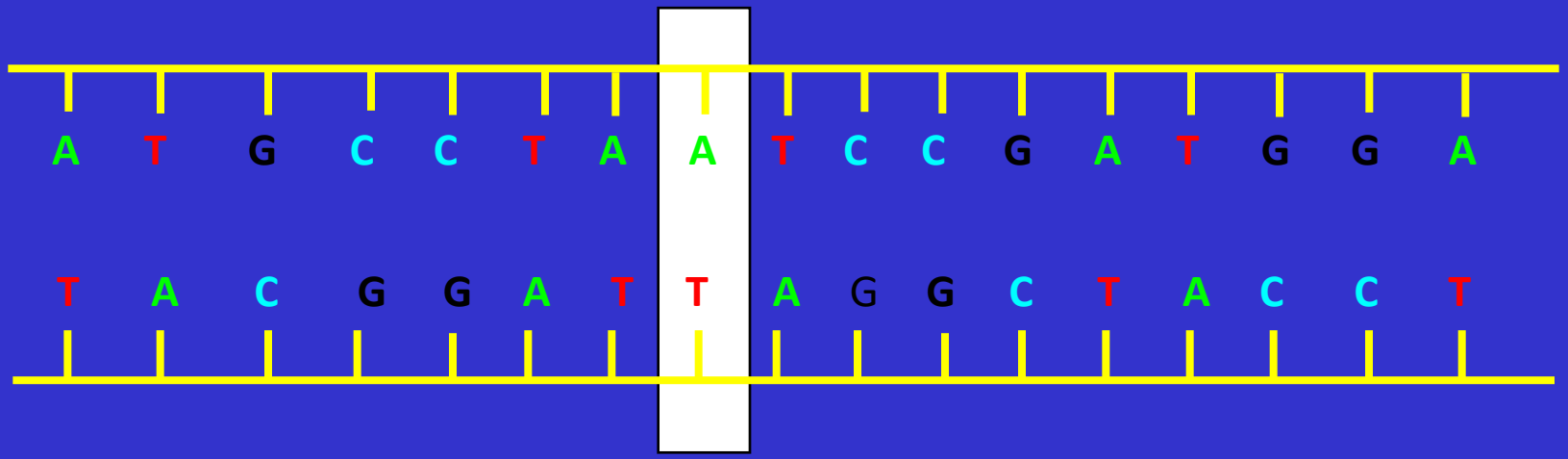


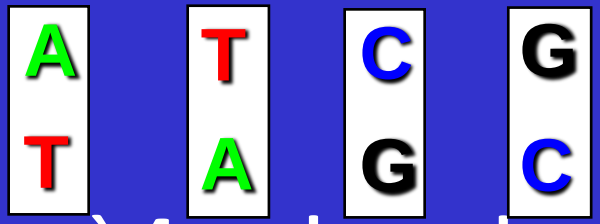
Schéma simplifié de la double-hélice

L'INFORMATION GENETIQUE : l' ADN



L'UNITE = une paire de nucléotides = paire de bases = paire de perles

4 possibilités



L'ADN = 2 collier de perles à 4 couleurs de perles.

Il y a 3,2 Milliards de perles par collier chez l'Humain

L'ordre des perles le long du collier= séquence de l'ADN

L'information génétique réside dans la séquence des perles

3,2 Miliards de lettres correspondent à environ 3.500 volumes de 500 pages chacun

Prix Nobel 1980

Fred SANGER en 1977 met au point la méthode de séquençage de l'ADN, qui porte son nom, et qui permet de déterminer l'ordre des bases ATGC le long des chaînes d'ADN.

A cette époque, le séquençage se fait par électrophorèse et la longueur unitaire des « lecture s » fait 100 bases, les appareils rudimentaires permettant de déterminer en parallèle la séquence de 4 fragments d'ADN de 100 bases chacun.

Le premier génome séquencé est celui d'un virus de bactérie, qui fait 5386 bases.

Dès ce premier séquençage, le challenge du séquençage du génome humain était engagé; ce n'était plus qu'une question de temps, d'argent, de progrès technologies et de bras!

Ce challenge apparaissait pharaonique: si l' ADN était représenté sous forme d'un ruban de magnétophone de 1 cm de large, la longeur correspondante serait d'environ 6.000 km , soit la distance de Paris à New-York, qu'il faudrait parcourir en déterminant la nature de la base, A, T, G ou C présente tous les 2 mm !!!!!

Les étapes:

1°) découper l'ADN humain en fragments de 150.000 pb en taille moyenne et séquencer chacun d'eux complètement.

2°) Assembler ces séquences pour reconstituer la séquence d'ADN de chacun des 24 types de chromosomes

3°) interpréter l'information de séquence ATGC en termes biologiques=identifier les fonctions exercées dans les cellules.

Des laboratoires sur les differents continents ont commencé à séquencer des petits bouts d'ADN humain, chacun de leur côté , sans aucune coordination, fragments qu'il était impossible de localiser les uns par rapport aux autres.

Il est vite apparu que pour mener à bien cet ordonnancement, il fallait disposer d'une cartographie du génome humain suffisamment détaillée pour pouvoir assigner chaque fragment d'ADN séquencé sur un endroit très précis de cette carte.

Dans les années 1980, les chercheurs disposaient d'une seule cartographie: le caryotype

Les 46 chromosomes humains

Caryotype humain

22 chromosomes présents chacun en 2 exemplaires

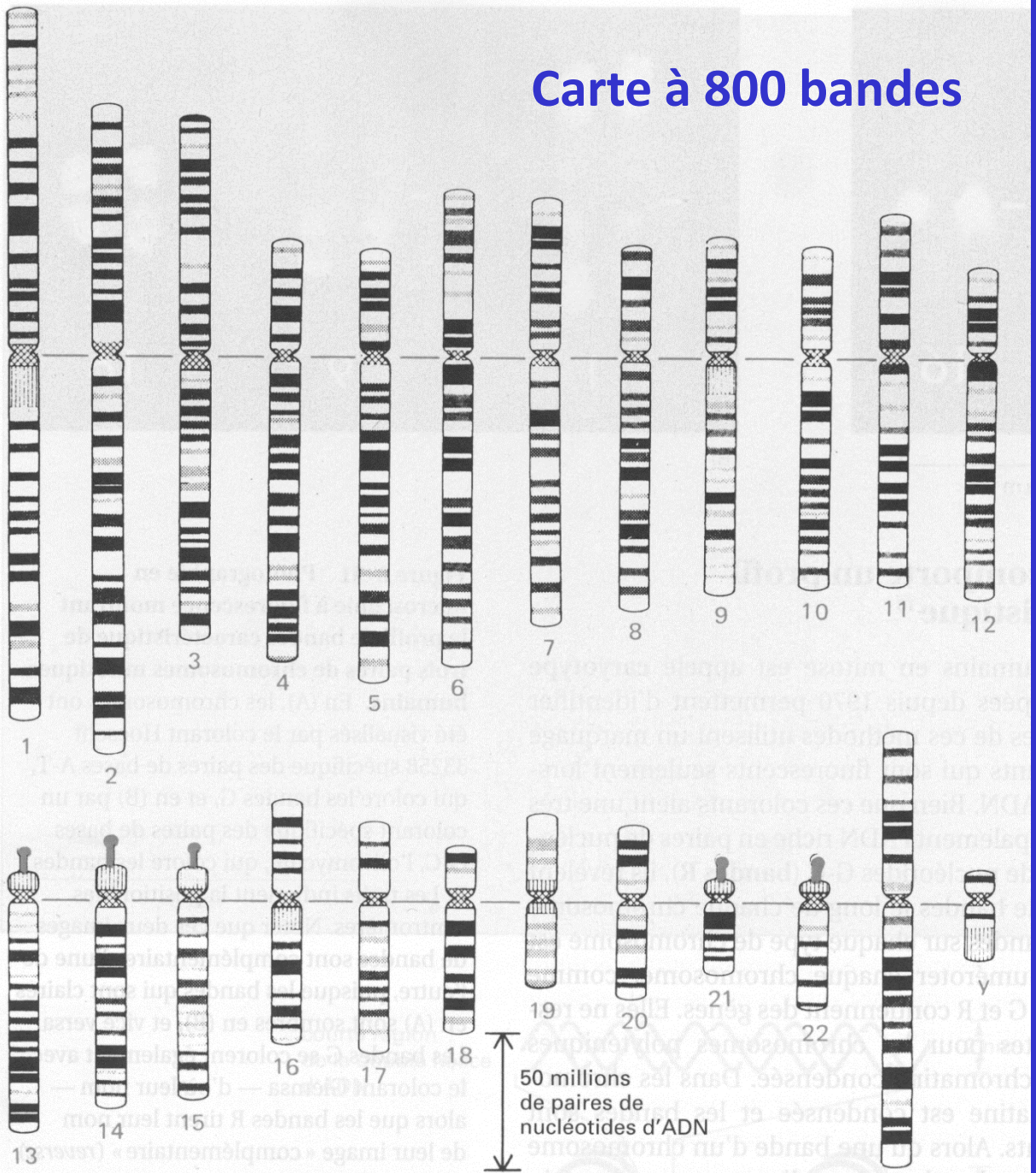
+

soit 2 X (femme)

soit 1 X et 1 Y (homme)

1 carte avec 800 bandes

1 carte avec 1.400 bandes



La résolution de cette carte physique était bien trop faible pour pouvoir mener à bien l'ordonnancement des tous les petits fragments d'ADN.

En 1986, Anthony Monaco réussit une première, l'identification du gène responsable de la myopathie de Duchenne. Cette découverte majeure pousse l'AFM à faire de la recherche génétique l'un des fers de lance de son action..

En 1986, le « Human Genome Project » est conçu par le NIH aux USA et piloté par James.D. Watson les premières années; il démarrera effectivement en 1989 et sera dirigé par Francis Collins.

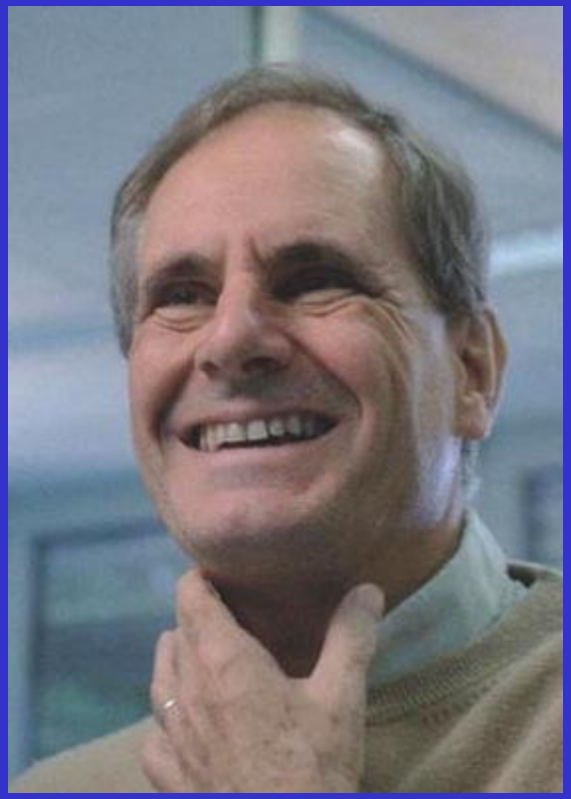
En 1986, Bernard Barataud et Pierre Birambeau, à l'AFM , décident une mission aux USA pour faire le point sur l'état de la recherche sur les maladies. Pïerre BIRAMBEAU entame un périple et rencontre l'acteur Jerry Lewis, qui est père d'un enfant myopathe. L 'acteur a lancé depuis 1966 une soirée annuelle avec une chaîne de télévison pour recueillir des dons en faveur de la recherche médicale.

Bernard Barataud et Pierre Birambeau convainquent alors Antenne 2 de lancer le Téléthon en France en 1987 et le succès est immédiat et durable.

En 1990, Bernard Barataud crée le Laboratoire Généthon à Evry. Il recrute Daniel Cohen et Jean Weissenbach et pour y construire respectivement, pour le génome humain,

- une carte physique très détaillée
- une carte génétique très détaillée

Daniel
COHEN



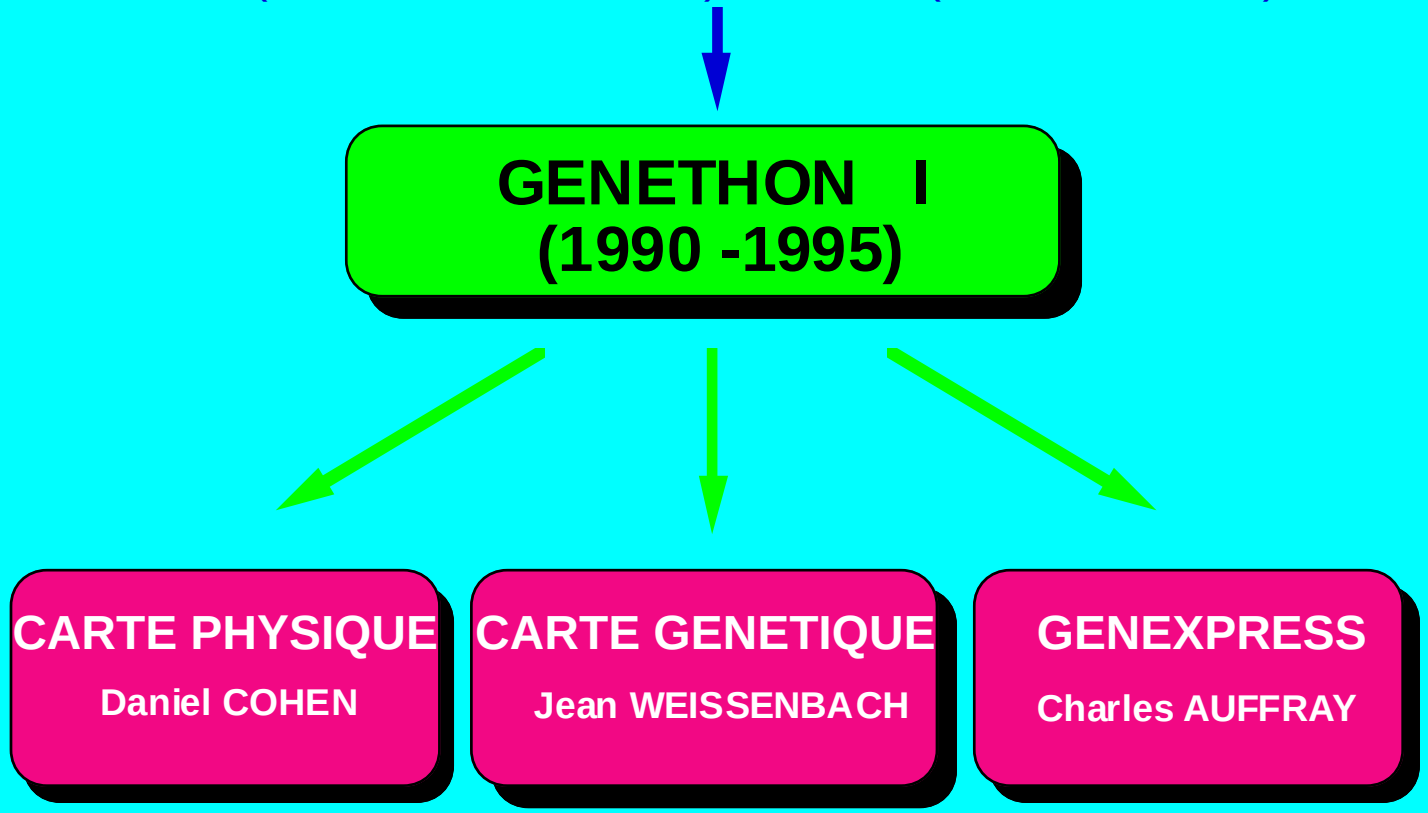
Jean
WEISSENBACH

Charles
AUFFRAY

CEPH - GENETHON

Collection of > 180 families by Jean DAUSSET

A.F.M. (Bernard BARATAUD) & CEPH (Daniel COHEN)



Les différents types de banques de fragments d'ADN

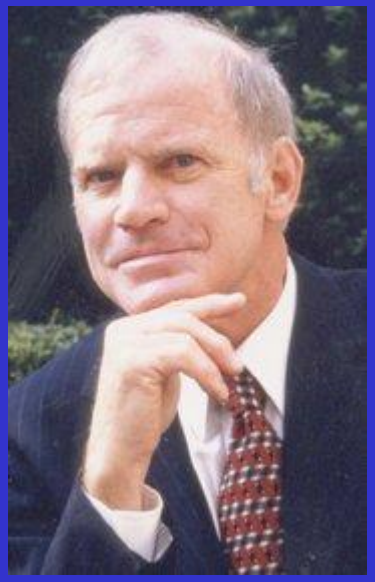
Vecteur	insert	Nombre de clones nécessaires	
		1 X	15 X
YAC	500 kb	6.000	90.000
BAC	150 kb	20.000	300.000

300.000

Clones en microplaques de 384 puits = 782 micro-plaques ,soit 1/2 congélateur -80°C.

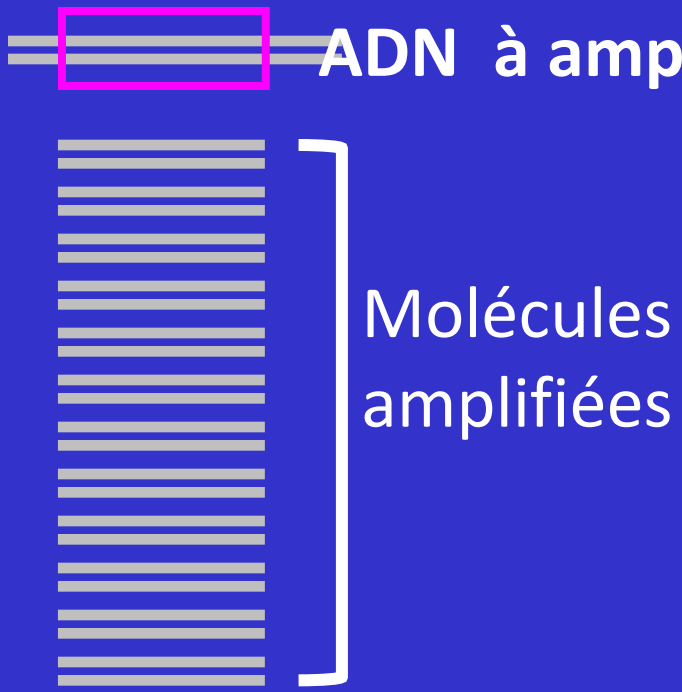
Microplaque à 96 puits
13 cm x 8 cm

Robot de distribution automatique de réactifs liquides dans les microplaques



L’ Amplification en chaîne par polymérase = PCR est une méthode d'amplification d’ADN *in vitro*, qui permet de dupliquer en grand nombre (avec un facteur de multiplication de l'ordre du milliard) une séquence d’ADN connue, à partir d'une faible quantité (de l'ordre de quelques picogrammes) d’ADN.

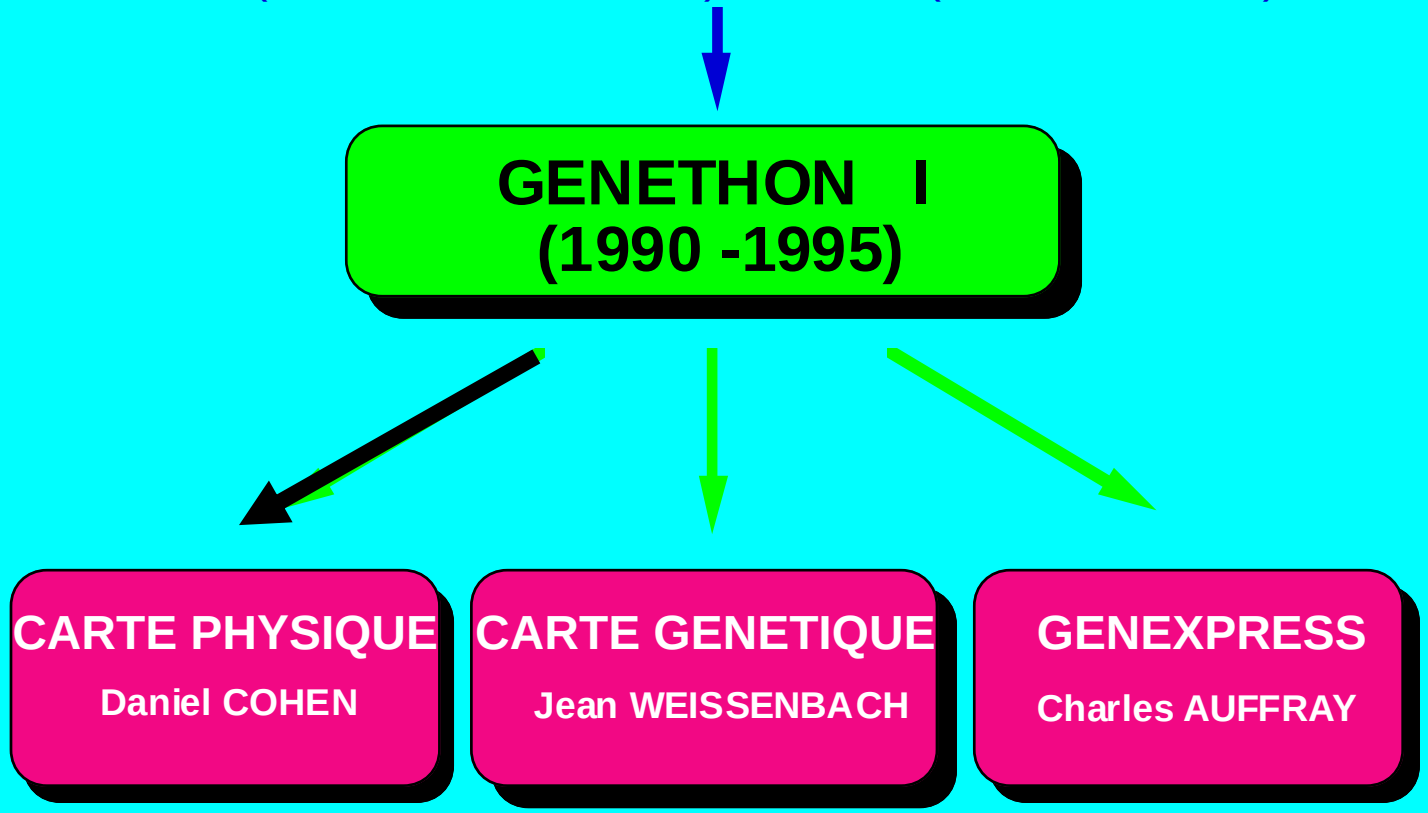
Kary MULLIS
Prix Nobel 1993



CEPH - GENETHON

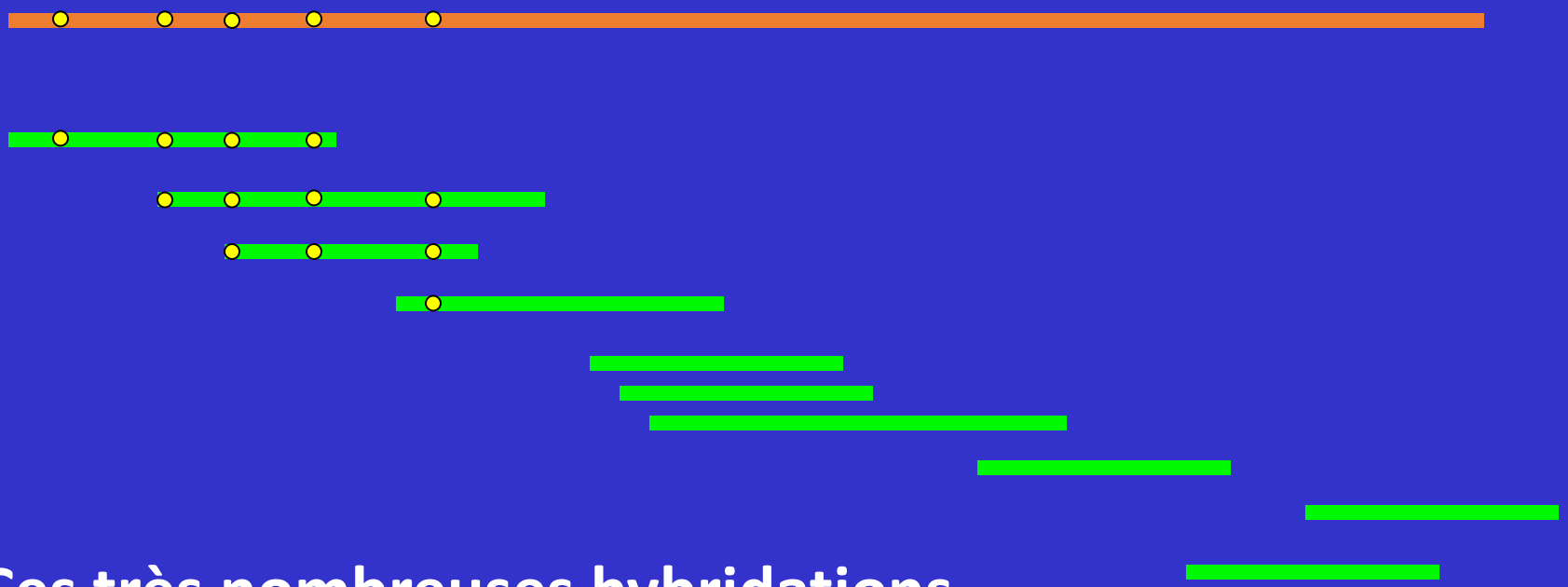
Collection of > 180 families by Jean DAUSSET

A.F.M. (Bernard BARATAUD) & CEPH (Daniel COHEN)



CARTOGRAPHIE PHYSIQUE

Il faut reconstituer l'ordre des inserts dans le génome. Pour la cartographie YACs, cela a été réalisé par hybridations moléculaires ADN-ADN



Ces très nombreuses hybridations moléculaires ont été réalisées sur des robots spécialement conçus à Généthon..

La cartographie physique a couvert 70% du génome humain et publiée en 1995.

Mais la moitié des YACs se sont révélés inutilisables car les fragments d'ADN qu'ils véhiculaient avaient subi des remaniements de séquences et leur séquence n'était plus co-linéaire avec l'ADN originel des prélèvements.

Mais beaucoup de portions de cartes ont été très utiles Pour commencer localiser des gènes impliqués dans des maladies.

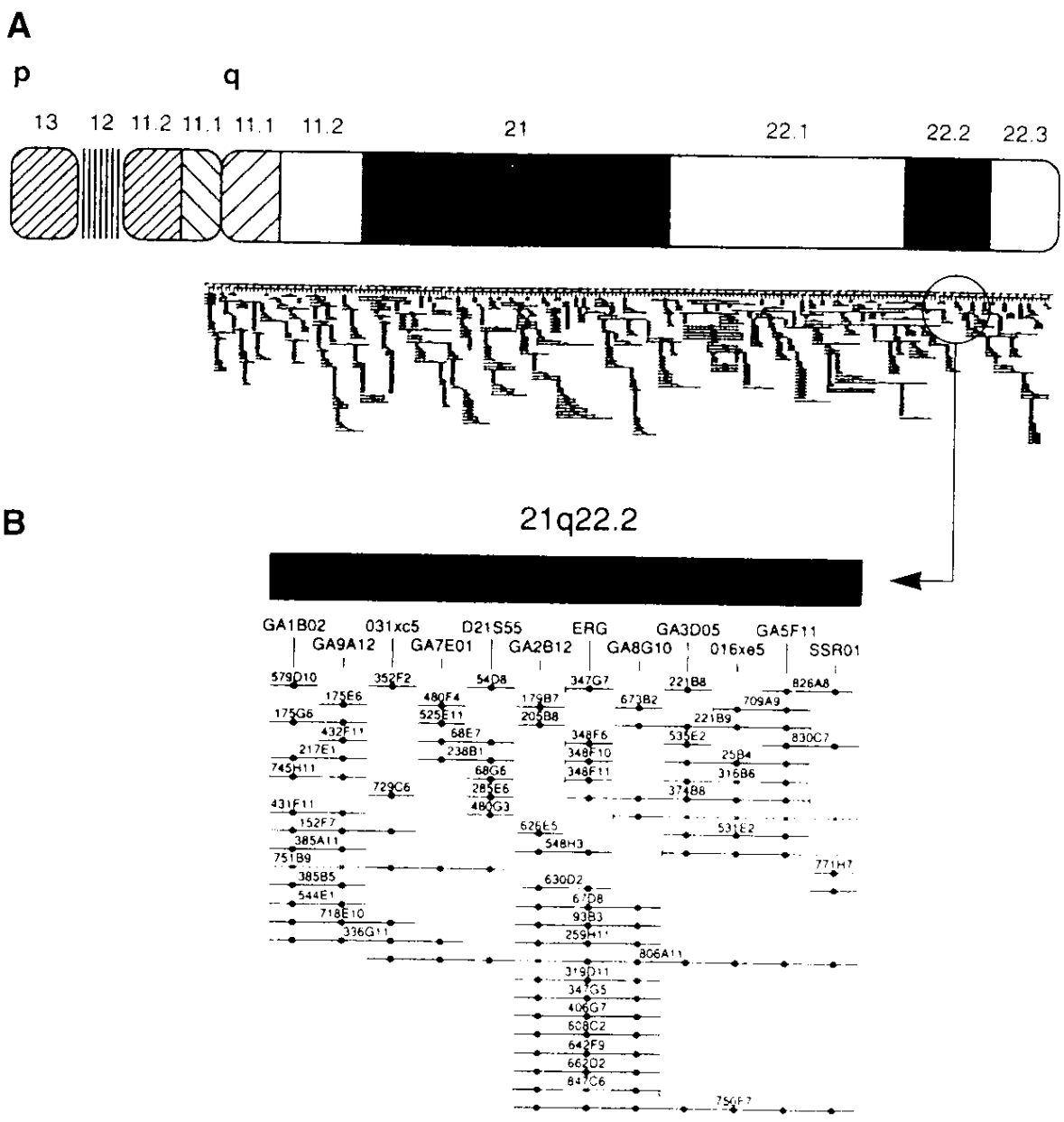
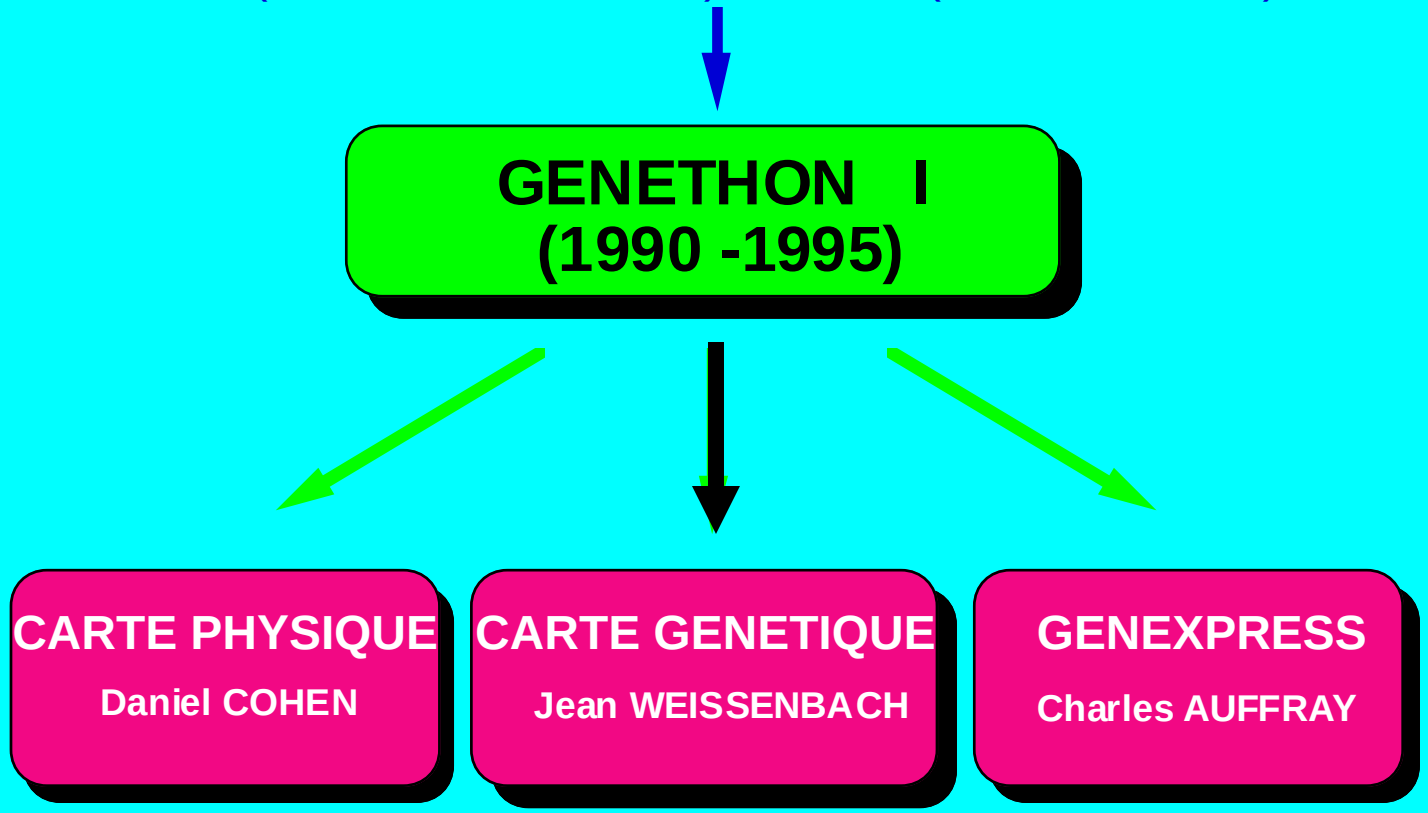


Figure 25. *Carte physique du chromosome 21.* **A** *: sous le chromosome est représenté l'ensemble des YAC cartographiés.* **B** *: détail montrant comment les YAC sont ordonnés en fonction du contenu en STS.*

CEPH - GENETHON

Collection of > 180 families by Jean DAUSSET

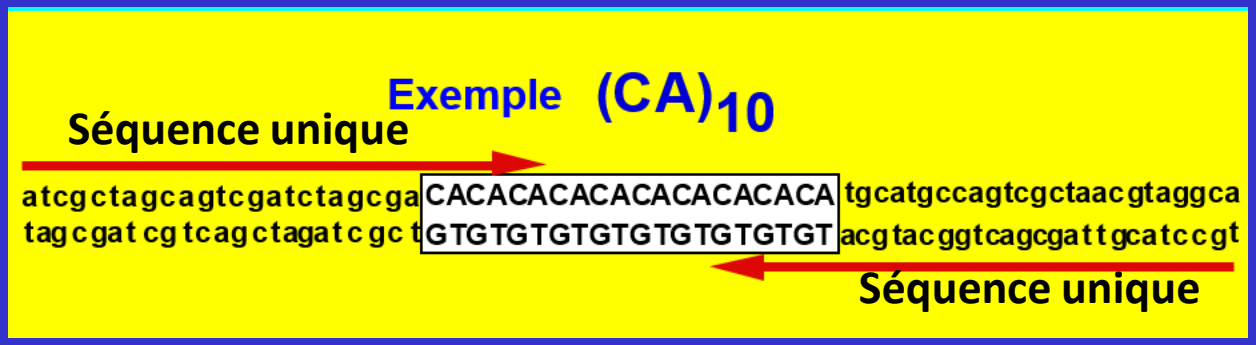
A.F.M. (Bernard BARATAUD) & CEPH (Daniel COHEN)



Deux individus non apparentés diffèrent en moyenne au niveau de leur génome, par 1 paire de bases sur 1.000= polymorphisme des génomes individuels.

Un polymorphisme particulièrement important porte sur des petit séquences appelées microsatellites, qui sont constituées de répétions de 2 paires de bases dont le nombre varie d'un individu à l'autre.

Les microtatellites CA (n) sont très nombreux (70.000 dans le génome humain), bien dispersés le long des chromosomes (1 tous ls 45 kb). Ils sont, comme tous le sautres microsatellites, amplifiables par PCR.



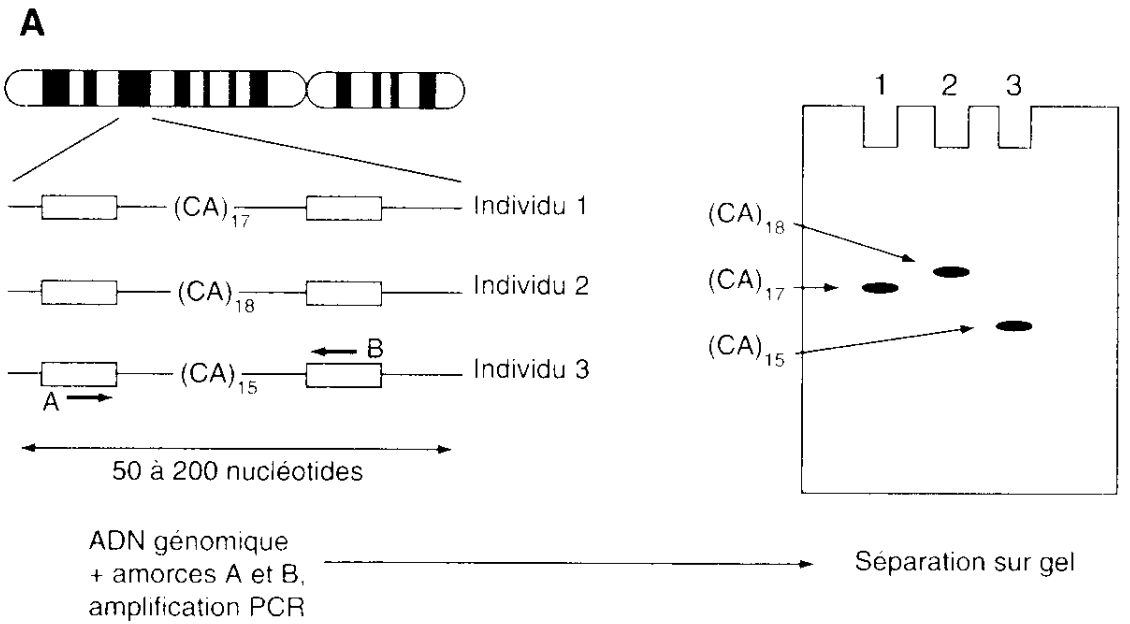
Ce sont les marqueurs moléculaires qui ont été utilisés par l'équipe de Jean WEISSENBACH à Généthon pour construire la première carte génétique humaine complète et à haute résolution.

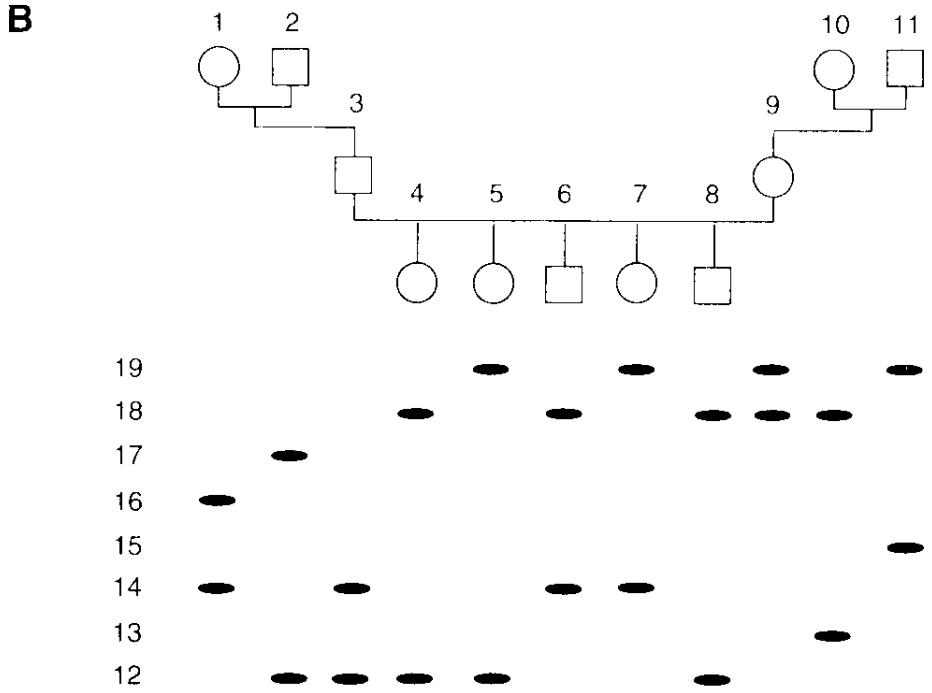
Le Pr. Jean DAUSSET

Prix Nobel 1970

Directeur du CEPH

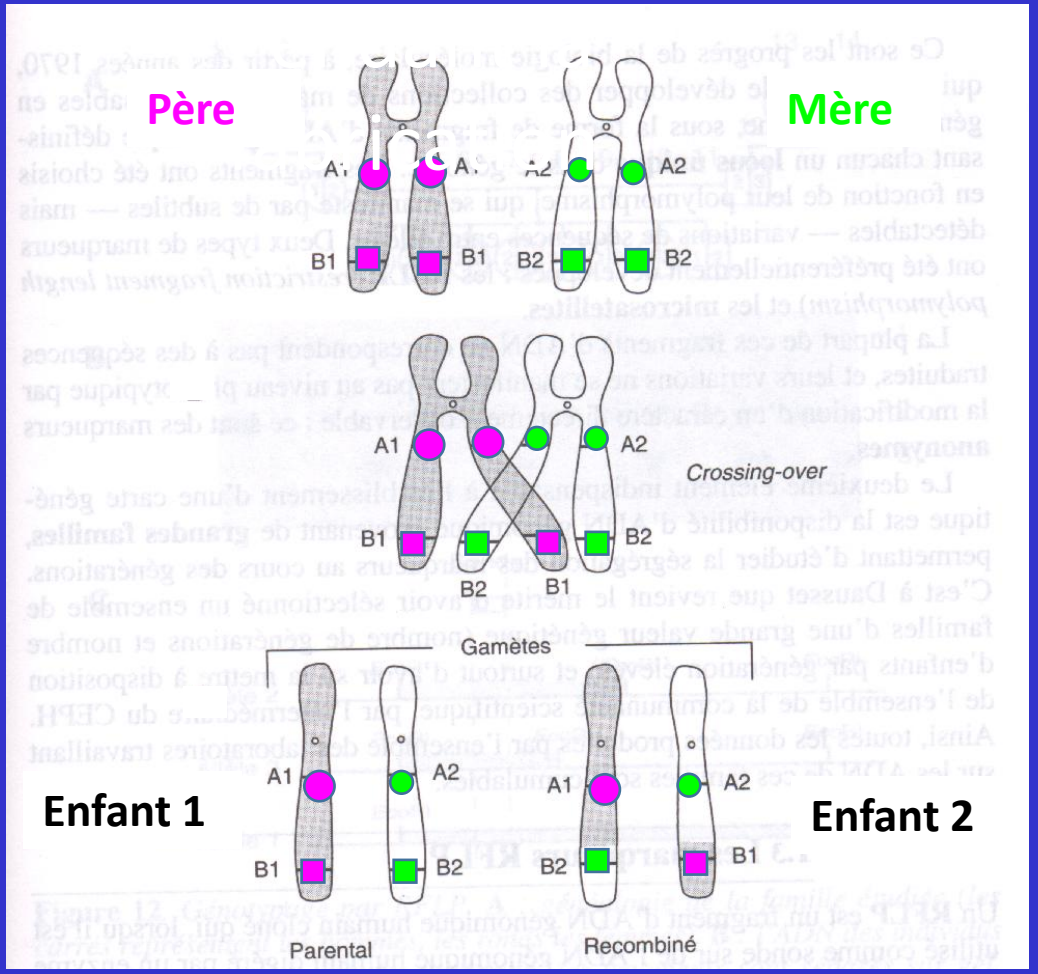
a collecté 180 frateries dont chaque membre avait fait l'objet d'un prélèvement de sang pour extraction d'ADN.





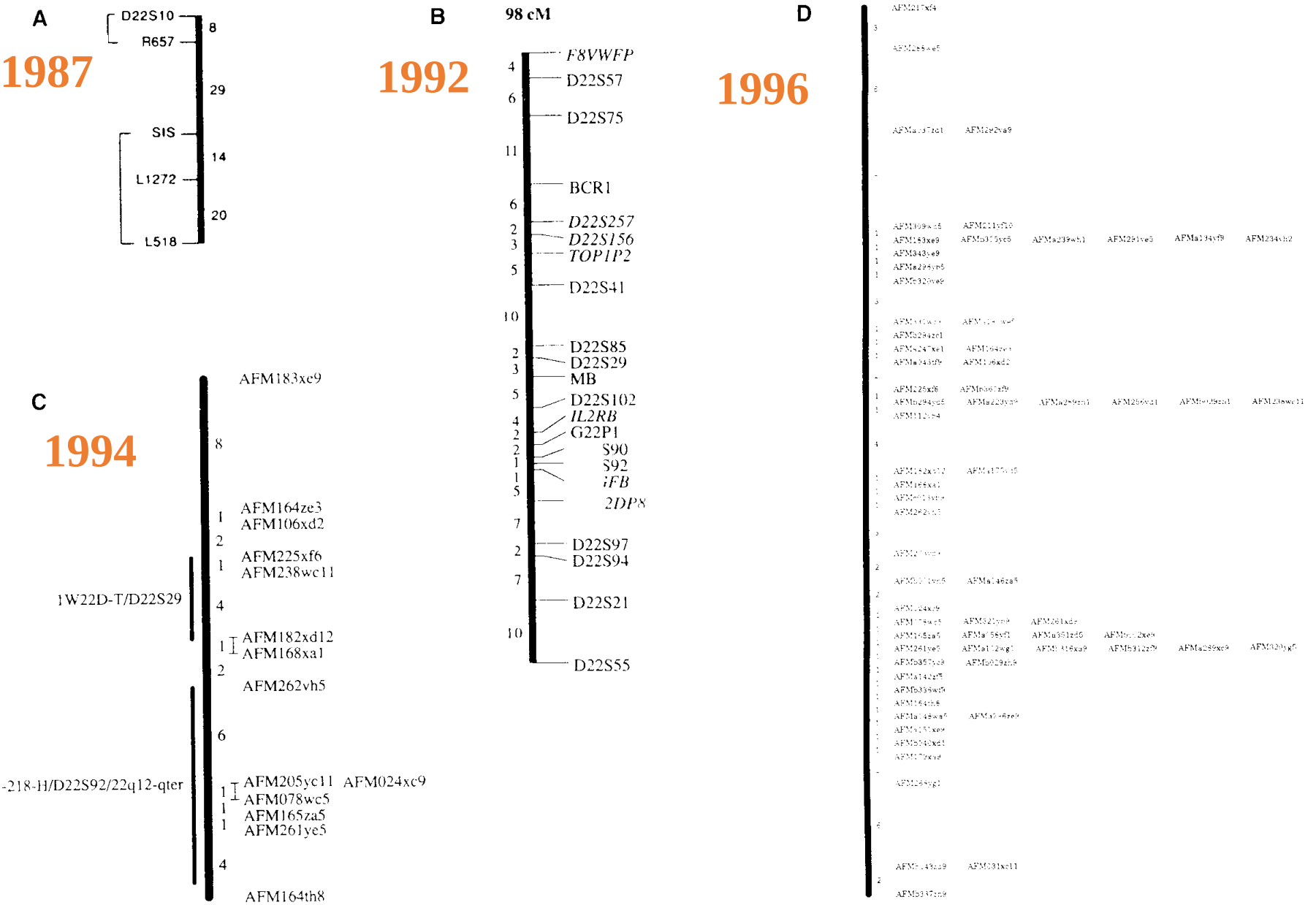
Pendant les divisions cellulaires pour la fabrication des spermatozoïdes et des ovocytes, quand les chromosmes paternels et maternels forment 23 paires, il y a en moyenne 2 ou 3 échanges de morceaux de chromosomes. Ces crossing-over peuvent internir n'importe où le long de chaque paire de chromsomes.

Plus la distance entre 2 marqueurs est faible, moins il y a de chances pour que 1 des 2 ou 3 crossing over intervienne entre les 2 marqueurs.



2 marqueurs qui sont proches seront rarements séparés l'un de l'autre et seront transmis en ensemble. La distance génétique est donc proportionnelle au % observé d'associations recombinées. Unité= CentiMorgan➔ 1%.

La progression des cartes génétiques : exemple sur le chromosome 22



La carte génétique obtenue par l'équipe de Jean WEISSENBACH à Généthon a été publiée en avril 1996 et comporte 5264 marqueurs génétiques ordonnés.

CARTE GENERALE DU GENOME HUMAIN

WHITEHEAD-MIT & GENETHON-EVRY

1996

TOTAL= 15.0686 loci

5264 Marqueurs Généthon
3.954 STS divers*
2.400 EST*
1.838 divers
1.595 Marqueurs CHLC
827 Genbank

Maiilage= 199kb

900 pags format A4 en impression

Disponibles en accès libre sur le site web

Une étude retrospective menée par Généthon en 1997 a montré que la carte génétique avait permis de localiser 900 gènes impliqués dans des maladies et de manière déjà assez precise (chacun dans une zone de 200 kb)

Durant cette période, une nouvelle stratégie de cartographie, avec un très haut débit, a été mise au point :

la cartographie par hybrides d'irradiation

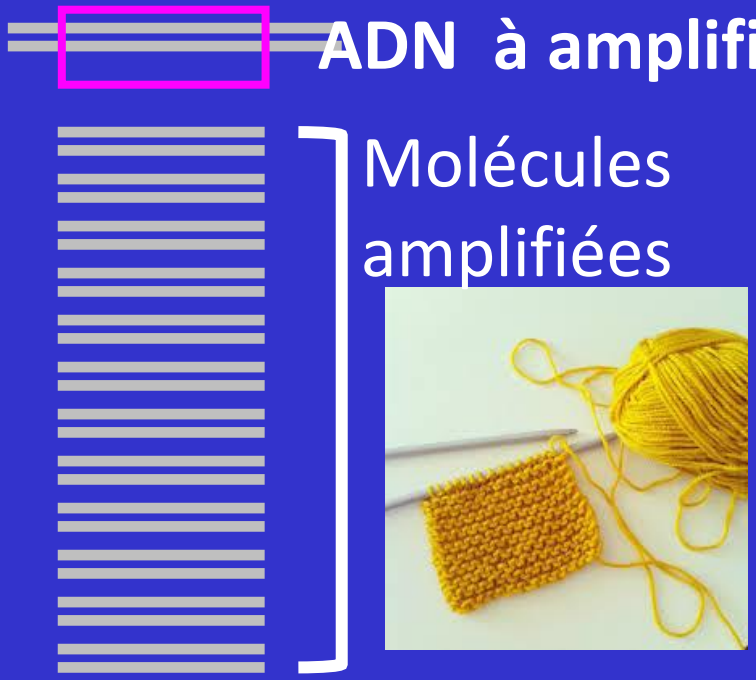
Le RH Mapping a été mené par un consortium interne du HGP= Genoscope/Sanger Institute/ Whitehead/Baylor College Medicine/ Wash-U.

En janvier 1999, le nombre de marqueurs moléculaires a dépassé les 30.000, ce qui a été jugé suffisant pour accélérer la phase de séquençage massif de l'ADN du génome humain.

La phase de séquençage de l'ADN

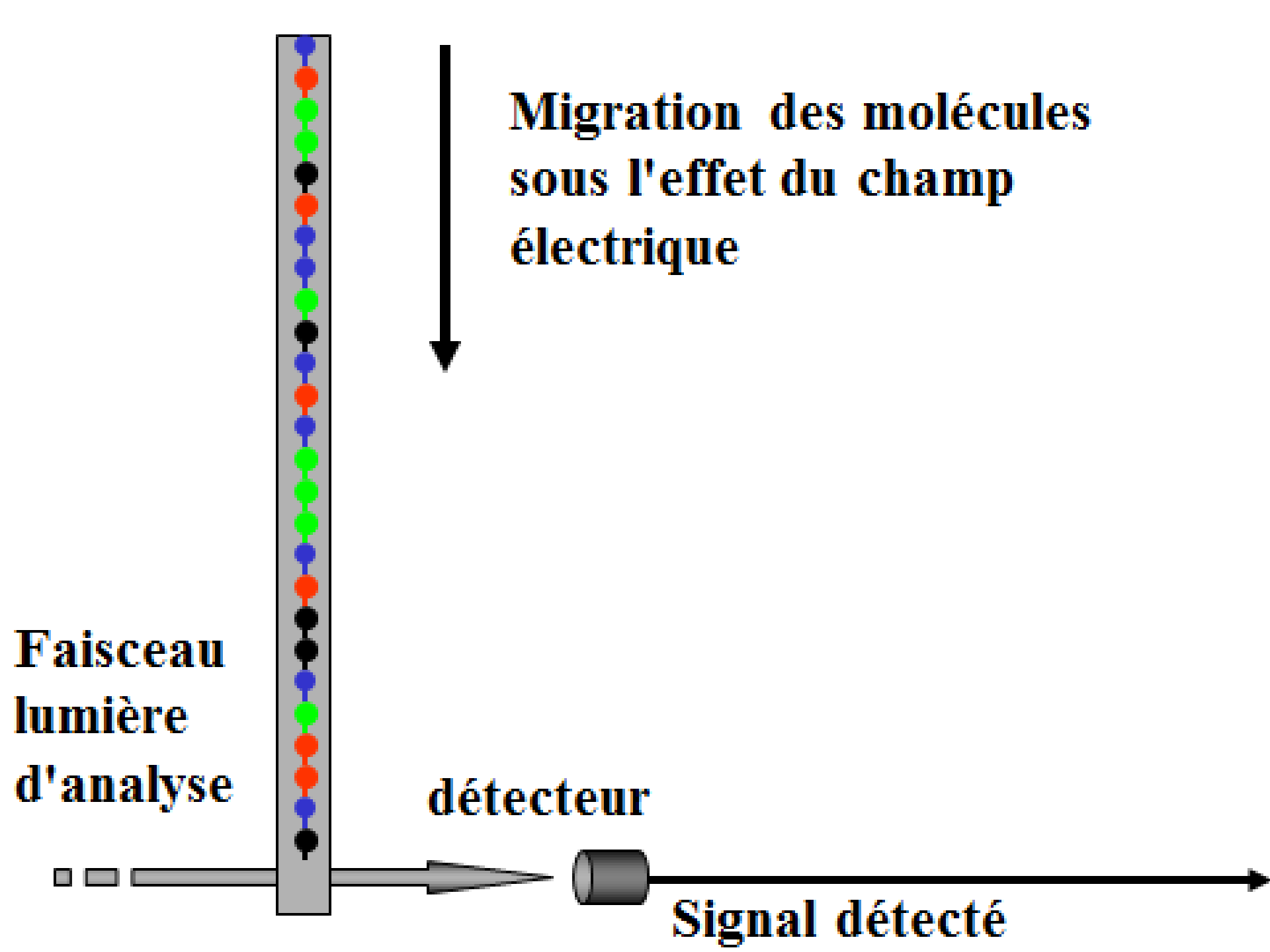
La technique de Séquençage de l'ADN à cette époque:

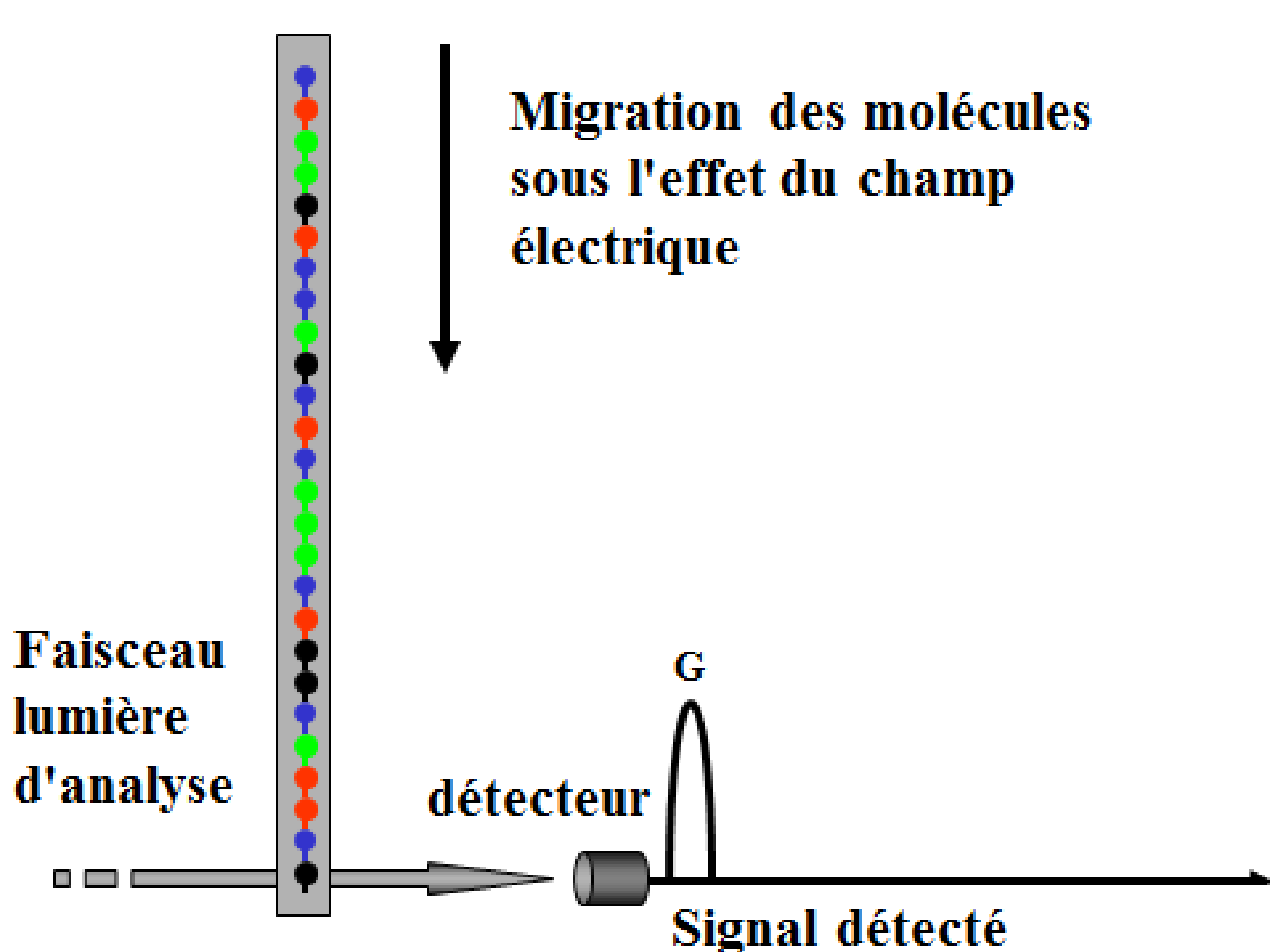
Chaque petit fragment d'ADN devait être copié un grand nombre de fois parle technologie PCR ,vue précédemment, pour atteindre une quantité nécessaire pour la détection.

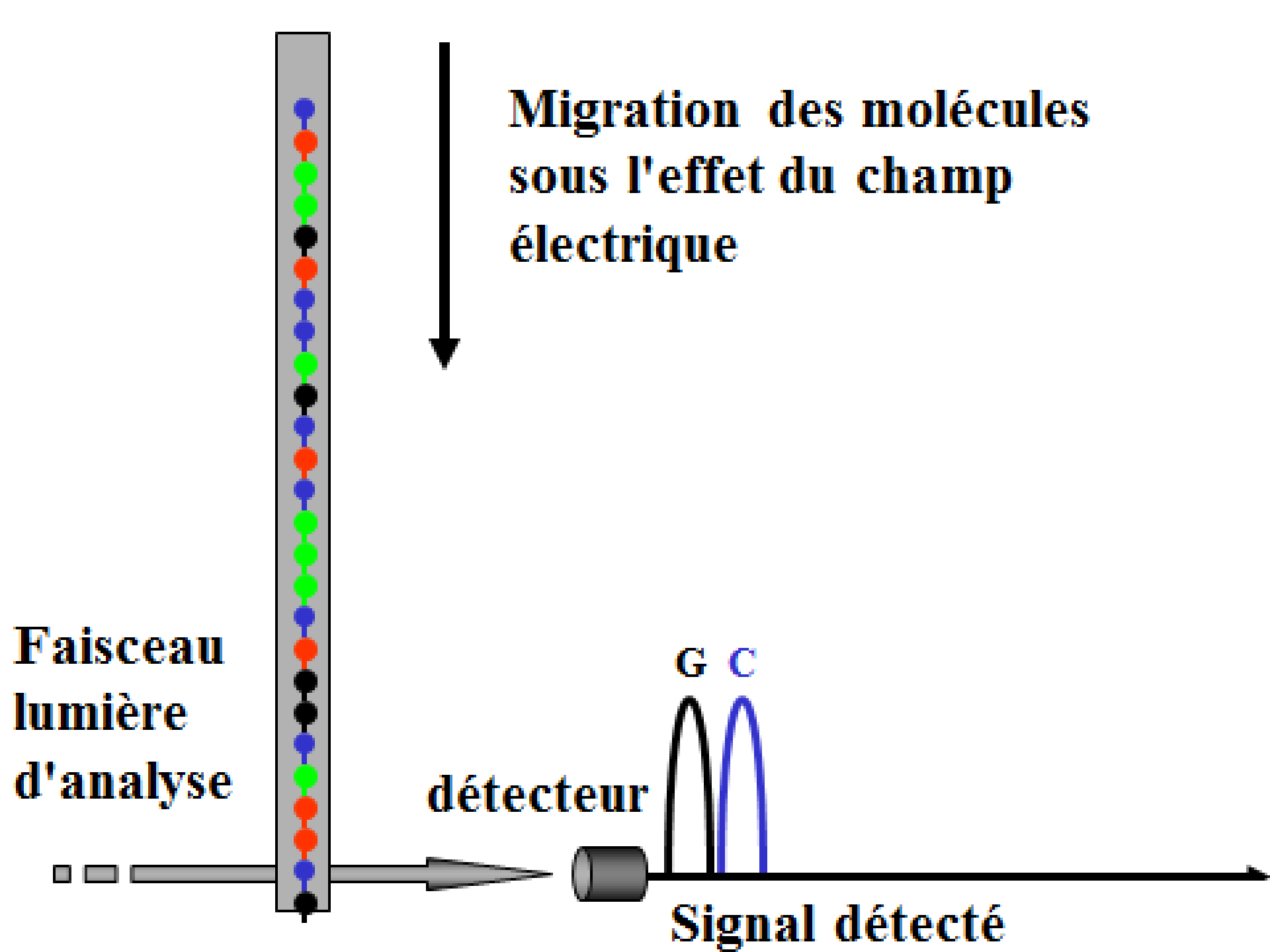


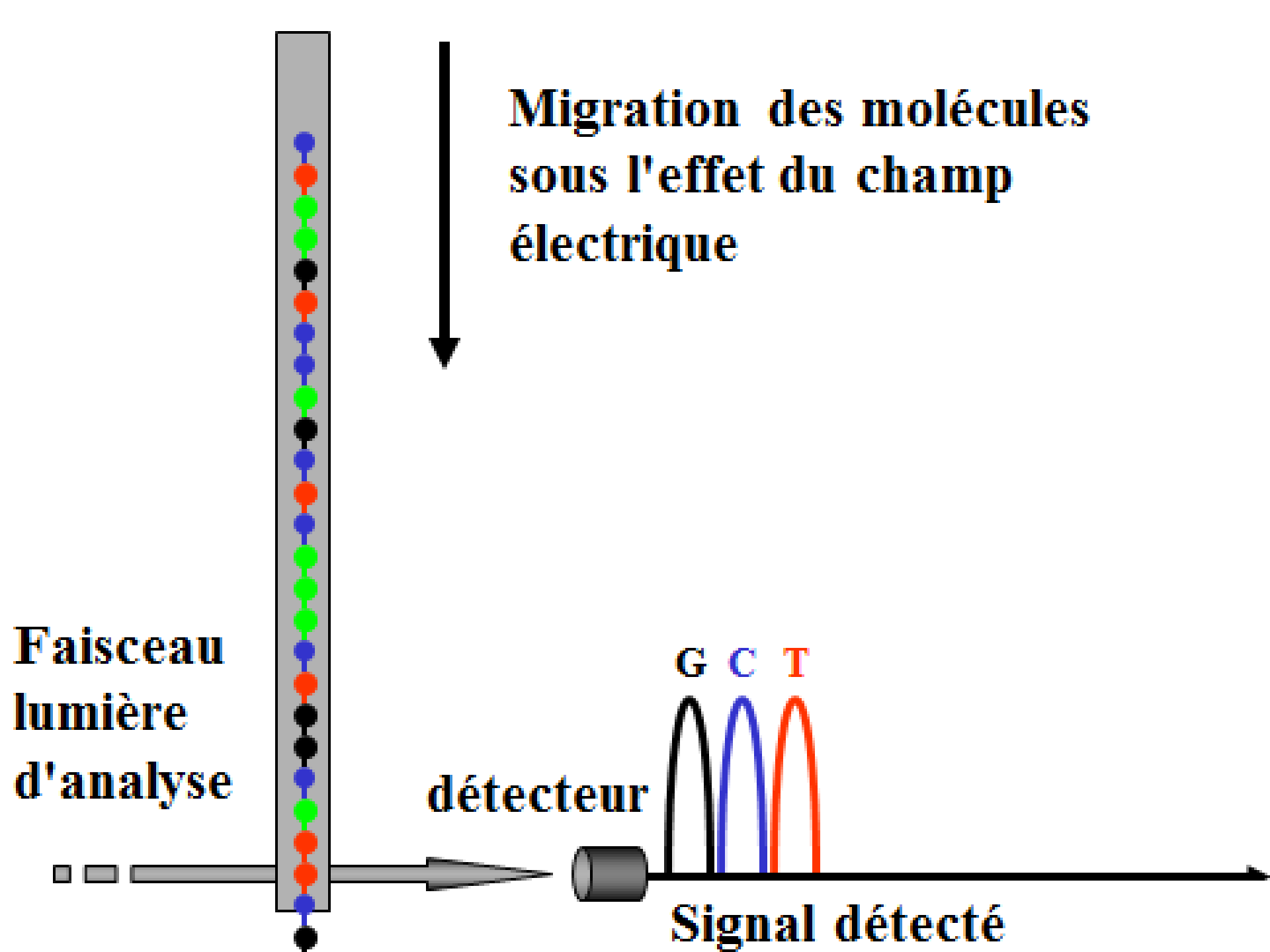
Les bases/nucléotides/perles ATGC utilisées pour fabriquer les copies de ces ADN étaient préalablement greffées avec molécules fluorescentes de couleurs spécifiques:

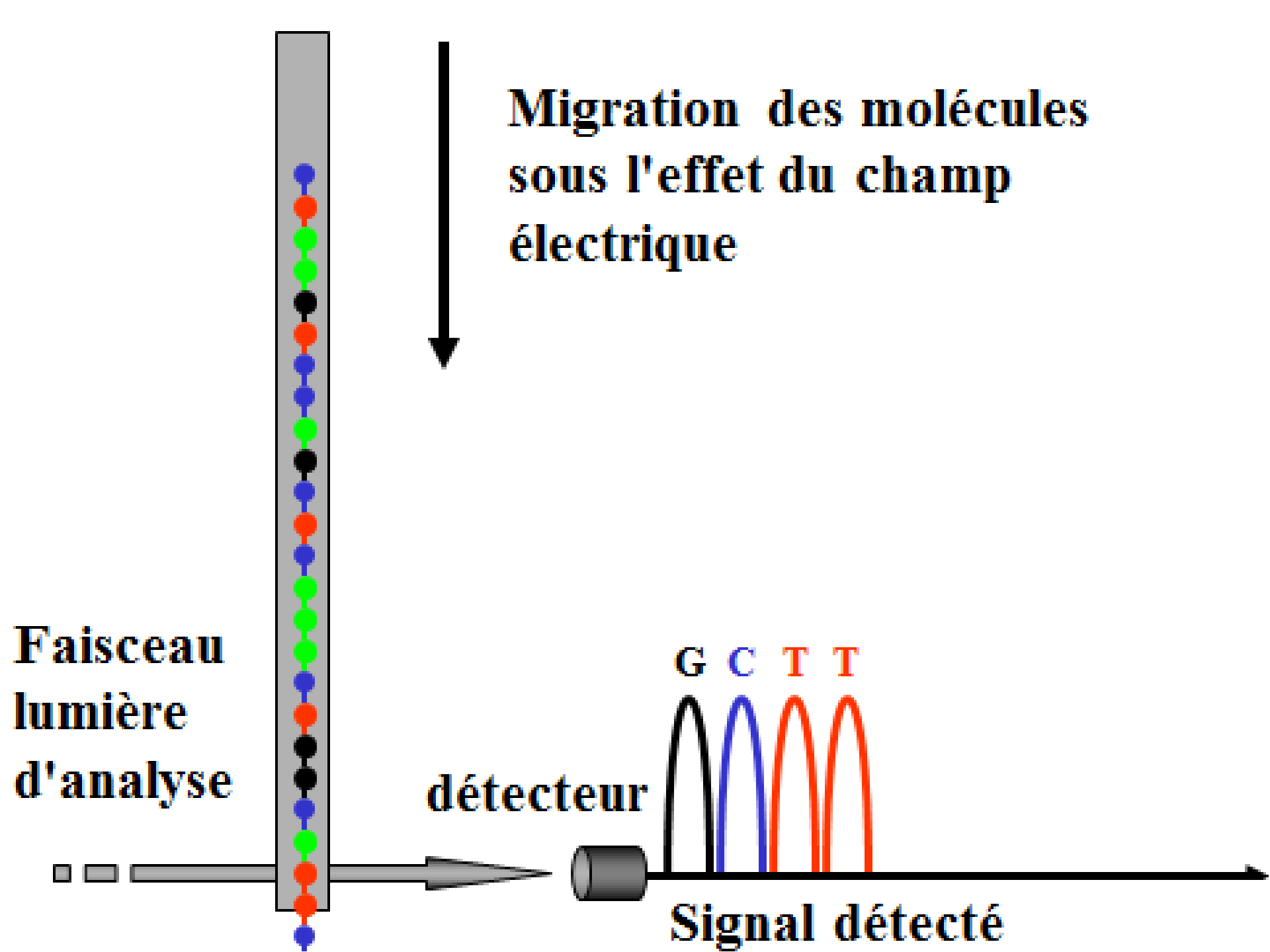
Les A en Vert
Les T en Rouge
Les C en Bleu
Les G en Noir.

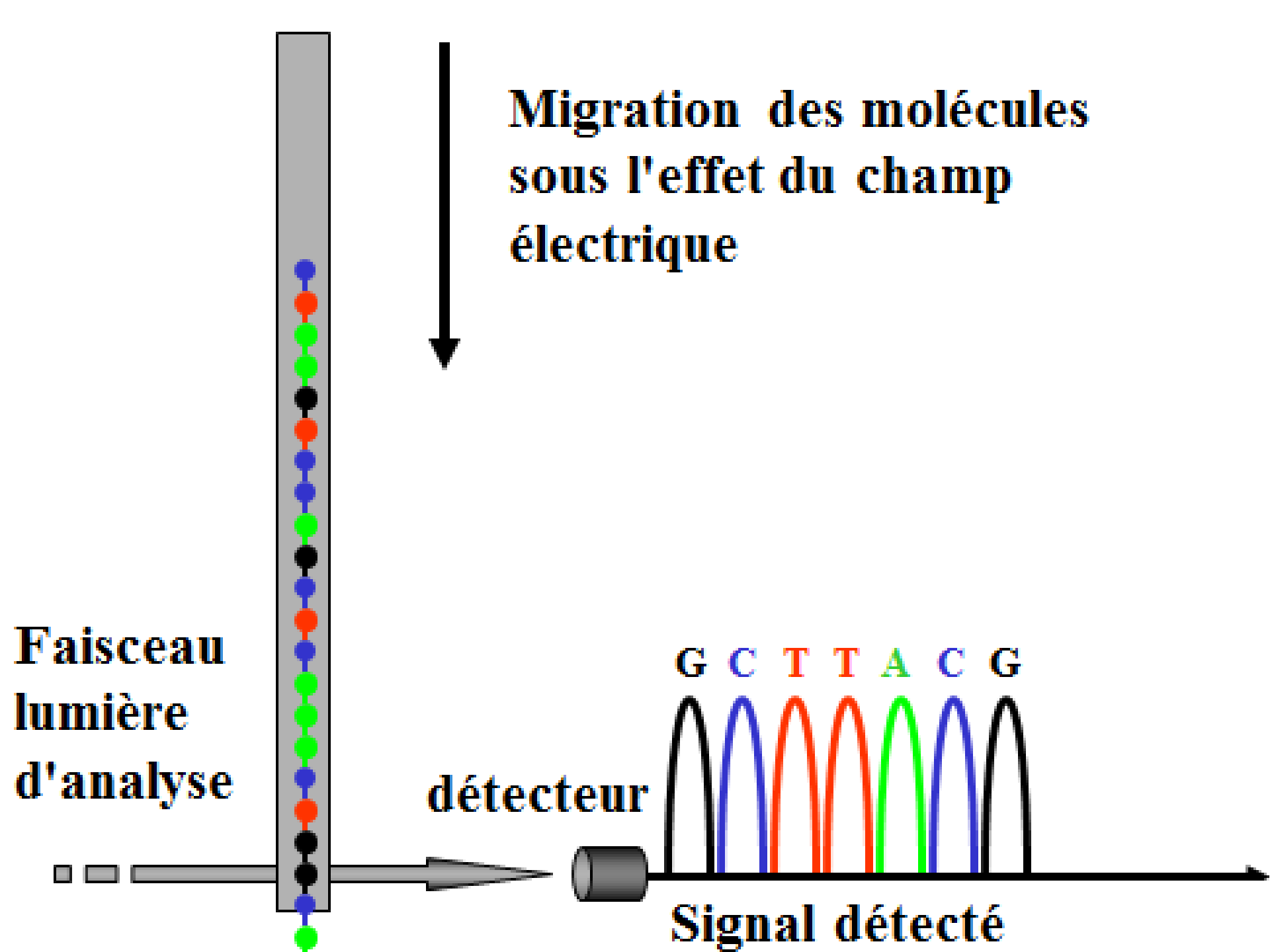

Migration des molécules sous l'effet du champ électrique
Faisceau lumière d'analyse
détecteur
Signal détecté

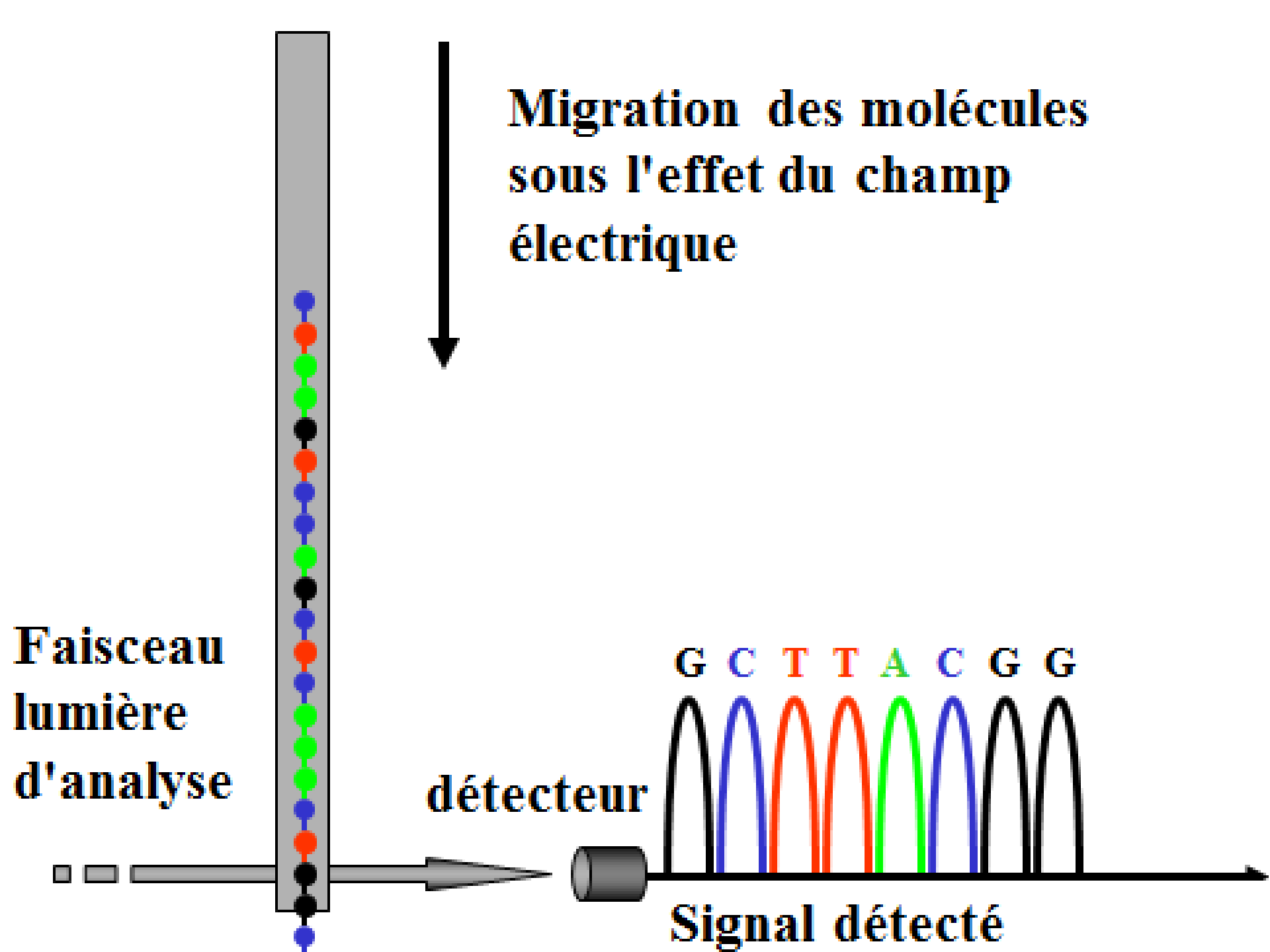

Migration des molécules sous l'effet du champ électrique
Faisceau lumière d'analyse
G
détecteur
Signal détecté

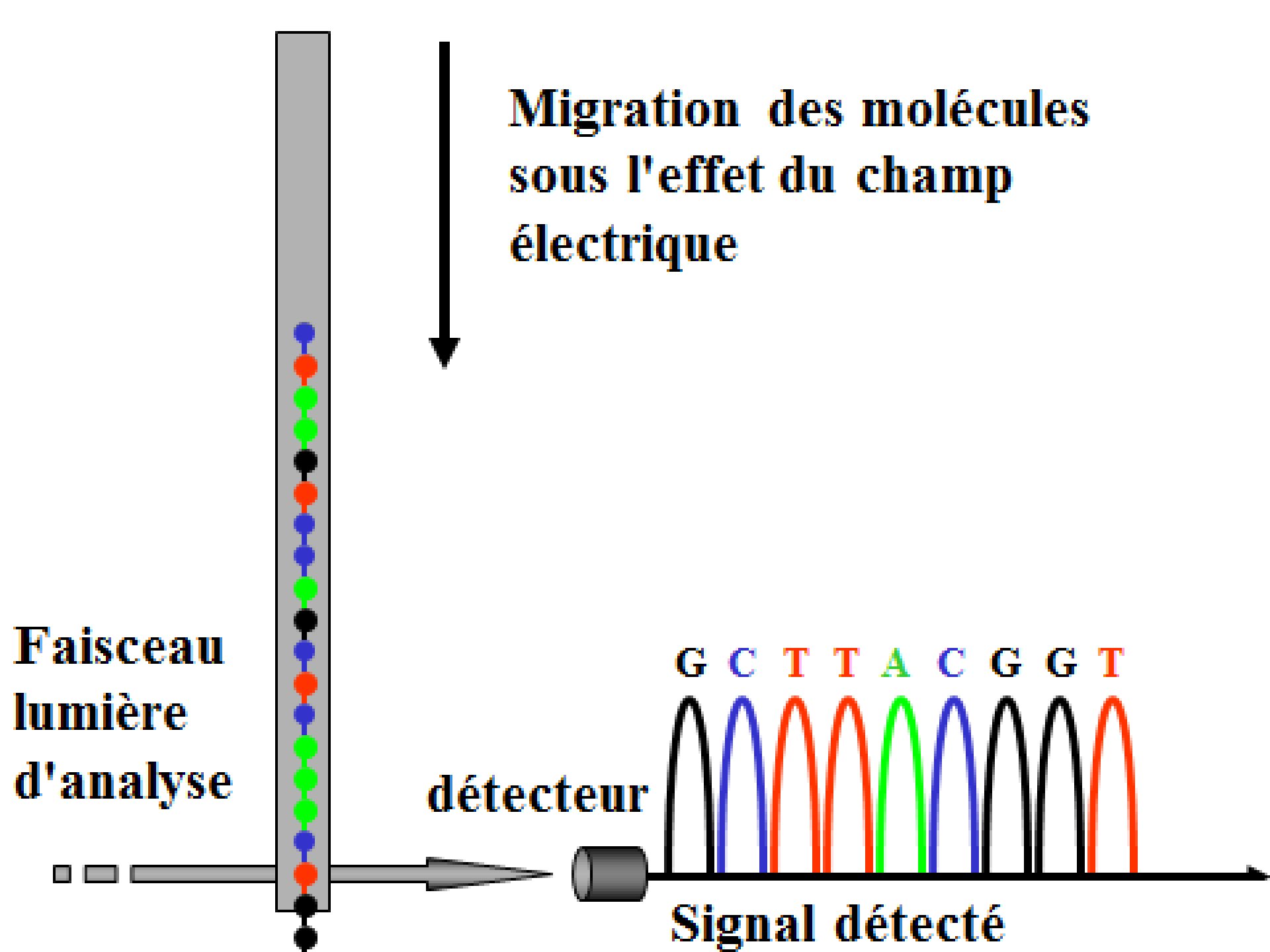

Migration des molécules sous l'effet du champ électrique
Faisceau lumière d'analyse
détecteur
G
C
Signal détecté

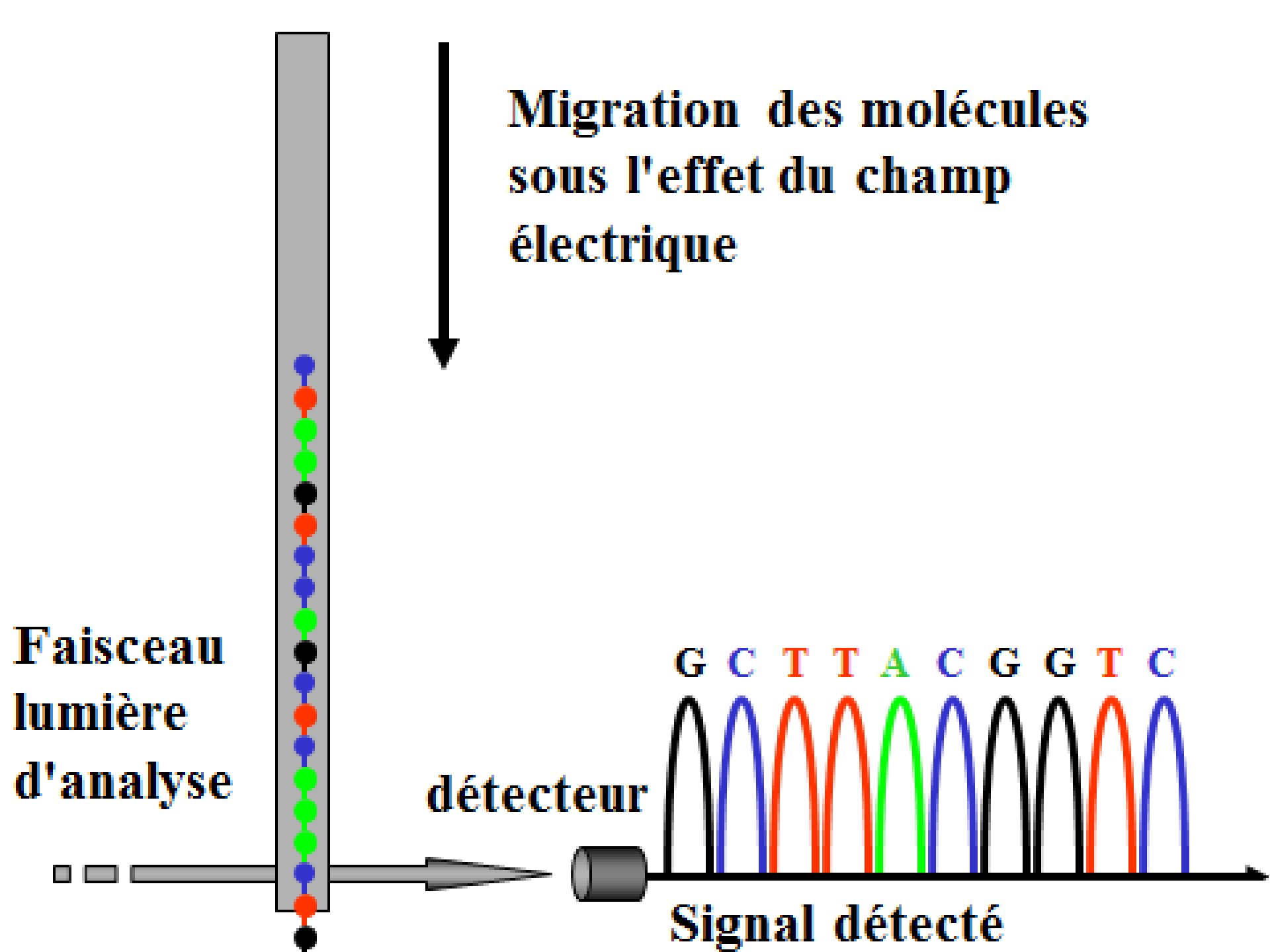

Migration des molécules sous l'effet du champ électrique
Faisceau lumière d'analyse
détecteur
G C T
Signal détecté

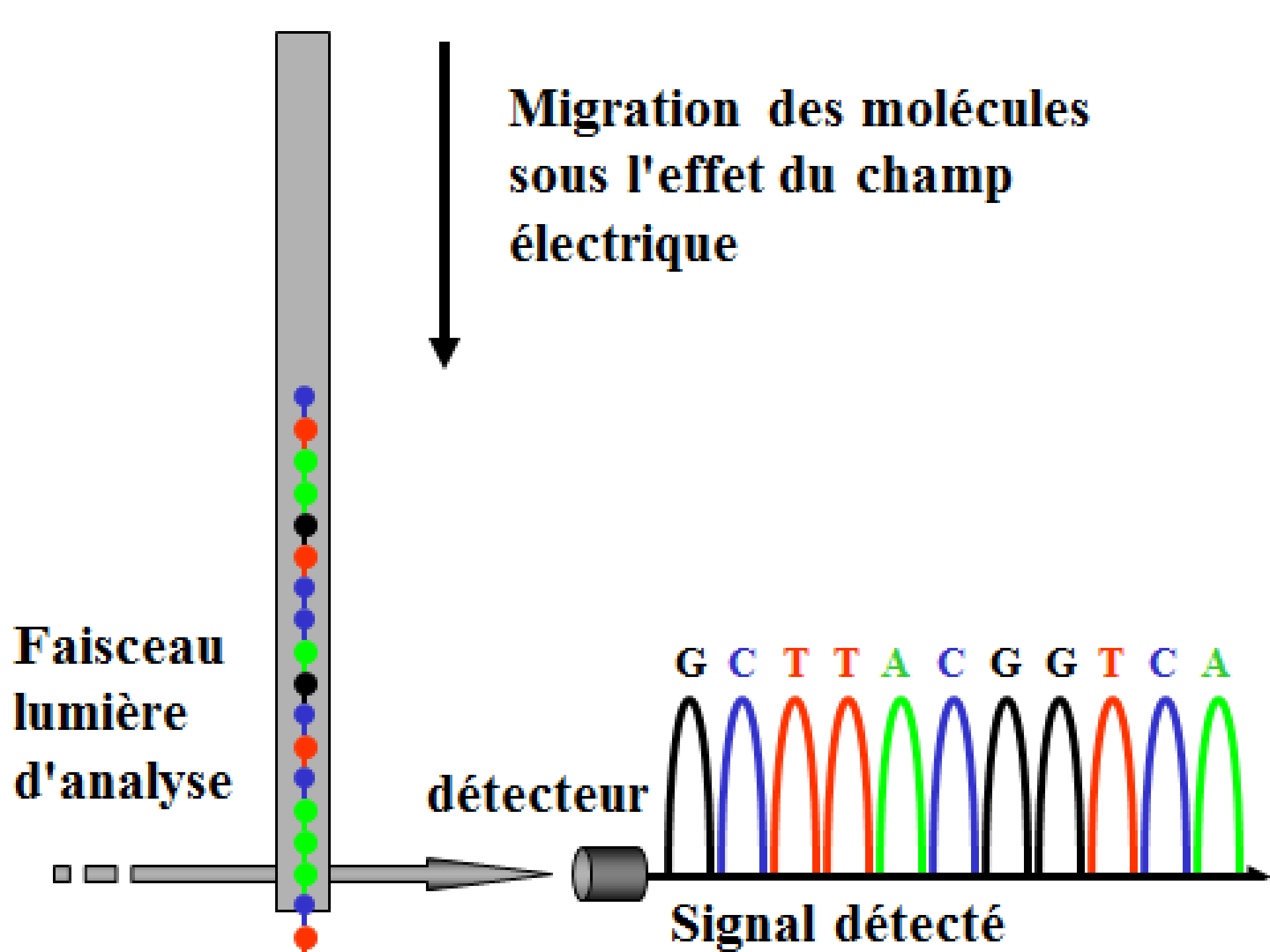

Migration des molécules sous l'effet du champ électrique
Faisceau lumière d'analyse
détecteur
G C T T
Signal détecté

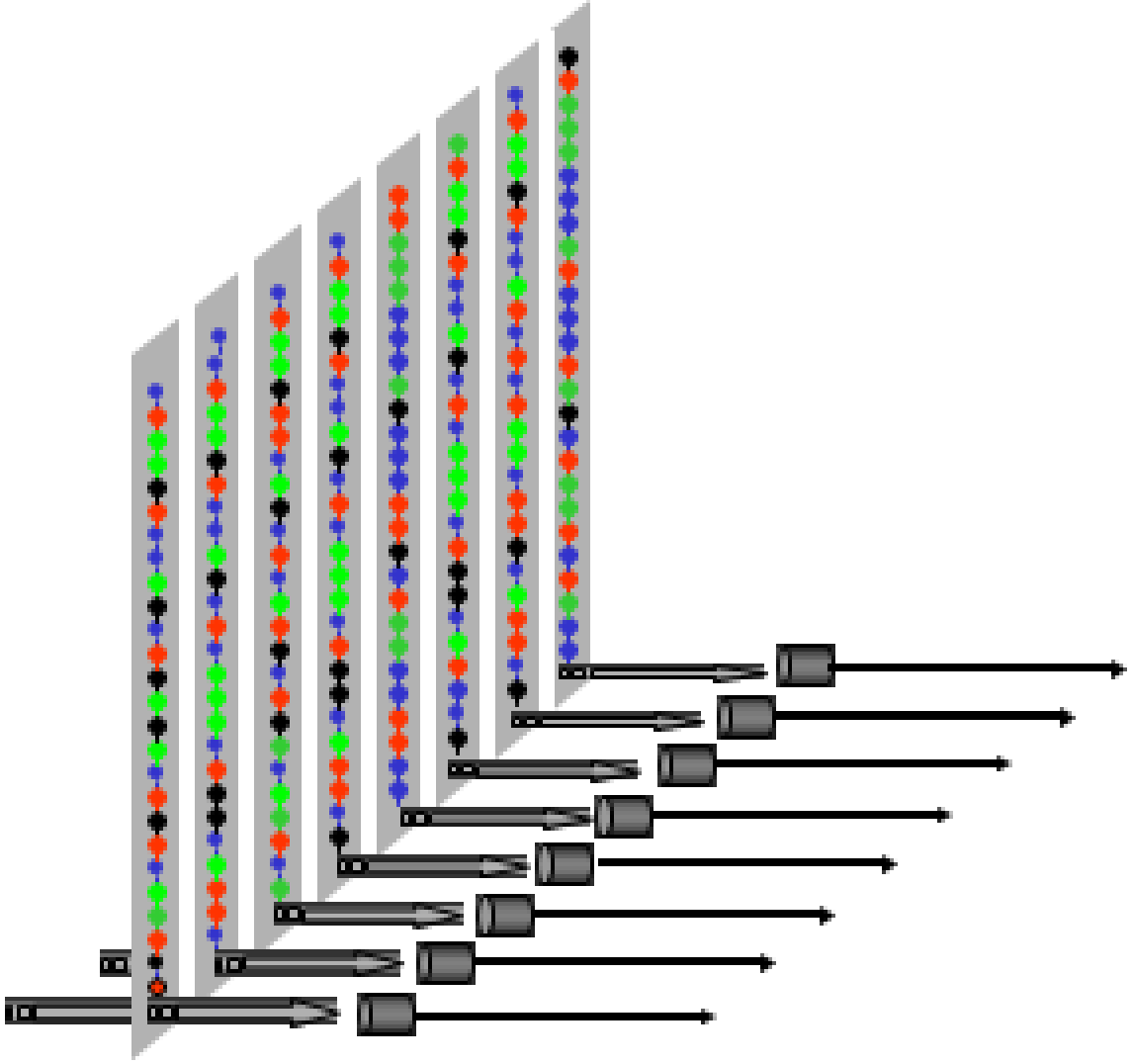

Migration des molécules sous l'effet du champ électrique
Faisceau lumière d'analyse
détecteur
G C T T A C G
Signal détecté


Migration des molécules sous l'effet du champ électrique
Faisceau lumière d'analyse
détecteur
G C T T A C G G
Signal détecté


Migration des molécules sous l'effet du champ électrique
Faisceau lumière d'analyse
détecteur
G C T T A C G G T
Signal détecté


Migration des molécules sous l'effet du champ électrique
Faisceau lumière d'analyse
détecteur
G C T T A C G G T C
Signal détecté

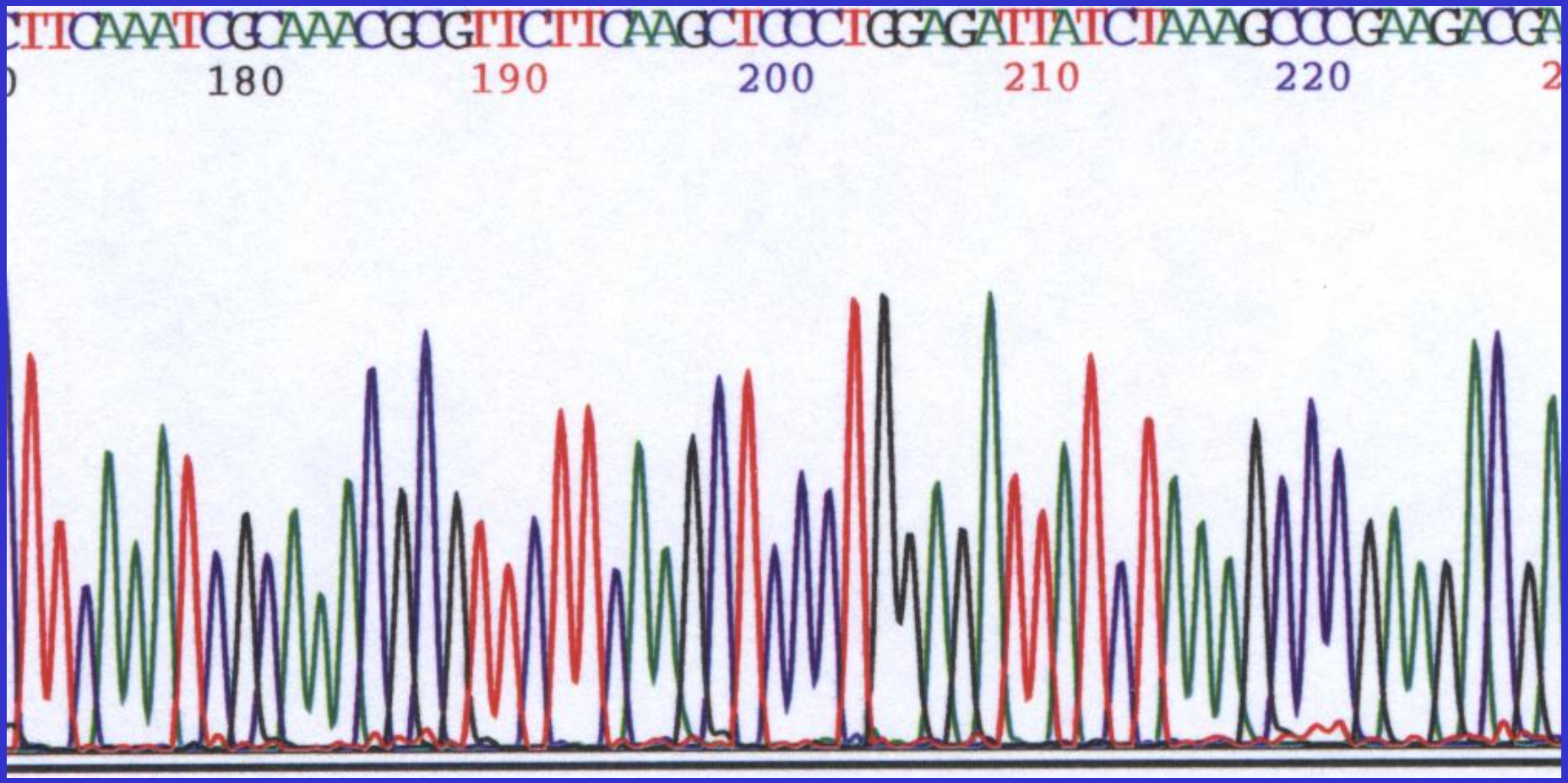

Migration des molécules sous l'effet du champ électrique
Faisceau lumière d'analyse
détecteur
G C T T A C G G T C A
Signal détecté



Analyse simultannée de plusieurs échantillons

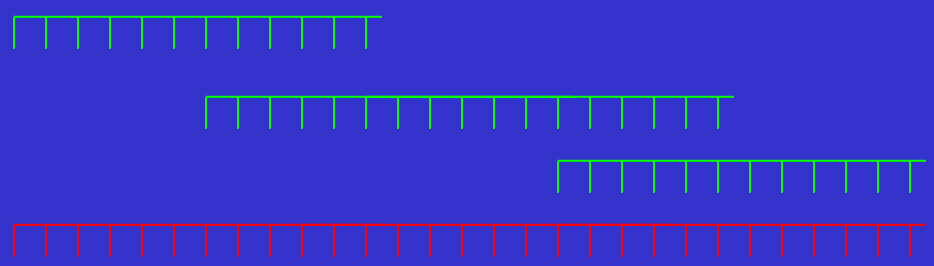
Les 83 séquenceurs LiCor de Genoscope, le seul modèle donnant des séquences unitaires de plus de 900 bases (les autres marques donnaient des séquences de 500 à 600 bases).

GENOSCOPE Centre National de Séquençage



Centre National de Séquençage
GENOSCOPE

Les programmes d ’assemblage établissent des contigs de séquence en recherchant des séquences chevauchantes entre les différents fragments.



contig

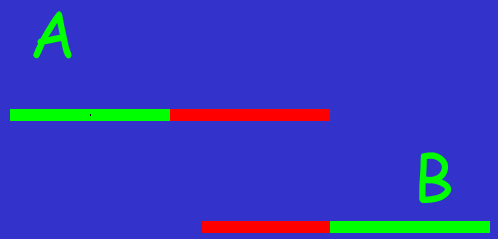
Le processus est mené en mode itératif pour étendre le contig à la longueur maximum.

Le problème des génomes eucaryotes (plantes, animaux, homme) réside dans le fait qu'un grande proportion de ces génomes est constituée de petites séquences de 1 kb à plusieurs dizaines de kb de long qui sont chacune présentes entre 1000 et 100.000 copies dispersées le long des chromosomes.

Ces séquences répétées posent une réelle difficulté dans le phase d'assemblage , qui consiste à étendre les contigs pour finir avec les 24 chromosomes:

Les robots de séquençage donnant des morceaux unitaires de 600 pb en moyenne (800 à 900 pb sur les Licor de Genoscope), les séquences répétées conduisent à des contigs chimériques qui n'ont aucune réalité, comme indiqué ci après.

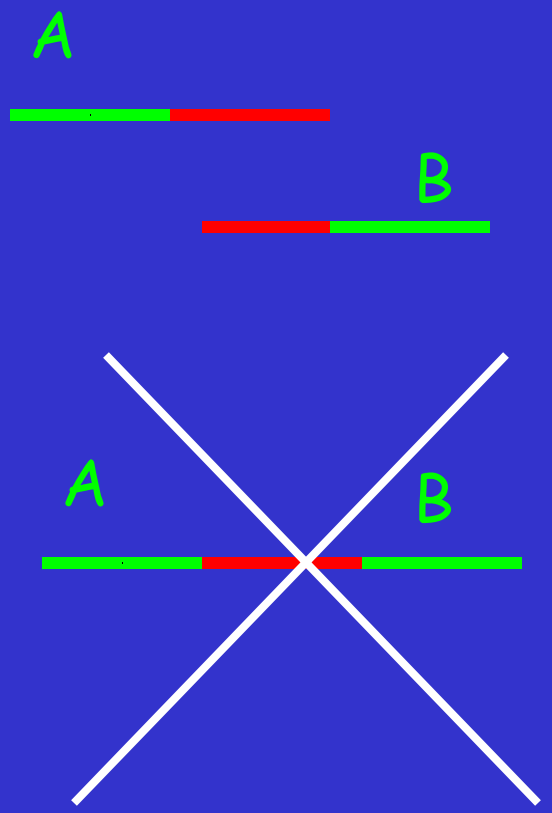
Le problème....



En vert, séquences uniques

En rouge, séquence répétée

Le problème….



En vert, séquences uniques

En rouge, séquence répétée

Il faut donc disposer de fragments d'ADN plus longs que les séquences répétées les plus longues.

Les différents types de banques de fragments d'ADN

Vecteur	insert	Nombre de clones nécessaires	
		1 X	15 X
YAC	500 kb	6.000	90.000
BAC	150 kb	20.000	300.000

300.000

Clones en microplaques de 384 puits = 782 micro-plaques ,soit 1/2 congélateur -80

Microplaque à 96 puits
13 cm x 8 cm

Le séquençage version public

- Stratégie «hierachique BAC à BAC », avec cartographie de pré et per séquençage.

1. - construction de la banque BAC
2. - Séquençage des extrémités des BACs
3. - séquençage individuel de BACs
4. - stratégies de marche BAC à BAC le long des chromosomes.

Stratégie de séquençage choisie par HGP : le séquençage par shotgun hiérachique.

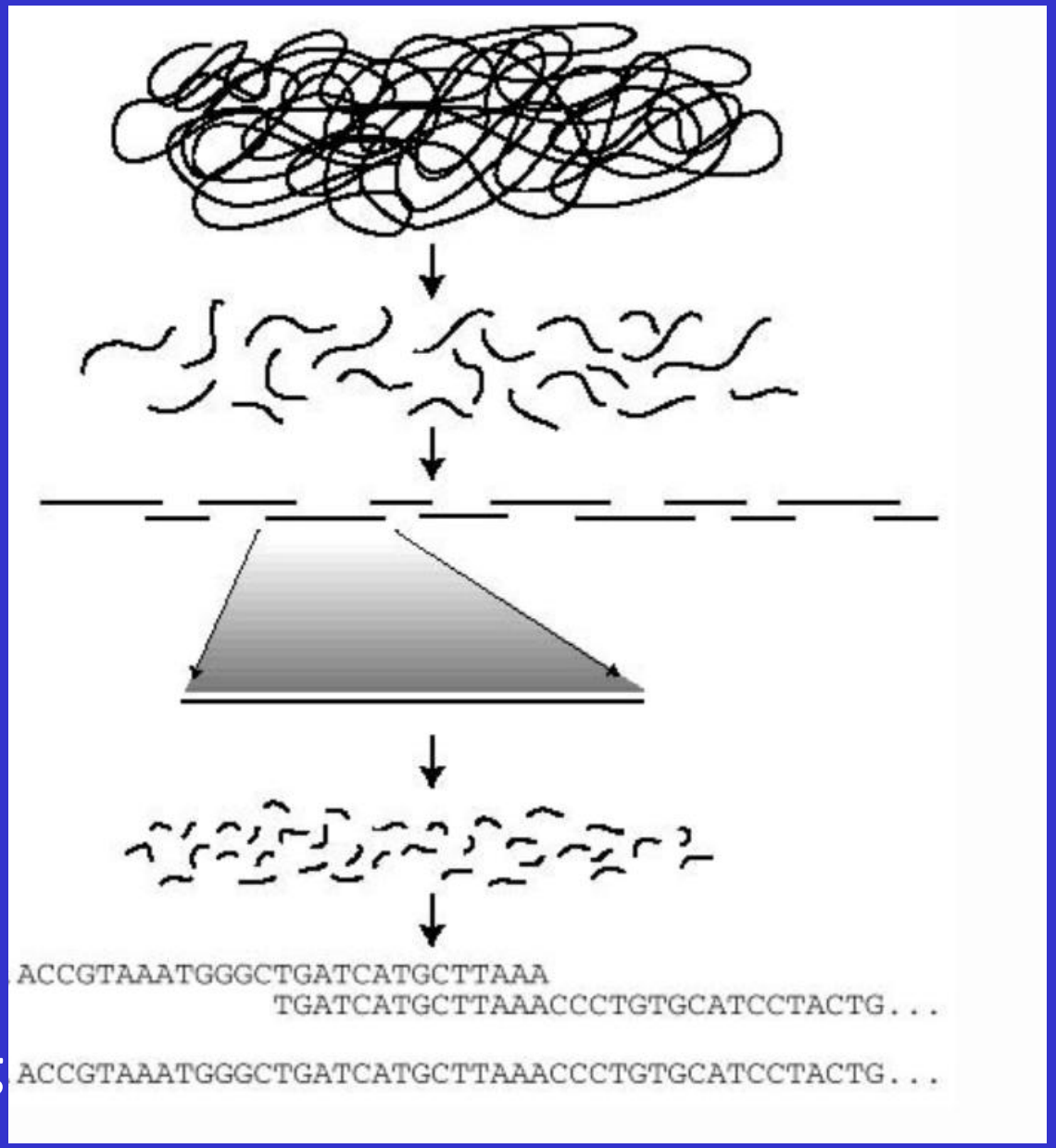
ADN isolé de noyaux

Construction d'une banque de BACs 150kb

Choix des BACs recouvrants à séquencer pour couvrir le chromosome.

Séquençage individuel des BACs par sous-clonage

Assemblage des séquences en 1 seul contig = séquence du BAC



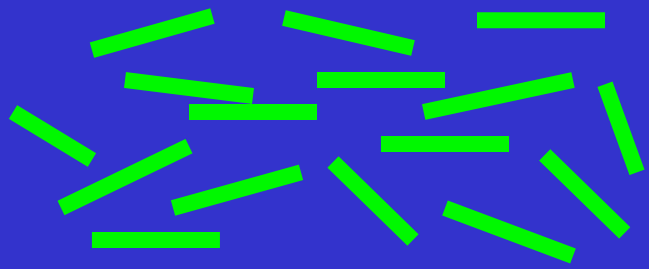
BAC end sequencing

Les 2 extrémités de chaque BAC sont séquencées sur tous les BACs de la banque BAC → Database des BES.

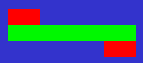
Pour 300.000 BACs, il y aura 600.000 BES appariés dans la Database des BES

Le séquençage d'un BAC

150 kb



1°) l'ADN d'un BAC est découpé par cassures mécaniques aléatoires (seringue avce aiguille fine)en fragments de 3 kb. Ces fragments sont clonés dans un vecteur.

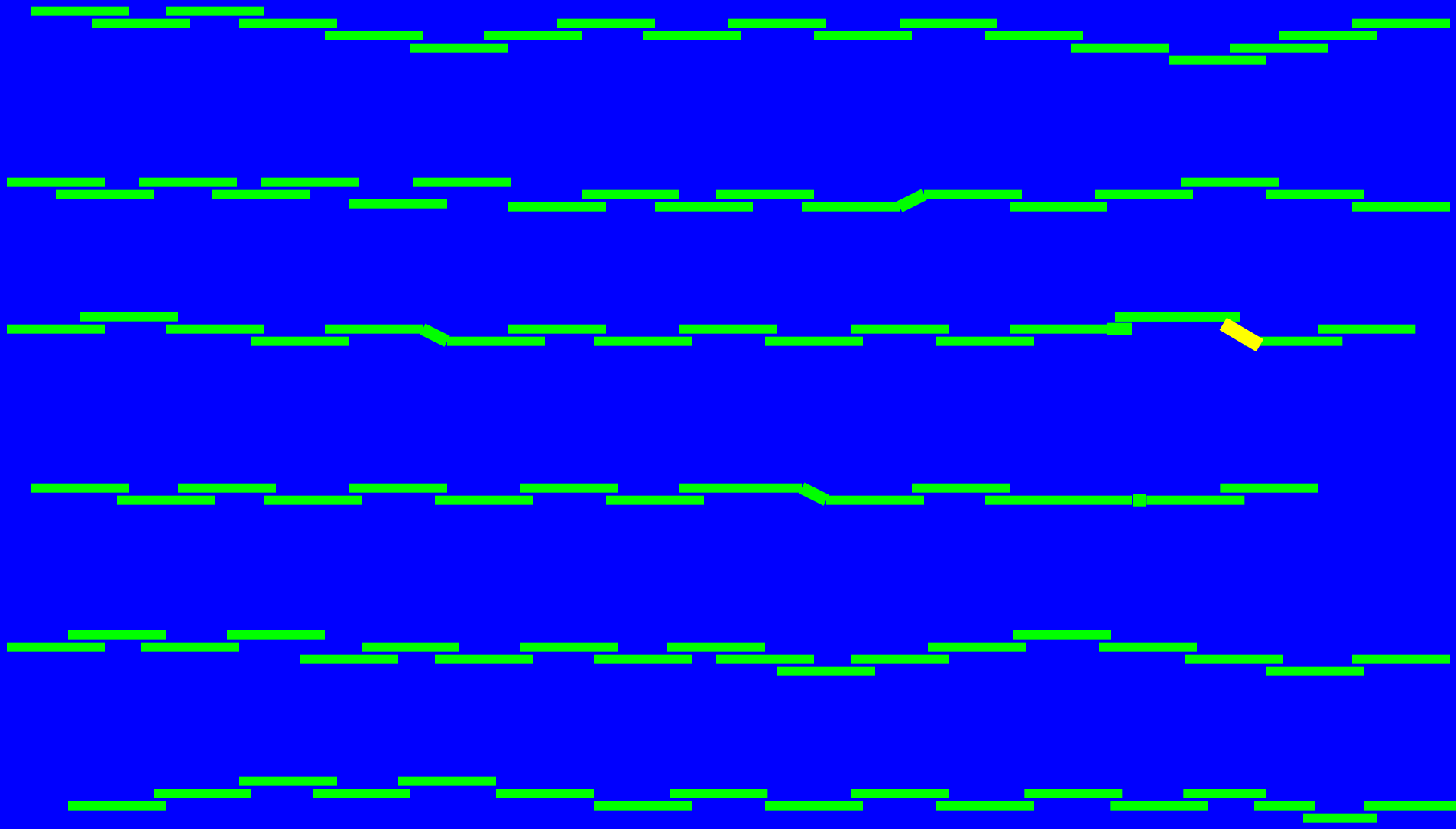


2°) chaque sous-clone est séquencé aux 2 extrémités (600 à 900 pb suivant les machines). L'opération est menée jusqu'à une couverture de 12 fois la longueur du BAC avec la sous-banque 3 kb. Les lectures sont contiguées par les séquences recouvrantes. La séquence du BAC est amenée à la qualité standard décidée par le HGP (erreurs < 10^{-5}) si nécessaire

Pour commencer le séquençage d'un chromosome, un jeu de BACs est choisi de manière la plus équidistante entre eux en utilisant les marqueurs moléculaires de la carte génétique de ce chromosome qui s'hybrident sur les BACs.

Ancrage de BACs sur une carte génétique

Couverture d'un chromosome par BACs recouvrants, grâce aux séquences d'extrémités de BACs + quelques PCR…



1 BAC

PCR

Projet public HGP:

- séquençage hiérarchisé BAC par BAC

- Choix des BACs chevauchant pour couvrir
- chaque chromosome= pas de trou final.

- bon assemblage des séquences (qualité
- HGP)

- mise sur le web des résultats au fur et à mesure, accessibles à toute personne.

Accélérations successives du programme de séquençage du génome humain par le consrtium public HGP

2015

…

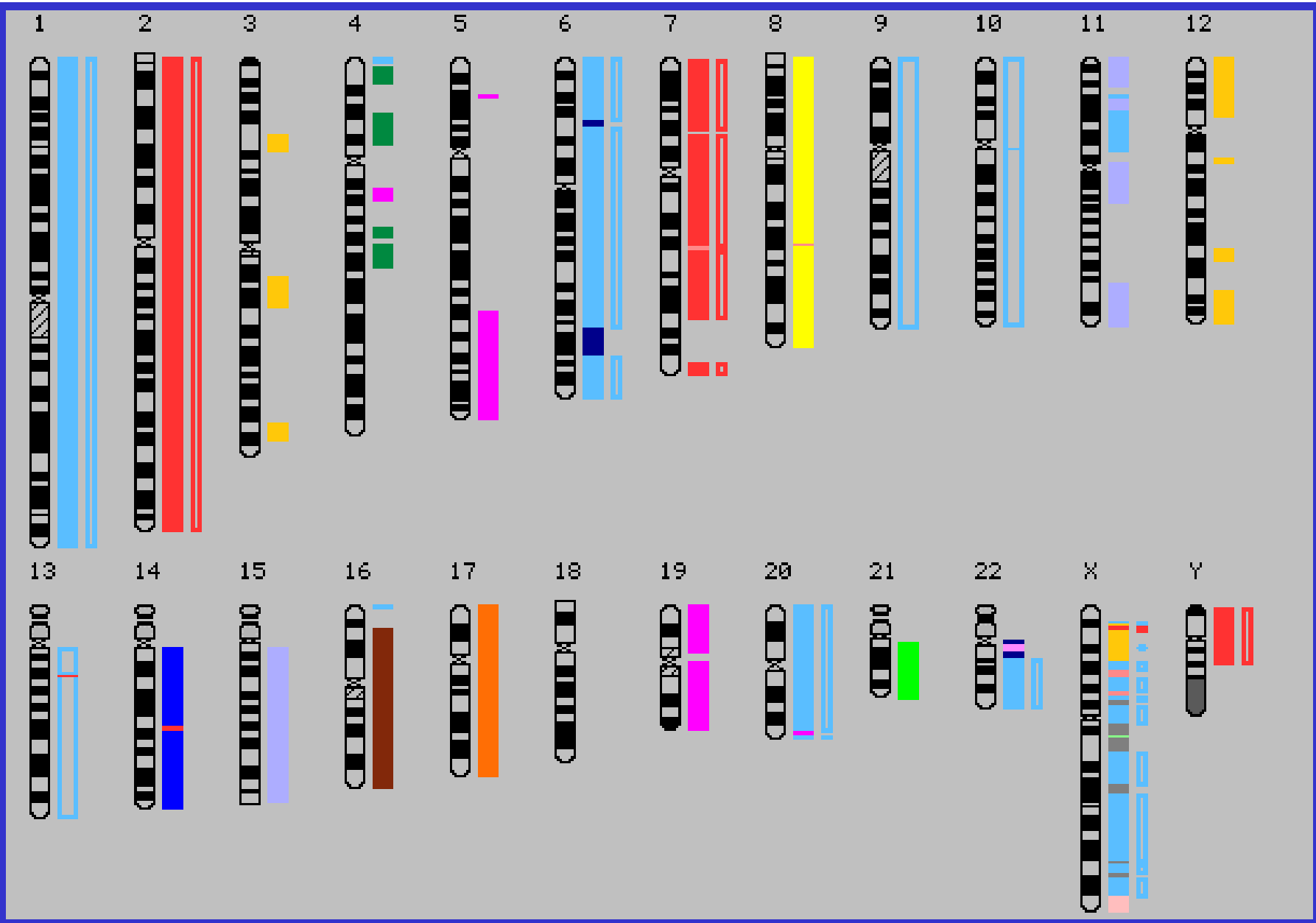
2010

2008

2005

2003

" YALTA " des chromosomes humains initial (pour 2003)



Accélérations successives de la fin prévue pour le séquençage du génome humain par le consortium public HGP.

Annonce faite en janvier 1999

Annonce par une firme privée, CELERA (USA), d'un programme de séquençage du génome humain pour 2001, dans une version d'ébauche, déjà très élaborée, mais l'accès aux résultats seront payants...

Projet privé CELERA :Whole Genome Sequencing (WGS)

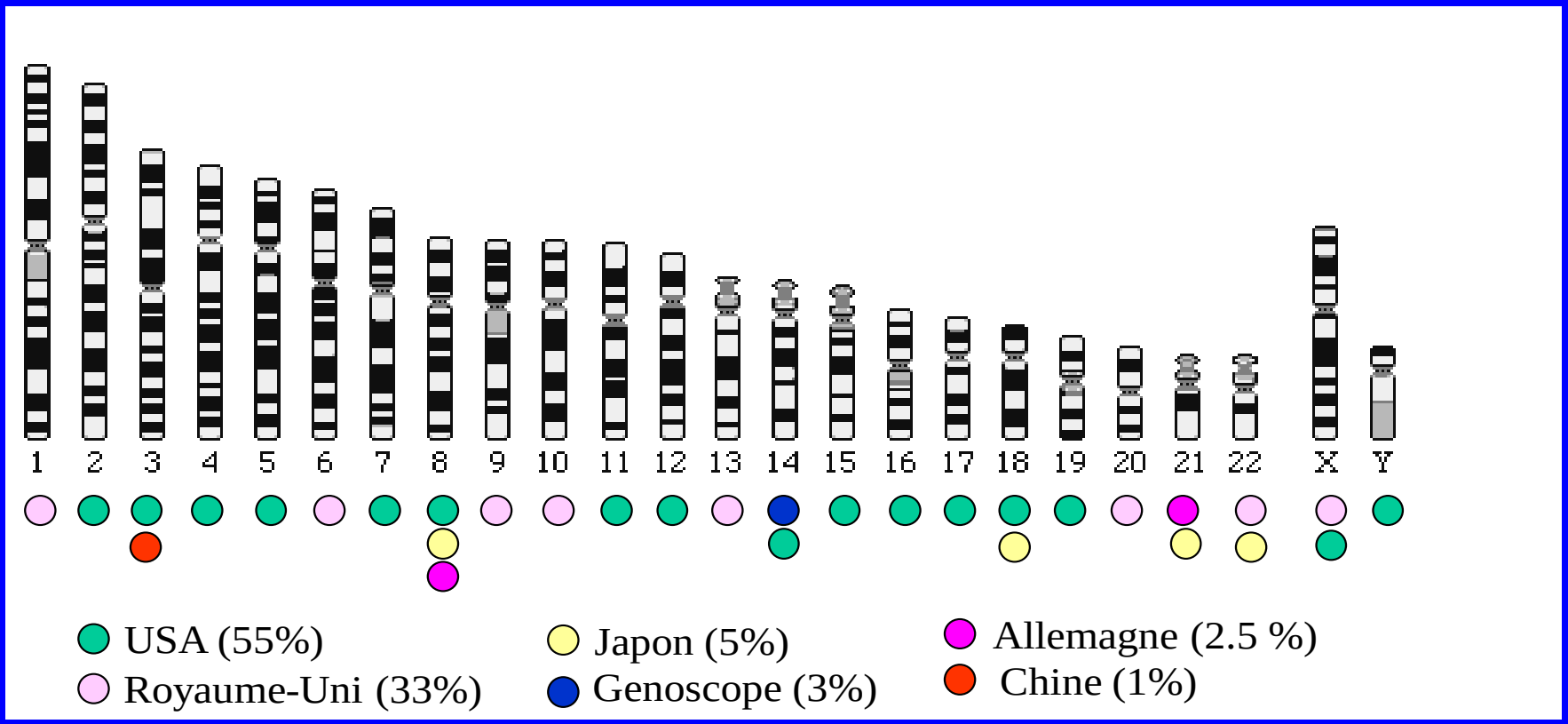
- séquençage en "whole shotgun", annoncé comme pouvant être assemblé par de gros moyens informatiques, **sans aucun besoin de cartographie…**

- Dans ces conditions, HGP prédit que le probleme des séquences répétées vresus la longueur des lectures <1000pb ne pourra pas être surmonté par CELERA

- résultats non déposés régulièrement sur le web public,.

- **L'accès aux données, quand elles seront publiées, en 2001, sera payant= 20 Millions $.**

En réaction, HGP annonce une accélération par concentration sur 9 centres de séquençage répartis sur 9 pays et augmentation tres significative du débit, pour atteindre la complétion à 90% en 2000, soit une année plus tôt que l'annonce faite par CELERA.*



- **Le Genoscope reste responsable du séquençage du chromosome 14**

9 Décembre 1999

Le séquençage des 24 chromosomes du génome humain, soit 3,2 milliards de paires de bases, en version "brouillon", déjà de bonne qualité et à 90 % de complétion a été obtenu le

26 Juin 2000

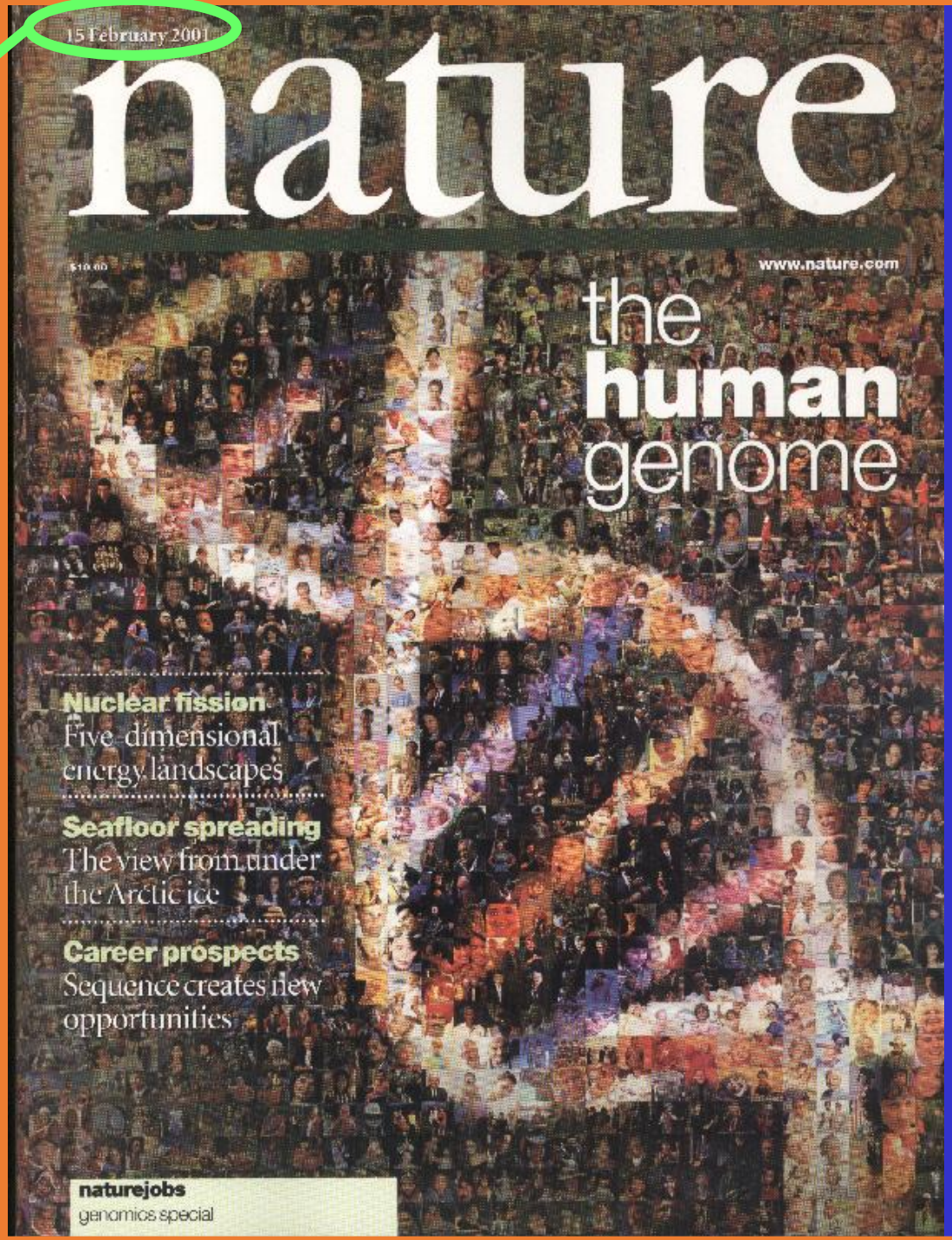
Craig VENTER Bill CLINTON Francis COLLINS

The DNA sequence and comparative analysis of human chromosome 20

P. Deloukas, L. H. Matthews, J. Ashurst, J. Burton, J. G. R. Gilbert, M. Jones, G. Stavrides, J. P. Almeida, A. K. Babbage, C. L. Bagguley, J. Bailey, K. F. Barlow, K. N. Bates, L. M. Beard, D. M. Beare, O. P. Beasley, C. P. Bird, S. E. Blakey, A. M. Bridgeman, A. J. Brown, D. Buck, W. Burrill, A. P. Butler, C. Carder, N. P. Carter, J. C. Chapman, M. Clamp, G. Clark, L. N. Clark, S. Y. Clark, C. M. Clee, S. Clegg, V. E. Cobley, R. E. Collier, R. Connor, N. R. Corby, A. Coulson, G. J. Coville, R. Deadman, P. Dhami, M. Dunn, A. G. Ellington, J. A. Frankland, A. Fraser, L. French, P. Garner, D. V. Grafham, C. Griffiths, M. N. D. Griffiths, R. Gwilliam, R. E. Hall, S. Hammond, J. L. Harley, P. D. Heath, S. Ho, J. L. Holden, P. J. Howden, E. Huckle, A. R. Hunt, S. E. Hunt, K. Jekosch, C. M. Johnson, D. Johnson, M. P. Kay, A. M. Kimberley, A. King, A. Knights, G. K. Laird, S. Lawlor, M. H. Lehvaslaiho, M. Leversha, C. Lloyd, D. M. Lloyd, J. D. Lovell, V. L. Marsh, S. L. Martin, L. J. McConnachie, K. McLay, A. A. McMurray, S. Milne, D. Mistry, M. J. F. Moore, J. C. Mullikin, T. Nickerson, K. Oliver, A. Parker, R. Patel, T. A. V. Pearce, A. I. Peck, B. J. C. T. Phillimore, S. R. Prathalingam, R. W. Plumb, H. Ramsay, C. M. Rice, M. T. Ross, C. E. Scott, H. K. Sehra, R. Shownkeen, S. Sims, C. D. Skuce, M. L. Smith, C. Soderlund, C. A. Steward, J. E. Sulston, M. Swann, N. Sycamore, R. Taylor, L. Tee, D. W. Thomas, A. Thorpe, A. Tracey, A. C. Tromans, M. Vaudin, M. Wall, J. M. Wallis, S. L. Whitehead, P. Whittaker, D. L. Willey, L. Williams, S. A. Williams, L. Wilming, P. W. Wray, T. Hubbard, R. M. Durbin, D. R. Bentley, S. Beck & J. Rogers

The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK

15 Février 2001

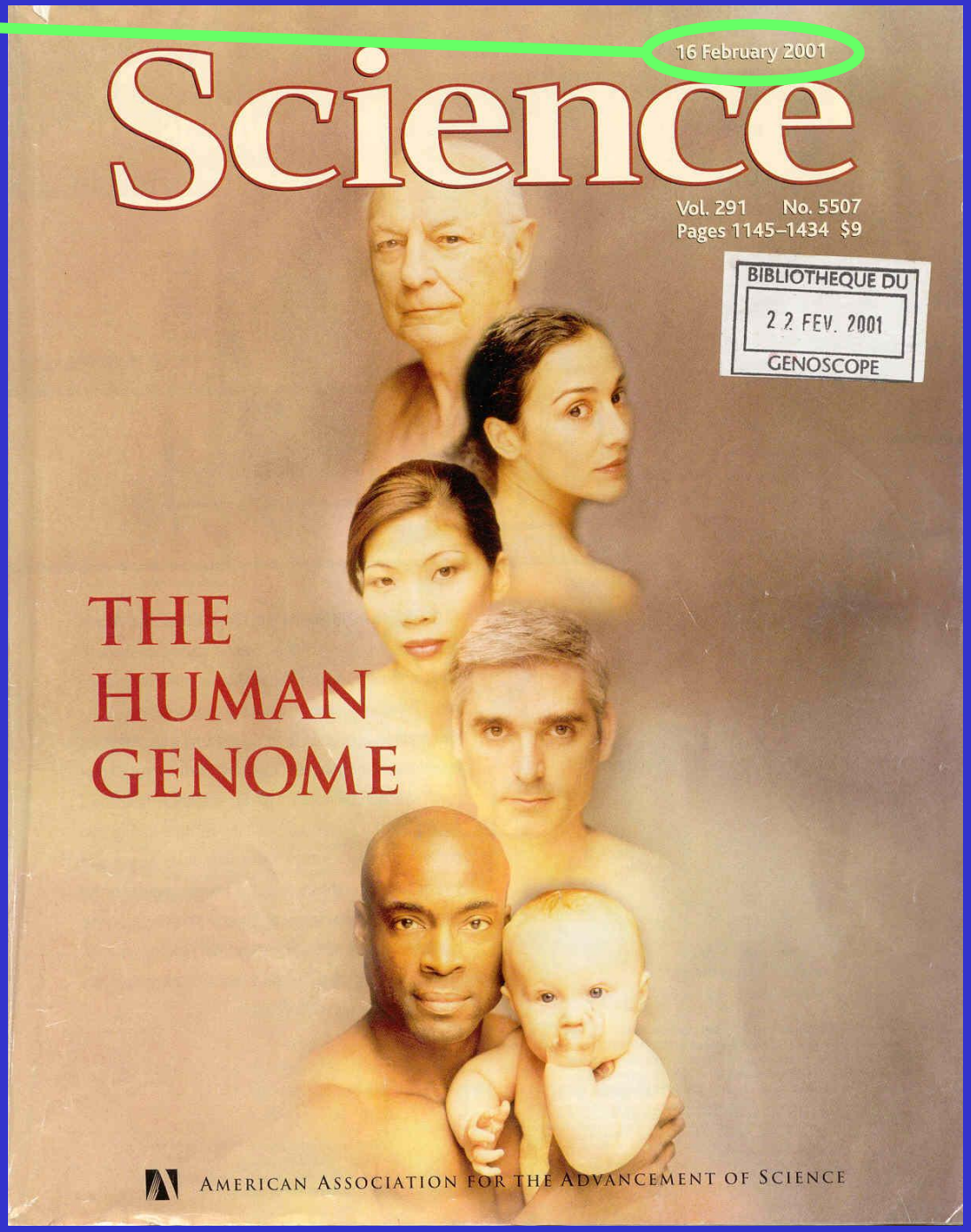


16 Février 2001

La qualité des séquences dans les contigs est bonne (erreurs<1. 10-4)

mais

il y a près de 200.000 trous ….dont les tailles ne sont pas connues!



Projet privé CELERA :

- séquençage en "whole shotgun", annoncé comme pouvant être assemblé par de gros moyens informatiques sans aucun besoin de cartographie.
(ce dont on a pu douter au vu des résultats obtenus sur la Drosophile…)

IMPOSTURE

- résultats non déposés sur le web public, séquences non publiées (accés payant). Accessibilité éventuelle et partielle pour les chercheurs du public.

On the sequencing of the human genome

Robert H. Waterston*†, Eric S. Lander‡, and John E. Sulston§

*Genome Sequencing Center, Washington University, Saint Louis, MO 63108; ‡Whitehead Institute/Massachusetts Institute of Technology Center for Genome Research, Cambridge, MA 02142; and §Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom

Communicated by Aaron Klug, Medical Research Council, Cambridge, United Kingdom, December 21, 2001 (received for review November 7, 2001)

Two recent papers using different approaches reported draft sequences of the human genome. The international Human Genome Project (HGP) used the hierarchical shotgun approach, whereas Celera Genomics adopted the whole-genome shotgun (WGS) approach. Here, we analyze whether the latter paper provides a meaningful test of the WGS approach on a mammalian genome. In the Celera paper, the authors did not analyze their own WGS data. Instead, they decomposed the HGP's assembled sequence into a "perfect tiling path", combined it with their WGS data, and assembled the merged data set. To study the implications of this approach, we perform computational analysis and find that a perfect tiling path with 2-fold coverage is sufficient to recover virtually the entirety of a genome assembly. We also examine the manner in which the assembly was anchored to the human genome and conclude that the process primarily depended on the HGP's sequence-tagged site maps, BAC maps, and clone-based sequences. Our analysis indicates that the Celera paper provides neither a meaningful test of the WGS approach nor an independent sequence of the human genome. Our analysis does not imply that a WGS approach could not be successfully applied to assemble a draft sequence of a large mammalian genome, but merely that the Celera paper does not provide such evidence.

problem is straightforward, because reads with overlapping sequence can typically be merged together without risk of misassembly. The relatively few gaps and problems can be solved to produce complete sequences. The approach has been applied successfully to produce complete sequences of simple genomes such as plasmids, viruses, organelles, and bacteria. Whole-genome shotgun data alone also has been applied with an almost 15-fold redundancy (5) to produce a draft sequence of the euchromatic portion of the *Drosophila* genome (3% repeat content), although a clone-based strategy is being applied to convert this to a finished sequence.

A greater challenge arises in tackling complex genomes with a large proportion of repeat sequences that can give rise to misassembly. Two alternative approaches (Fig. 1) can be taken.

Hierarchical shotgun (HS) assembly. In this approach, the genome is first broken up into an overlapping collection of intermediate clones such as bacterial artificial chromosomes (BACs). The sequence of each BAC is determined by shotgun sequencing, and the sequence of the genome is obtained by merging the sequences of the BACs. The HS approach provides a guaranteed route for producing an accurate finished genome sequence, because the sequence assembly is local and anchored to the genome. But it requires some additional preliminary work

The DNA sequence and analysis of human chromosome 14

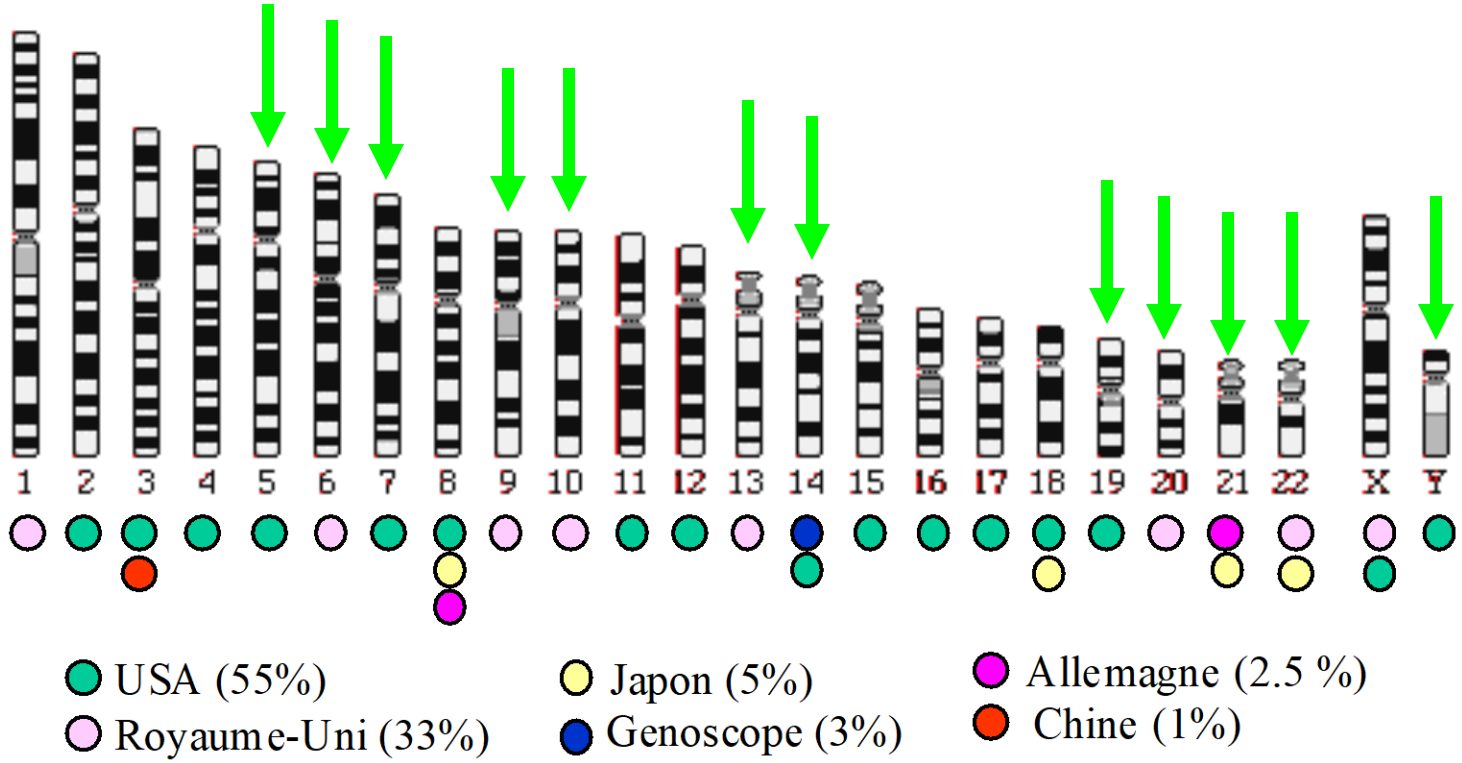
Q1

Roland Heilig*†, Ralph Eckenberg*†, Jean-Louis Petit*†, Núria Fonknechten*†, Corinne Da Silva*†, Laurence Cattolico*, Michaël Levy*, Valérie Barbe*, Véronique de Berardinis*, Abel Ureta-Vidal*, Eric Pelletier*†, Virginie Vico*, Véronique Anthouard*, Lee Rowen‡, Anup Madan‡, Shizhen Qin‡, Hui Sun§, Hui Du§, Kymberlie Pepin§, François Artiguenave*, Catherine Robert*, Corinne Cruaud*, Thomas Brüls*, Olivier Jaillon*†, Lucie Friedlander*, Gaelle Samson*†, Philippe Brottier*, Susan Cure*, Béatrice Ségurens*, Franck Anière*, Sylvie Samain*, Hervé Crespeau*, Nissa Abbasi‡, Nathalie Aiach*, Didier Boscus*, Rachel Dickhoff‡, Monica Dors‡, Ivan Dubois*, Cynthia Friedman‡, Michel Gouyvenoux*, Rose James‡, Anuradha Madan‡, Barbara Mairey-Estrada*, Sophie Mangenot*, Nathalie Martins*, Manuela Ménard*, Sophie Oztas*, Amber Ratcliffe‡, Tristan Shaffer‡, Barbara Trask‡, Benoit Vacherie*, Chadia Bellemere*, Caroline Belser*, Marielle Besnard-Gonnet*, Delphine Bartol-Mavel*, Magali Boutard*, Stéphanie Briez-Silla*, Stephane Combette*, Virginie Dufossé-Laurent*, Carolyne Ferron*, Christophe Lechaplais*, Claudine Louesse*, Delphine Muselet*, Ghislaine Magdelenat*, Emilie Pateau*, Emmanuelle Petit*, Peggy Sirvain-Trukniewicz*, Arnaud Trybou*, Nathalie Vega-Czarny*, Elodie Bataille*, Elodie Bluet*, Isabelle Bordelais*, Maria Dubois*, Corinne Dumont*, Thomas Guérin*, Sébastien Haffray*, Rachid Hammadi*, Jacqueline Muanga*, Virginie Pellouin*, Dominique Robert*, Edith Wunderle*, Gilbert Gauguet*, Alice Roy*, Laurent Sainte-Marthe*, Jean Verdier*, Claude Verdier-Discalla*, LaDeana Hillier§, Lucinda Fulton§, John McPherson§, Fumihiko Matsuda||, Richard Wilson§, Claude Scarpelli*, Gábor Gyapay*, Patrick Wincker*, William Saurin*, Francis Quétier*†, Robert Waterston§, Leroy Hood‡ & Jean Weissenbach*†

** Genoscope-Centre National de Séquençage; † UMR-3080, CNRS et Université d'Evry; and || Centre National de Génotypage, 91000, Evry, France*
‡ Institute for Systems Biology, Seattle, Washington 98103, USA
§ Genome Sequencing Center, Washington University School of Medicine, St Louis, Missouri 63108, USA

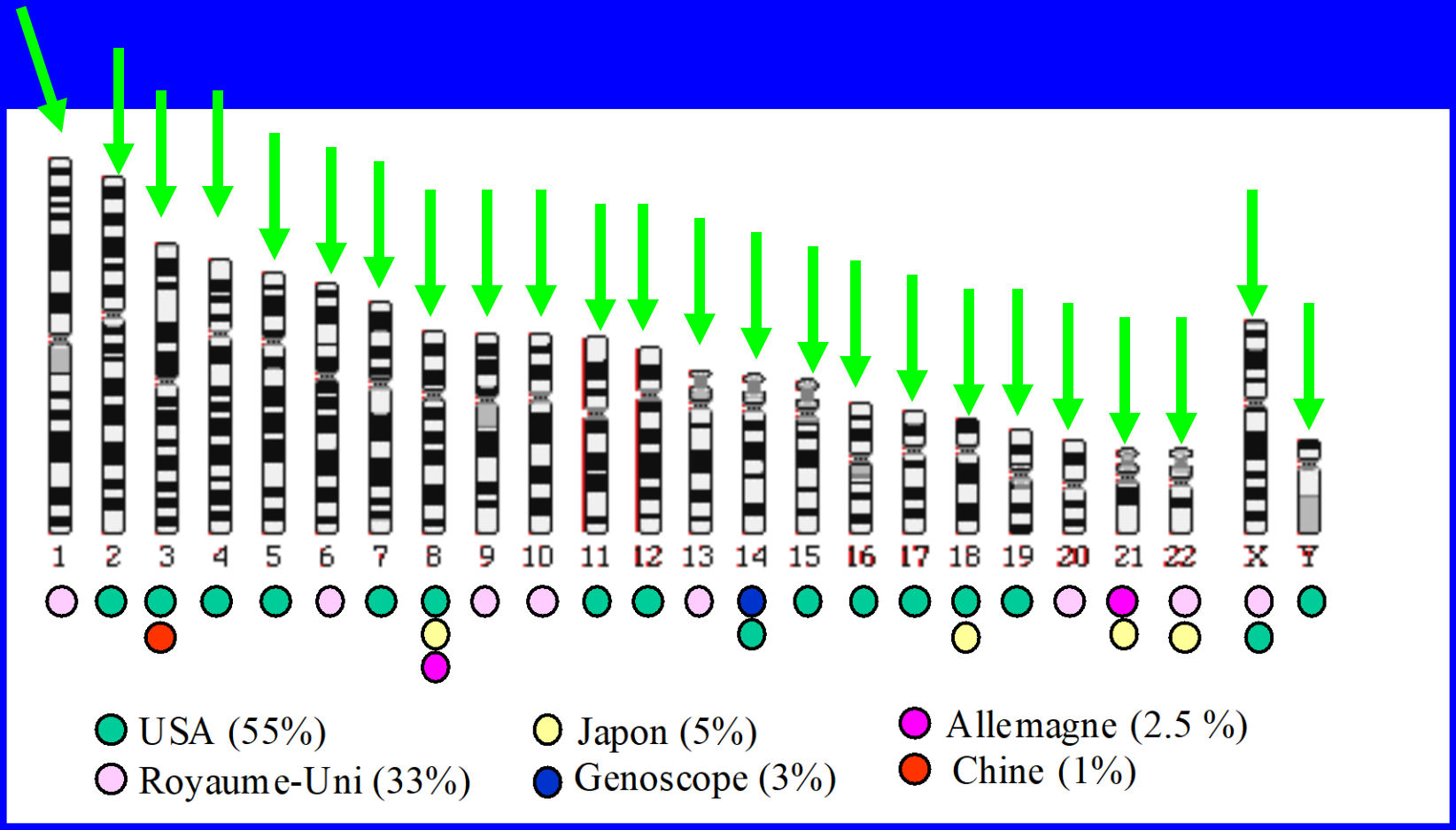
Chromosome 14 is one of five acrocentric chromosomes in the human genome. These chromosomes are characterized by a heterochromatic short arm that contains essentially ribosomal RNA genes, and a euchromatic long arm in which most, if not all, of the protein-coding genes are located. The finished sequence of human chromosome 14 comprises 87,410,661 base pairs (bp), representing 100% of its euchromatic portion, in a single continuous segment covering the entire long arm with no gaps. Two loci of crucial importance for the immune system, as well as more than 60 disease genes, have been localized so far on chromosome 14. We identified 1,050 genes and gene fragments, and 393 pseudogenes. On the basis of comparisons with other vertebrate genomes, we estimate that more than 96% of the chromosome 14 genes have been annotated. From an analysis of the CpG island occurrences, we estimate that 70% of these annotated genes are complete at their 5′ end.

ETAT DES CHROMOSOMES FINIS novembre 2004



= chromosomes finis (nov 2004)

Le dernier des 24 chromosomes a été fini en mai 2006



= chromosomes finis (mai 2006)

Les gènes du genome humain

2 étapes

- Séquençage de l ’ADN ==> 24 fichiers informatiques ATGC.....

- Annotation des séquences : identification des régions d ’ ADN codant les gènes pour les
  - rARN r, tARNt, miARN (facile ou relativement facile)
  - protéines (assez difficile au début)

Jnhfeliftehhsncitagjiltionsnjuegsnvousahtegmmpsavezidenghyskjey

feffienjloids tousnjuhsd lesnbejdnsnnnsimessagesbenrfdconggefenu

hegdtttdfehakisnusndnhiieuevshdansvgefrdfdfcefichierhenhsgffdfs.

Jnhfeli ftehhsncita gjil tionsnjuegsnvous ahtegmmpsavezs idens tighyskjey

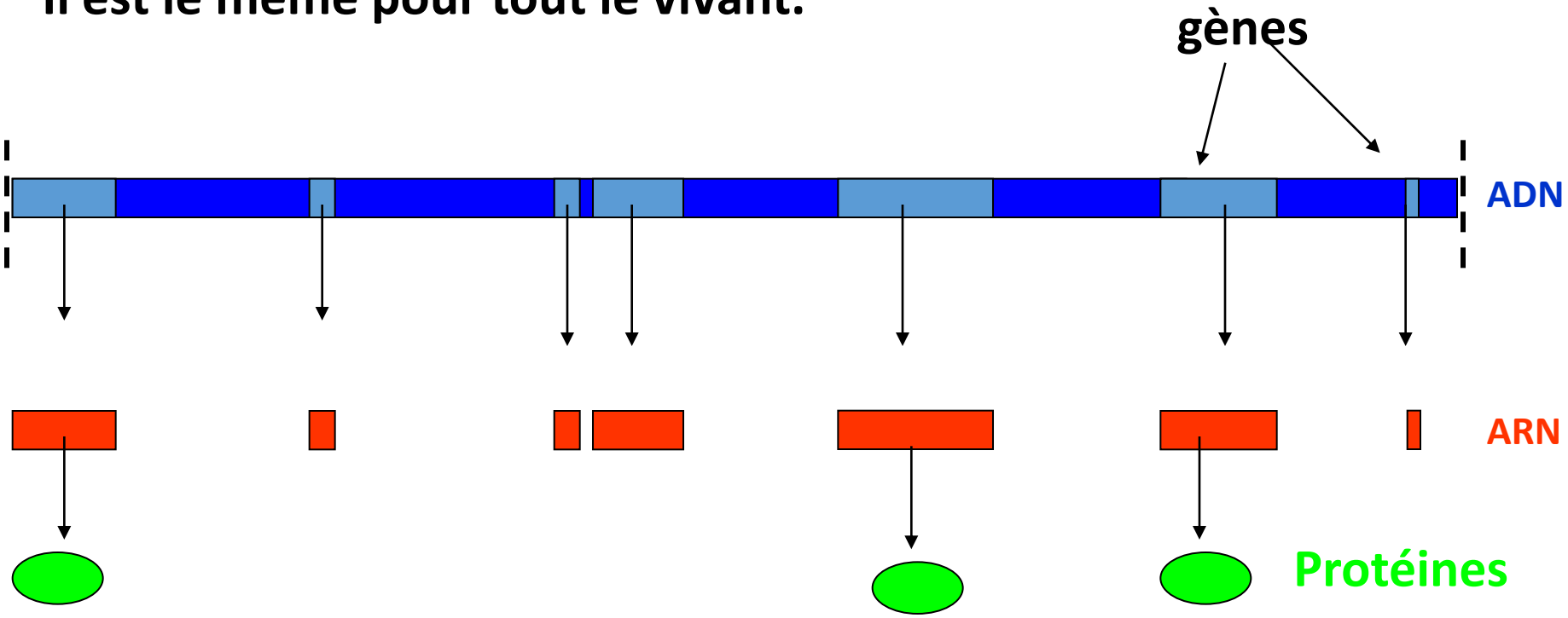
fe fie njloids tous njuhsd lesnbejdnsnnnsi messagesbenrfd contggef enus

kegdtttdfehakisnusndnhiieuevsh dansce vgefrdfdfce fichier henhsgffdfs

Felicitations vous avez identifie tous les messages dans ce fichier

Le flux général de l'information génétique

Il est le même pour tout le vivant.



Une portion de l'ADN n'est ni transcrite ni traduite

Une portion de l'ADN est transcrite en ARN mais pas traduite

Une portion de l'ADN est transcrite en ARN qui sont traduits en protéines.

Jusqu ’en Mai 2000, l ’estimation du nombre de gènes codant des protéines chez l ’homme et les vertébrés en général variait entre

100.000 et 120.000

GENOSCOPE a publié en Mai 2000 une nouvelle méthode pour identifier les gènes codant des protéines

Cette méthode est basée sur un simple alignement des séquences de l ’ADN humain sur l ’ADN d ’un autre vertébré, distant de 400 millions d ’années sur l ’arbre évolutif.

Tetraodon nigroviridis

Cette méthode est basée sur un simple alignement des séquences de

l ’ADN humain sur l ’ADN d ’un autre vertébré, distant de

400 millions d ’années sur l ’arbre évolutif.

L’estimation du nombre de gènes codant des protéines est dans la fourchette provisoire de seulement

28.000 à 34.000

Tetraodon nigroviridis

Une autre méthode, basée sur un principe différent et publiée aussi en Mai 2000, aboutit à une estimation maximale de **36.000** gènes codant des protéines

Le nombre maximal de 30,000 gènes codant des protéines a été alors accepté par la communauté scientifique.

L'estimation actuelle est de 20.444 gènes codant des protéines (nombre ne variant plus beaucoup).

Attention, 20.444 gènes codant des protéines ne signifie pas qu'il y aurait 20.444 protéines !

Dans le même temps où le nombre de gènes codant des protéines

a été largement surestimé, la fraction de ces gènes qui utilisent

l'épissage alternatif pour donner chacun plusieurs protéines

a été largement sous- estimée (> 35 %).

Le nombre total de protéines est resté estimé aux alentours de

150 000 . Au 16 juin 2016, il est chiffré à **154.434**

Une nouvelle classe de gènes a été découverte en 2000. Il s'agit des microARN (miRNA), qui sont des gènes modulant négativement la traduction de 60% des protéines humaines. On en a dénombré 2.812 à ce jour.

Sur les 20.313 gènes codant des protéines, plus de 10.000 sont connus pour entraîner une maladie quand ils portent une certaine mutation, tandis que près de 5.000 autres sont en suspicion.

Il y a un consensus pour penser que la majorité des autres gènes puissent être également impliqués dans des maladies, mais la relation entre gene et maladie n'est pas encore établie.

Quant aux 1881 gènes correspondant aux miRNAs, une grande partie d'entre eux sont déjà associés à des maladies.

La médecine personnalisée basée sur le séquençage de génomes individuels

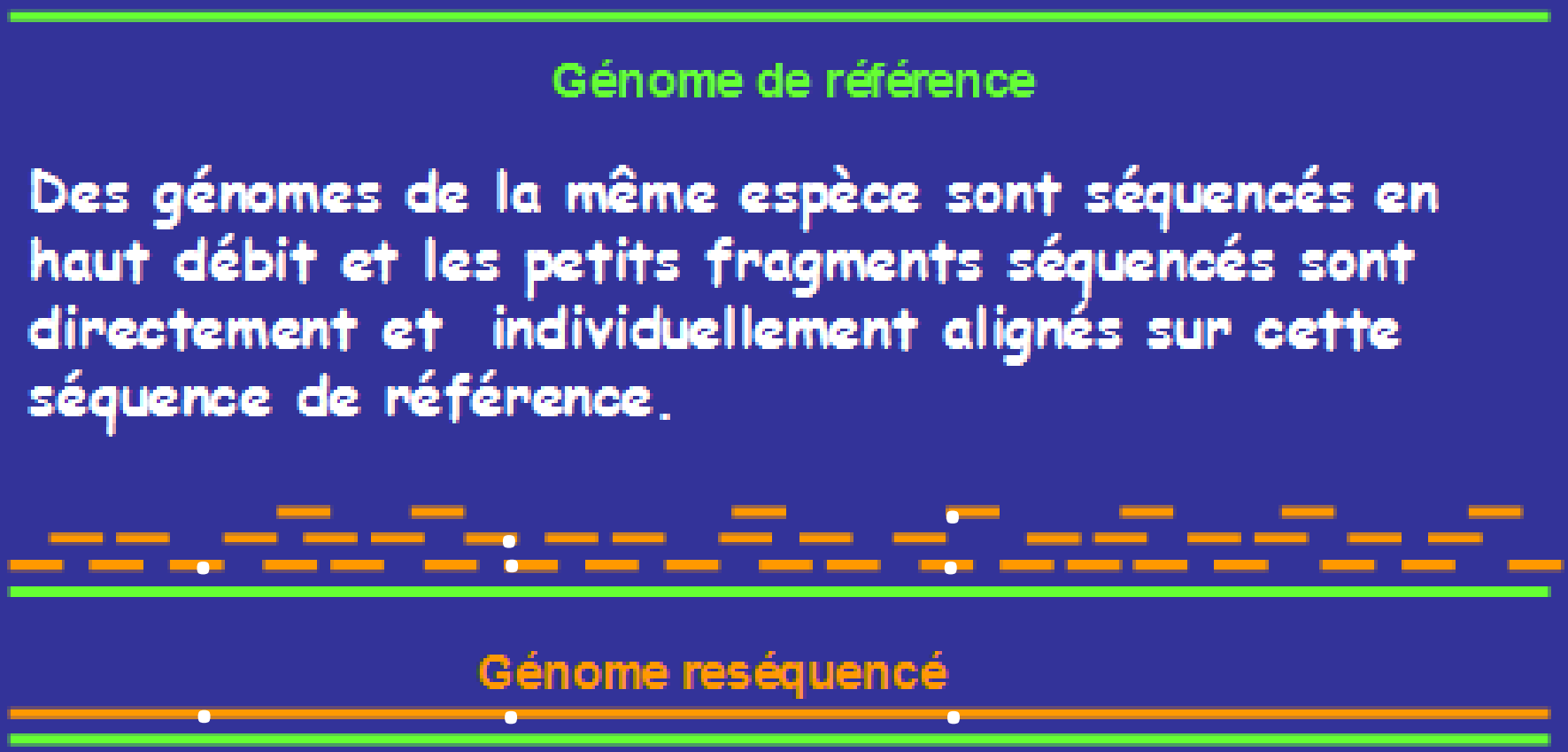
Le séquençage *de novo* d'un génome humain a été accompli entre 1992 et 2006 et a coûté près de 3 Milliards de $.

Des vagues successives d'améliorations technologiques ont permis entre 2006 et 2016

- **d'augmenter le débit d'un robot de séquençage de plus de 2 Millions de fois.**

- **de diminuer le prix du séquençage d'un génome individuel de plus de 2 Millions de fois.**

Le re-séquençage est une opération qui peut être réalisée dès que la génome de référence d'un organisme a été établi.



Des génomes de la même espèce sont séquencés en haut débit et les petits fragments séquencés sont directement et individuellement alignés sur cette séquence de référence.

- Différence ponctuelle (1 à quelques bases) En moyenne, chaque individu diffère de la séquence de référence par 1 base sur 1000, soit environ 2 millions de différences ponctuelles, différentes d'un individu à un autre.

Whole –genome sequencing power

Maximum throughput for population- and production-scale genomics

Le Pack ILLUMINA X-10 comprend 10 séquenceurs Hi-Seq X en réseau.

La puissance de séquençage permet un débit de 18.000 génomes humains par an. Soit 1 genome toutes les demi-heures!

Le prix de revient d'un génome humain est < 1.000 $

Le plus puissant des Séquenceurs d'ADN.

- **8.000 génomes humains par an**
- **1 génome toutes les 30 minutes (en fichier ATGC brut).**
- **Sa taille est un peu plus grosse qu'une machine à laver**

Le NovaSeq 6000 fabriqué par la Sté Illumina (USA)*

La Médecine Personnalisée

Faire bénéficier le médecin/clinicien, le plus rapidement possible, des retombées des avancées applicatives de la recherche pour qu'il les utilise chez ses patients à des fins

✓de prévention des risques

✓d'établissement de diagnostics sûrs et précis

✓de thérapies adaptées pour chaque patient:

la bonne thérapie/molecule au bon patient, à la bonne dose au bon moment, sans effet secondaire (et au moindre coût).

Actuellement, 1 médicament sur 2 donné en première intention par un clinicien est inefficace/inutile.

La succession des étapes de la génomique clinique

Séquences

Identification mutations (SNPs, Indels, CNVs)
Sélection de la (des) mutation(s) causale(s)
Interrogation des bases de données des drogues actionnables pour ces mutations (approuvées ou en développement clinique)

La Médecine Personnalisée

Le pas de temps: plusieurs années
(essais précliniques sur animaux et essais cliniques sur l'homme

Chercheurs

Remontée des observations de N cliniciens sur M patients vers les chercheurs pour générer de nouvelles connaissances qui engendreront de nouveaux médicaments de de nouvelles thérapies

De la recherche au lit du patient

Utilisation **immédiate** des applications **disponibles** et **légalement utilisables** des résultats de la recherche au cas <u>individuel</u> du patient, sous l'encadrement du clinicien.

Des lits de patients vers la recherche

Patient + clinicien

Le pas de temps: quelques semaines

L'un des centres les plus performants actuellement.

Région de NEW-YORK= 19 millions d'individus

16 hopitaux/centres de santé

Responsable: Pr. Robert DARNELL

400 personnes

Séquenceur Illumina X10 + Bionangenomics (cartographie optique)

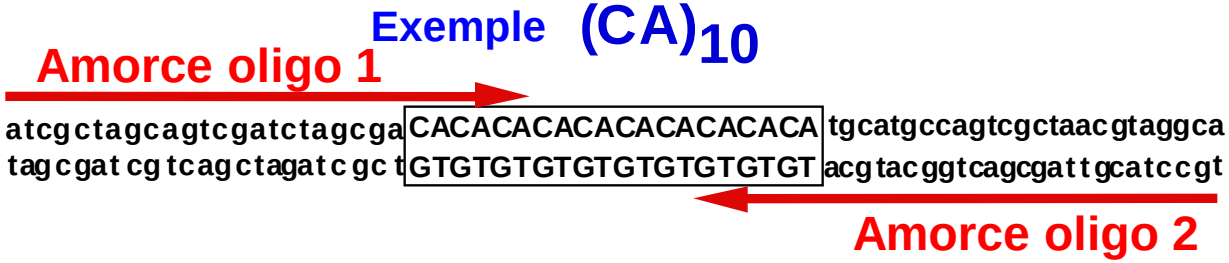
Temps de retour au clinicien = 3-4 semaines (2016).

Activité pour une douzaine des hopitaux de la région de New-York

- **2/3 pour du séquençage de génomes individuels dans des grandes cohortes (2.000-3000 personnes) pour l'identification de gènes de maladies.**
- **1/3 pour la clinique. Pour les diagnostics non établis ou pas fiables d'un enfant atteint, le séquençage sur 60 trios (parents à 60x et enfant à 100x) aboutit à un succès de 50 %.**
- **La plupart des mutations sont des mutations *de novo*.**

Les empreintes génétiques

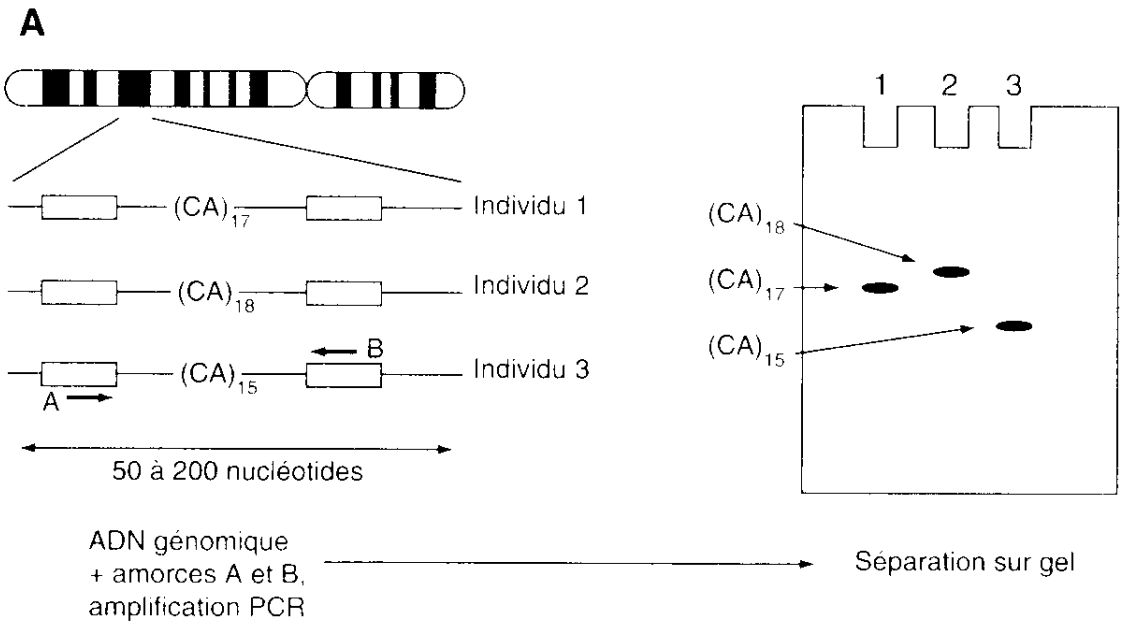
Les microsatellites sont des répétitionsen tandem de courtes séquences (de 1 à 4 nucléotides), hautement polymorphiques, fréquentes et bien dispersées le long des chromosomes et facilement amplifiables par PCR

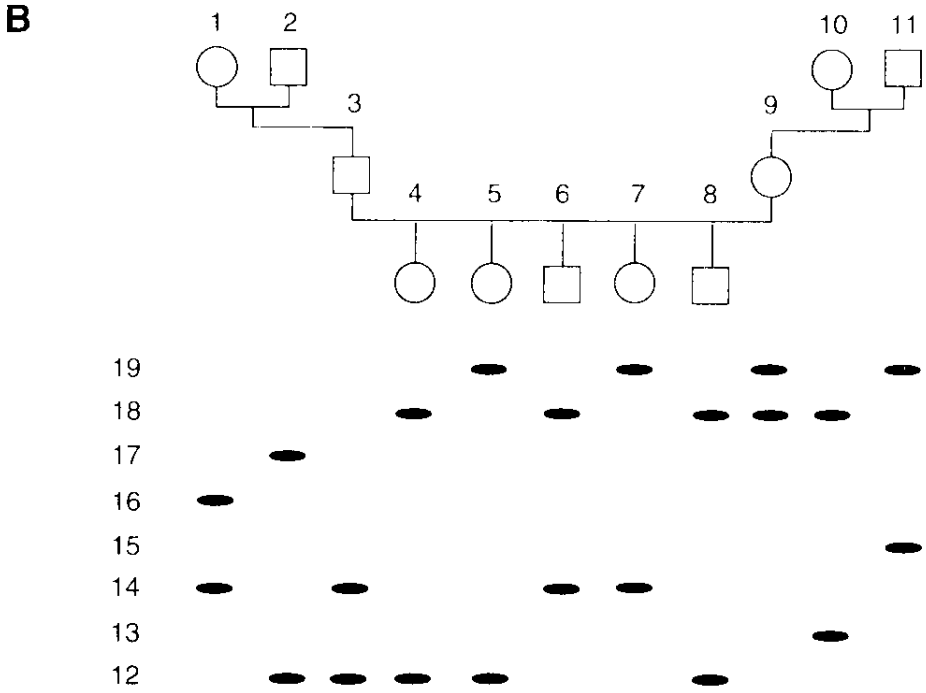


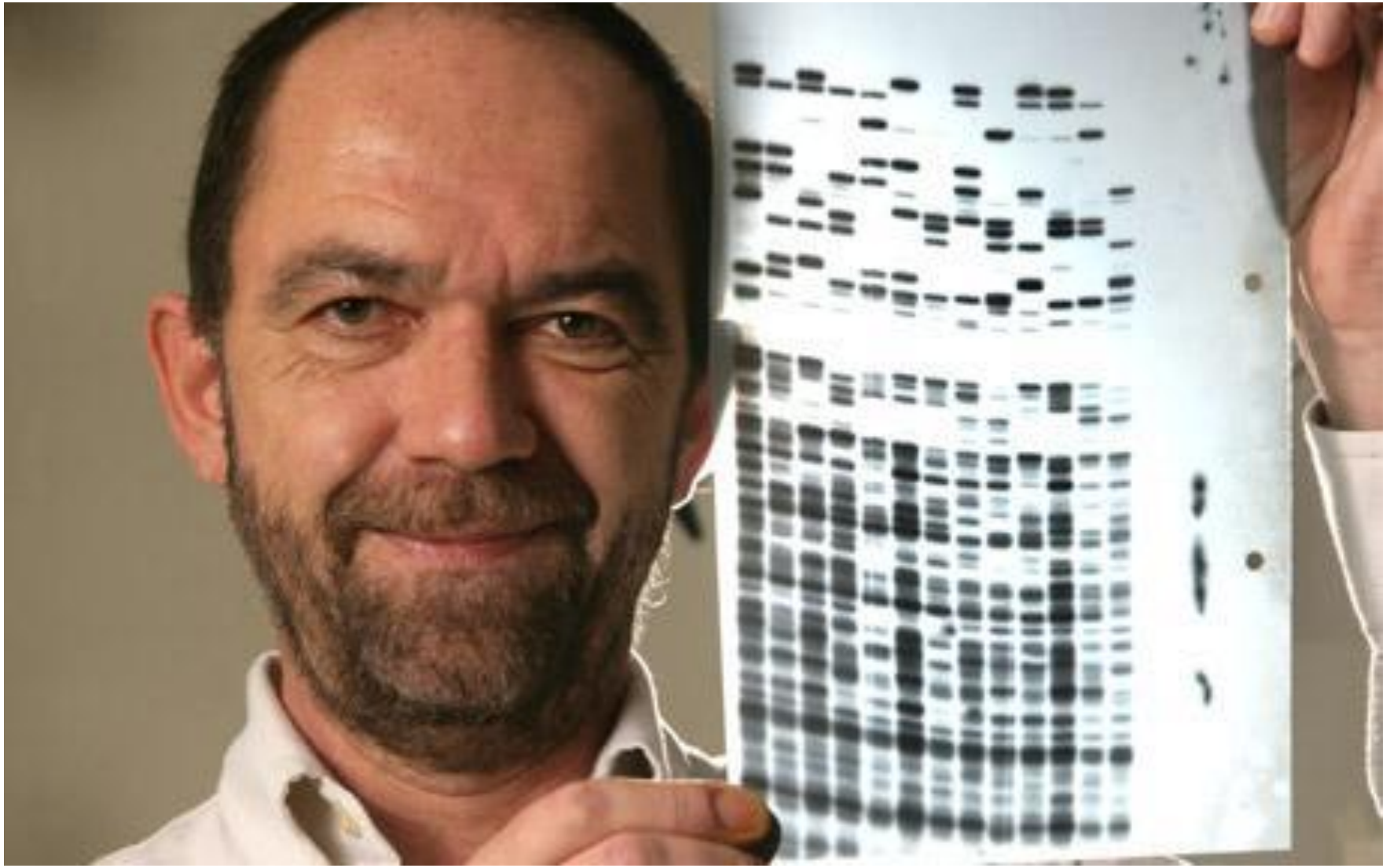
>20 répétitions: A, AC, AAAN, AAN, AG

1 Microsatellite / 6 kb

$(CA)_n$ = 50.000 à 100.000 dans le génome humain, bien dispersés


A
1 2 3
(CA)17
Individu 1
(CA)18
Individu 2
B
(CA)15
Individu 3
A
50 à 200 nucléotides
(CA)18
(CA)17
(CA)15
ADN génomique
+ amorces A et B,
amplification PCR
Séparation sur gel


B
1 2
3
4 5 6 7 8
9
10 11
19
18
17
16
15
14
13
12



Sir Alec JEFFREYS

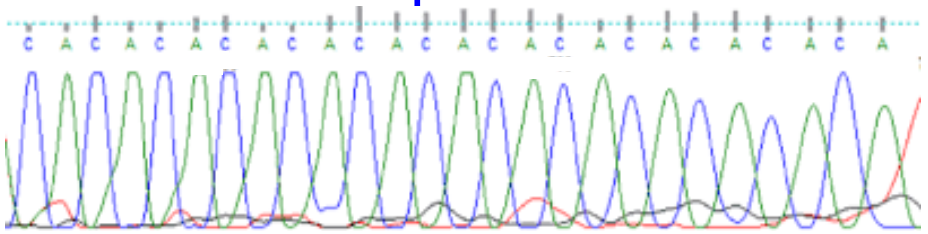
Technologie créée en 1985, commercialisation en 1987

L’ADN peut provenir de prélèvements sanguins**, de fragments de peaux, de cheveux (mais il faut le bulbe, la partie basse du cheveu qui contient le noyau), cellules provenant de desquammation naturelle (linge de corps). Il faut au moins 20 cellules.

Les services scientifiques de gendarmerie en France ont été les premiers à réaliser des empreintes génétiques à partir de sang séché!

Le test est réalisé par PCR: l’ADN est extrait et 13 aliquotes sont réparties dans 13 puits de microplaques. Chaque tube reçoit un des 13 couples d’amorces (oligos) du jeu et les réactions PCR sont menées en parallèle sur des automates PCR.

Un exemple de résultat pour un marqueur génétique à microsatellites



La probabilité que 2 individus aient la même empreinte (le même élecrophorégramme du jeu de microstallites) par hasard est de 1 chance sur 10^{18} .

Mais si on ne sait pas si les 2 individus sont apparentés ou non, la probabilité est de 1 chance sur $3x\ 10^{12}$, car la fréquence des jumeaux monozygotes* est de 0,2% dans la population mondiale.

AUX USA, le test est basé sur l'utilisation de 13 microsatellites, et seulement 10 au Royaume-Uni.

Tenant compte par ailleurs des recommandations internationales que nous évoquerons plus loin, l'arrêté du 18 juin 2000 a fixé comme suit la liste des sept loci utilisables :

- D21S11 (chromosome n° 21)
- VWA (chromosome n° 12)
- TH01 (chromosome n° 11)
- FGA (chromosome n° 4)
- D8S1179 (chromosome n° 8)
- D3S1358 (chromosome n° 3)
- D18S51 (chromosome n° 18)

Les analyses portent également sur le gène de l'amélogénine, marqueur spécifique du sexe.

Ils présentent en particulier une capacité de discrimination suffisante du fait du polymorphisme des loci. Ainsi cette capacité est-elle de l'ordre de 1 X 10-9 au sein de la population caucasienne qui, du point de vue de la génétique des populations, comprend les différentes populations européennes. En d'autres termes, les chances de voir deux individus non apparentés présenter le même profil sont de l'ordre d'une sur un milliard.

La mise en oeuvre des investigations sur un tel ensemble de loci permet par ailleurs de fournir des réponses dans un délai inférieur à 48 heures et en quelques heures pour les cas urgents.

En outre, l'ADN de ces locis est un ADN à structure répétitive dont la caractère non codant satisfait aux recommandations exprimées par la résolution du Conseil de l'Union Européenne du 9 juin 1997.

James D. WATSON & Francis H. CRICK