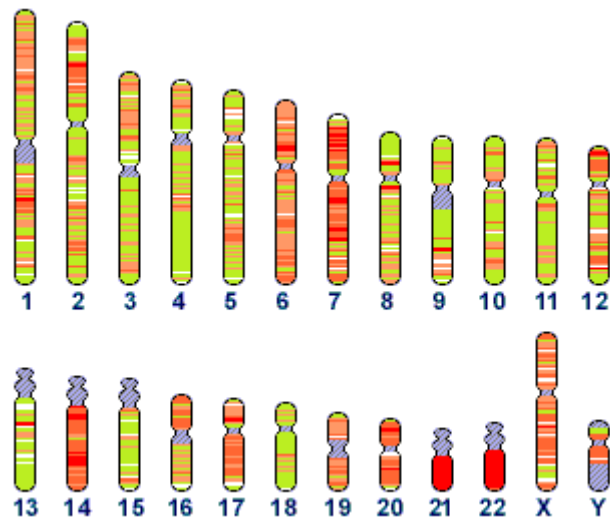


**MISE À PROFIT DU SÉQUENÇAGE DU  
GÉNOME HUMAIN POUR IDENTIFIER DE  
NOUVEAUX MINISATELLITES  
POLYMORPHES, VOIRE HYPERMUTABLES**



**France DENOEUDE**

Rapport de stage de DEA AGM2 (1999-2000) effectué à  
l'Institut de Génétique et Microbiologie d'Orsay, dans le laboratoire  
**Instabilité des Extrémités de Chromosomes Humains,**  
sous la responsabilité de **Gilles VERGNAUD**

## Résumé

Les minisatellites sont associés à de nombreux aspects intéressants de la biologie du génome humain tels que la régulation de gènes, les sites fragiles des chromosomes et l’empreinte parentale. Les connaissances sur la biologie des minisatellites ont beaucoup progressé ces dix dernières années, grâce à l’identification et à l’analyse de minisatellites hypermutables, qui sont apparus comme des biomarqueurs potentiels pour l’exposition à des agents génotoxiques comme les radiations ionisantes. Cependant, l’isolement de minisatellites d’intérêt (polymorphes ou hypermutables) était jusqu’à présent limité par les techniques expérimentales. Dans cette étude, nous avons mis à profit les données de séquençage du génome humain afin d’étudier la répartition des minisatellites (dont la définition a pu être précisée). Nous avons également pu trouver un moyen simple et efficace d’améliorer la détection de minisatellites polymorphes à partir de la séquence : il pourra être mis en œuvre dès que nous disposerons de deux allèles différents d’un même chromosome.

## Abstract

Minisatellites have been found in association with important features of human genome biology such as gene regulation, chromosomal fragile sites and imprinting. Our knowledge of minisatellites biology has greatly increased in the past ten years owing to the identification and careful analysis of human hypermutable minisatellites which have been put forward as potential biomarkers for the monitoring of genotoxic agents such as ionizing radiation. Nevertheless, identification of polymorphic or hypermutable minisatellites has been limited by experimental constraints. In this study, we take advantage of the recent whole chromosome sequence data release to provide an unifying view on minisatellites definition and distribution (which may facilitate the annotation of tandem repeats sequences) and to find an efficient method of identification of polymorphic minisatellites, based on their sequence : it will be applicable as soon as two different alleles of a human chromosome will be released.

# Table des matières

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Matériel et méthodes</b>	<b>3</b>
2.1	Elaboration d'une base de données des répétitions en tandem . . . . .	3
2.1.1	Séquences traitées . . . . .	3
2.1.2	Recherche des répétitions en tandem et importation dans la base de données . . . . .	4
2.1.3	Interrogation de la base de données par le web . . . . .	4
2.2	Etudes de population par PCR . . . . .	4
<b>3</b>	<b>Résultats et discussion</b>	<b>6</b>
3.1	Etude de la répartition des "minisatellites" sur le chromosome 22 humain, et comparaison avec d'autres génomes . . . . .	6
3.2	Tests de polymorphisme . . . . .	10
<b>4</b>	<b>Conclusion et perspectives</b>	<b>14</b>

# 1 Introduction

Les minisatellites sont des répétitions en tandem de courts motifs (de 6 à 100 paires de bases) plus ou moins bien conservés (60% à 100% d'homologie) sur quelques centaines à quelques milliers de paires de bases. Il s'agit d'intermédiaires entre les microsatellites (motif inférieur à 5 pb) et les satellites (longueur totale atteignant plusieurs Mégabases). De telles structures, bien qu'ayant tout d'abord été isolées chez l'homme, existent dans de nombreux organismes, y compris les bactéries.

Les minisatellites ont plusieurs intérêts biologiques :

- Certains font partie de phases ouvertes de lecture, ce qui peut engendrer du polymorphisme dans la population humaine
- D'autres lient des protéines avec des conséquences fonctionnelles variables (régulation transcriptionnelle, épissage, empreinte parentale...)
- Certains sites fragiles des chromosomes sont des minisatellites
- Les minisatellites polymorphes ont servi de marqueurs pour la cartographie du génome humain
- Grâce à leur localisation principalement sub-télomérique, ils ont pu servir de sondes pour détecter des micro-délétions des extrémités de chromosomes humains, impliquées par exemple dans le retard mental
- Ils sont utilisés pour l'identification d'individus (empreintes génétiques)
- Les minisatellites hypermutables (en méiose) permettent l'étude de phénomènes complexes de recombinaison, comme la conversion génique. Ils pourraient également servir de biomarqueurs d'exposition à des agents génotoxiques

Les minisatellites d'intérêt sont donc principalement les minisatellites polymorphes (également appelés VNTR : variable number of tandem repeats) et les minisatellites hypermutables (qui sont très rares mais présentent une extrême instabilité).

Les minisatellites hypermutables sont les éléments les plus instables du génome humain : par exemple, CEB1 a un taux de mutation de plus de 12% en méiose dans la lignée germinale masculine (voir Annexe 1 : description des principaux minisatellites hypermutables). On qualifie d'hypermutables les minisatellites pour lesquels on observe avec une fréquence de plus de 0.5%, des enfants ayant hérité d'un allèle de longueur différente de celles des allèles parentaux lors de l'étude de pedigrees (grandes familles du CEPH). Ces allèles mutants proviennent généralement d'événements complexes, intra-alléliques ou interalléliques du type conversion génique (Buard, 1994 ; Jeffreys, 1994), qui surviennent lors de la méiose (taux de mutations beaucoup plus élevés dans le sperme que dans le sang, où les rares allèles mutants résultent d'événements beaucoup plus simples et intra-

alléliques : Jeffreys, 1997). Plusieurs modèles ont été proposés depuis 1994 pour expliquer ces recombinaisons : ils suggéraient pour la plupart la génération de coupures double brin décalées à l'intérieur des répétitions en tandem (par exemple : Bois, 1998).

Cependant, plusieurs observations laissent supposer un rôle important de la séquence flanquante. Tout d'abord, pour la plupart des minisatellites hypermutables, on observe une polarité de localisation des mutations : les événements se produisent préférentiellement vers une extrémité du minisatellite (Jeffreys, 1994). D'autre part, une transversion de G en C dans la flanquante du minisatellite MS32 est associée à une forte réduction du taux de mutation : cette suppression agit en *cis* mais n'affecte pas la capacité de l'allèle à jouer le rôle de donneur lors de la conversion génique (Monckton, 1994). Enfin, MS32 est associé à un point chaud de crossing over, ce qui suggère que l'instabilité des minisatellites hypermutables pourrait être une conséquence de la proximité de points chauds de recombinaison méiotique (Jeffreys, 1998).

En 1999, il a été montré que si on insérait un minisatellite hypermutable humain dans la levure *Saccharomyces cerevisiae* à proximité d'un point chaud de recombinaison, on observait une instabilité du minisatellite dépendant de la topoisomérase méiotique *spo11* qui génère des cassures double brin franches (Debrauwère, 1999). Nous avons donc proposé un nouveau modèle d'instabilité des minisatellites hypermutables, (Vergnaud & Denoeud, 2000) : voir annexe 2. Il permet de rendre compte de la plupart des événements observés chez l'homme en initiant le mécanisme par une cassure double brin franche dans la flanquante du minisatellite.

L'idée actuelle que les minisatellites hypermutables ne sont pas intrinsèquement instables, mais résultent de la coïncidence entre un point chaud de recombinaison méiotique et un minisatellite "normal" (qui sont fréquents) promet d'intéressantes perspectives : la détection de minisatellites hypermutables pourra être associée à la mise en évidence et à l'analyse des points chauds de recombinaison du génome humain, qui restent pour l'instant encore très peu étudiés. Par exemple, une étude à grande échelle de la distribution des minisatellites hypermutables dans le génome permettrait d'appréhender la distribution des points chauds de cassures double brin.

Le stage effectué a pour but de tirer parti du séquençage du génome humain pour isoler des minisatellites d'intérêt : polymorphes en premier lieu, puis éventuellement hypermutables (les minisatellites hypermutables sont en effet extrêmement polymorphes). Les méthodes expérimentales (hybridation de sondes sur des Southern Blots : Vergnaud, 1989...) développées jusqu'alors ne permettaient d'isoler que les minisatellites dont la longueur totale excédait 800 pb environ, et dont le motif répété était peu complexe. Ces

séquences présentent un fort biais de répartition vers les extrémités télomériques. Une étude plus systématique des répétitions en tandem visant à mieux définir les minisatellites, sans *a priori* dû aux contraintes expérimentales, est maintenant rendue possible par les progrès du séquençage du génome humain : nous pouvons dorénavant isoler directement les répétitions en tandem à partir de la séquence, et étudier leur distribution à l'échelle d'un chromosome entier (le séquençage des chromosomes 21 et 22 est en effet achevé). Après avoir, par ce biais, isolé les minisatellites (dont on aura pu préciser la définition qui est encore peu claire et donc peu employée : par exemple, dans l'article traitant de la séquence du chromosome 22 humain (Dunham, 1999), les minisatellites ne sont pas annotés comme tels), il faudra être capable de repérer parmi eux les minisatellites polymorphes. Une étude menée sur un ensemble de minisatellites du chromosome 22 nous permettra d'estimer le polymorphisme moyen et d'en déduire des méthodes visant à détecter préférentiellement les minisatellites polymorphes.

## 2 Matériel et méthodes

### 2.1 Elaboration d'une base de données des répétitions en tandem

Afin de faciliter l'accès aux données concernant les minisatellites du chromosome 22 humain (et de chromosomes entièrement séquencés dans d'autres espèces), j'ai élaboré une base de données des répétitions en tandem interrogeable par le web sur le site <http://minisatellites.u-psud.fr>.

#### 2.1.1 Séquences traitées

Nous avons choisi de traiter des chromosomes dont le séquençage était achevé, afin de ne pas avoir à remettre la base de données à jour au fur et à mesure de l'avancée du séquençage : c'est pourquoi nous ne disposons actuellement que du chromosome 22 humain : Dunham, 1999 (le chromosome 21 est en cours d'importation). Afin de pouvoir effectuer des comparaisons avec d'autres organismes, nous avons également traité entre autres le chromosome 4 d'*Arabidopsis thaliana* (Mayer, 1999) et le chromosome 1 de *Caenorhabditis elegans* (*C. elegans* sequencing consortium, 1998). Les contigs traités pour ces trois chromosomes sont ceux qui étaient disponibles au mois de janvier 2000 :

- 12 contigs pour le chromosome 22 humain totalisant 34,6 Mb (leur liste est accessible sur le site <http://www.ncbi.nlm.nih.gov/genome/seq/> : ces séquences ont très peu évolué depuis leur importation)
- 2 contigs pour le chromosome 4 d'*A. thaliana* totalisant 17,8 Mb (<http://www.mips.biochem.mpg.de/proj/thal/>)

- 1 fichier pour le chromosome 1 de *C elegans* : 12,7 Mb  
(<http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/framik?db=Genome&gi=62>)

### 2.1.2 Recherche des répétitions en tandem et importation dans la base de données

Les contigs rapatriés ont été soumis au logiciel Tandem Repeats Finder (Benson, 1999), accessible sur le site <http://c3.biomath.mssm.edu/trf.html> qui identifie les répétitions en tandem et fournit différentes informations (longueur du motif, nombre de répétitions, lien vers un fichier d'alignement...) qui ont été importées dans une base de données créée sous Access, après un traitement préalable des tableaux de sortie par une macro Excel. (En effet, les résultats proposés par le TRF présentent souvent une redondance : pour un même locus, plusieurs consensus peuvent être proposés -par exemple :  $15 \times 20pb$  et  $30 \times 10pb$ -. Cette redondance a été éliminée en utilisant les critères suivants : parmi les consensus associés au meilleur taux d'homologie entre les motifs (+/- 5%), seul celui correspondant au plus petit motif répété, et à la plus grande longueur totale a été conservé). Les fichiers d'alignement fournis par le logiciel TRF, très volumineux donc inexploitable par le web, ont été scindés en autant de fichiers que d'alignements (c'est à dire un fichier d'alignement par répétition en tandem détectée) grâce à un programme codé en Perl, fourni par Mourad Sahbatou (CEPH).

### 2.1.3 Interrogation de la base de données par le web

La base de données des répétitions en tandem dans des chromosomes entièrement séquencés est accessible sur le site <http://minisatellites.u-psud.fr>. Son interrogation s'effectue grâce à la technologie ASP (Active Server Pages), qui permet la création de pages Web dynamiques et interactives issues de l'exécution de scripts par le serveur (le programme a été codé en VB script).

L'annexe 3 présente la page de soumission des requêtes ainsi que les résultats générés pour une requête "répartition" (visualisation de la répartition des répétitions en tandem le long d'un chromosome) et une requête "isolement" (obtention de toutes les caractéristiques des répétitions en tandem isolées) effectuées sur le chromosome 22 humain.

## 2.2 Etudes de population par PCR

Les minisatellites peuvent être polymorphes par la composition interne de leurs motifs ou par leur nombre d'unités répétées : nous ne nous intéressons ici qu'au polymorphisme de nombre de répétitions. Pour le mesurer, nous avons étudié par PCR (polymerase chain reaction) 96 individus pour la plupart non apparentés : parents de familles du CEPH

et quelques grands-parents afin de vérifier la ségrégation des allèles, ce qui permettra d'éliminer les éventuels artéfacts.

Les amorces ont été choisies à partir des 200 nucléotides de flanquante proposés par le logiciel **Tandem Repeats Finder** (Benson, 1999). Nous avons sélectionné des oligonucléotides d'environ 30 pb, grâce au logiciel **Primer3** disponible sur le site [http://www-genome.wi.mit.edu/cgi-bin/primer/primer3\\_www.cgi](http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi), et leur spécificité a été vérifiée par confrontation (Blasts) avec différentes banques de séquences : **genbank nr**, **htgs**, **REPBASE**, **REPEATS...**

Des tests ont permis de choisir pour chaque couple d'amorces :

- la machine PCR la plus efficace (**MJresearch PTC 200** ou **Perkin Elmer 9600** : ces machines comptent 96 loges)
- la température d'annealing (hybridation des amorces) optimale (la machine **PTC200** peut produire des gradients de températures, ce qui évite de multiplier les PCRs)
- la Taq polymérase la mieux appropriée (Taq Boehringer associée à Pwo Boehringer ou Taq Appligene)

Chaque PCR a été effectuée pour 96 tubes (individus) simultanément, dans 12 barrettes de 8 tubes.

La composition du mélange réactionnel dans chaque tube (15  $\mu L$ ) est la suivante :

- ADN : environ 50 ng (selon la quantité indiquée par le CEPH), déposés dans un volume de 5  $\mu L$  grâce à un appareil de distribution automatique à 96 aiguilles (**Robbins**)
- nucléotides tri-phosphate : concentration finale de 200  $\mu M$  chacun
- amorces : concentration finale de 0,6  $\mu M$  chacune
- Taq (**Appligene** ou **Boehringer**) : 0,5 U
- éventuellement Pwo (**Boehringer**, associée à la Taq **Boehringer**) : 0,05 U
- Tampon **Boehringer** ou **Appligene** selon la Taq, contenant 1,75 mM de  $Mg^{++}$  final

Le mélange (Taq, tampon, dNTP, amorces) a été réparti à raison de 10  $\mu L$  par tube dans les 96 tubes contenant déjà 5  $\mu L$  de solution d'ADN.

Après l'amplification des minisatellites, les échantillons ont été soumis à une électrophorèse sur un gel d'agarose **SIGMA** à 1%, dans un tampon de migration **TBE 0.5X**, à un voltage de 8V/cm pendant 1 heure 30 environ. Les gels ont ensuite été colorés au **BET** et visualisés sous les UV (l'annexe 4 présente un exemple de résultat obtenu).



## 3 Résultats et discussion

### 3.1 Etude de la répartition des “minisatellites” sur le chromosome 22 humain, et comparaison avec d’autres génomes

(voir Vergnaud & Denoeud, 2000, paragraphe "Taking advantage of the global view provided by large scale sequencing")

Les minisatellites isolés jusqu’à présent (en général grâce à leurs propriétés de cross-hybridation avec d’autres répétitions en tandem : Vergnaud, 1989) par l’intermédiaire de Southern Blots, avaient nécessairement une longueur compatible avec cette technique, soit plus de 800 pb environ. De tels minisatellites ont une fréquence similaire dans cinq génomes de mammifères explorés, mais leur distribution diffère beaucoup d’un génome à l’autre : chez l’homme, on observe un fort biais de répartition vers les extrémités de chromosomes, moins accentué chez le porc, et inexistant chez la souris et le rat. En fait, les relations de synténie entre ces espèces suggèrent que chez l’ancêtre commun, les minisatellites avaient une localisation télomérique (Amarger, 1998). Les répétitions en tandem qui peuvent être analysées par Southern Blot (>800 pb) sont donc majoritairement associées aux extrémités de chromosomes, et les régions interstitielles qui en sont riches sont probablement le résultat de réarrangements secondaires, comme la fusion de télomères. Cependant, ce type de "minisatellites" ne représente qu’une faible proportion des répétitions en tandem, dont la plupart ont une longueur totale inférieure à la limite arbitraire, adoptée jusqu’à présent, de 500 pb sans appartenir pour autant à la classe des microsatellites. La définition comme quoi les minisatellites doivent excéder 500 pb est trop restrictive pour au moins deux raisons :

- des répétitions en tandem qui sont clairement qualifiées de minisatellites ont souvent des allèles de moins de 500 pb dans la population
- lors de l’assemblage des données de séquence, la véritable longueur des minisatellites n’est pas toujours convenablement déterminée, surtout quand les motifs répétés sont fortement conservés

CEB1, le minisatellite le plus hypermutable caractérisé jusqu’alors (Vergnaud, 1991) illustre ces deux lacunes de la définition actuelle.

Dorénavant, les répétitions en tandem peuvent être isolées directement à partir des données du séquençage, et leur répartition peut être étudiée à l’échelle de chromosomes entiers. Nous avons donc créé une base de données des répétitions en tandem de plusieurs chromosomes entièrement séquencés (voir matériel et méthodes : paragraphe 2.1).

La figure 1 présente une étude de la répartition des répétitions en tandem du chromosome 22 humain. De la gauche vers la droite, sont représentés : la position du centromère

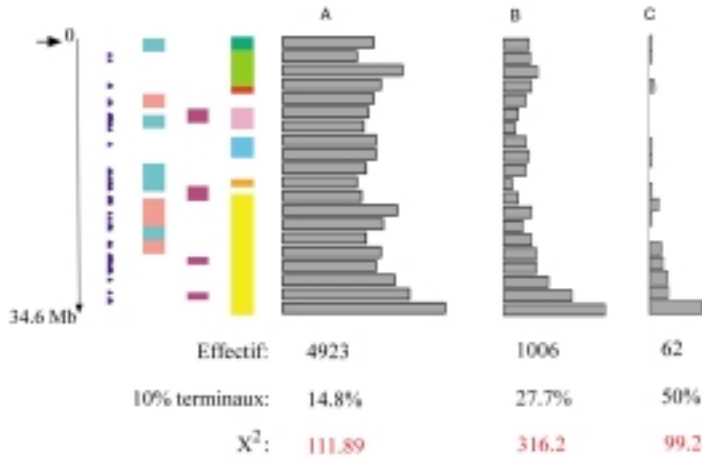


FIG. 1 – Répartition des répétitions en tandem dans le chr22

(flèche : le chromosome 22 est acrocentrique), la position des 51 (sur 67) marqueurs génethon identifiés par le logiciel TRF, les régions riches (rose) et pauvres (vert) en GC, les régions de recombinaison accrue (violet), et les régions de synténie avec la souris correspondant aux indications de Dunham *et al.*, 1999. Les histogrammes présentent les répartitions obtenues pour différentes requêtes : A :  $U \geq 6$ ,  $N \geq 3$ ; B :  $U > 16$ ,  $N \geq 3$ ,  $L \geq 100$ ; C :  $U > 16$ ,  $N \geq 3$ ,  $\%GC \geq 65\%$ ,  $B_{GC} \geq 0.3$ ,  $\%M \geq 85\%$  (avec  $U$ =longueur du motif,  $N$ = nombre de répétitions,  $L$ = longueur totale,  $B_{GC}$ = biais G/C ( $= \frac{|\%G-\%C|}{\%G+\%C}$ ),  $\%M$ = pourcentage d'homologie,  $\%GC$ = pourcentage en G+C). Le chromosome 22 été fragmenté en 20 segments de longueur égale. Les pourcentages rapportés sous les histogrammes correspondent à la proportion d'objets dans les deux derniers segments, qui comprennent 3.46 Mb soit 10% de la longueur totale. Les valeurs de  $\chi^2$  ont été calculées en comparant le nombre de répétitions en tandem dans les 10% terminaux du chromosome et le nombre moyen de répétitions en tandem dans 10% du chromosome. Le seuil de significativité est de 3.841 avec  $P = 0.5\%$  (les valeurs sont en rouge car elles excèdent ce seuil).

On constate qu'un tiers des 14000 répétitions en tandem du chromosome 22 satisfait la définition la plus large d'un minisatellite (motif d'au moins 6 pb répété au moins 3 fois) : figure 1A. Ces répétitions présentent déjà un biais significatif de répartition vers le télomère. Parmi elles, les minisatellites ayant des unités répétées de plus de 16 pb et une longueur totale supérieure à 100 pb montrent une distribution plus fortement biaisée vers le télomère : fig 1B (les répétitions en tandem de motifs de moins de 17 pb, et de longueur totale inférieure à 100 pb montrent quant à elles une répartition homogène : données

non montrées). La comparaison des unités répétées dans les minisatellites classiques a conduit à la notion de “core sequences” montrant des similitudes avec la séquence *Chi* du phage  $\lambda$  (GCTGTGG). Ces observations peuvent être résumées comme suit : la plupart des minisatellites classiques sont riches en GC, avec une forte asymétrie G/C entre les brins. La figure 1C montre le résultat d’une requête mimant ces caractéristiques (requête B + contenu riche en GC, fort biais G/C, et forte homogénéité interne). La moitié des 62 minisatellites correspondant à ces critères sont localisés dans les 10% terminaux du chromosome 22.

Ces requêtes simples démontrent qu’une fraction des répétitions en tandem, comprenant des centaines de locus sur le chromosome 22, se comporte comme les minisatellites classiques, présents majoritairement dans la bande R terminale des chromosomes humains (Amarger, 1998). Par contre, on ne trouve pas de corrélation apparente entre la distribution des minisatellites et les frontières de synténie avec la souris, la richesse en GC ou les régions de recombinaison accrue.

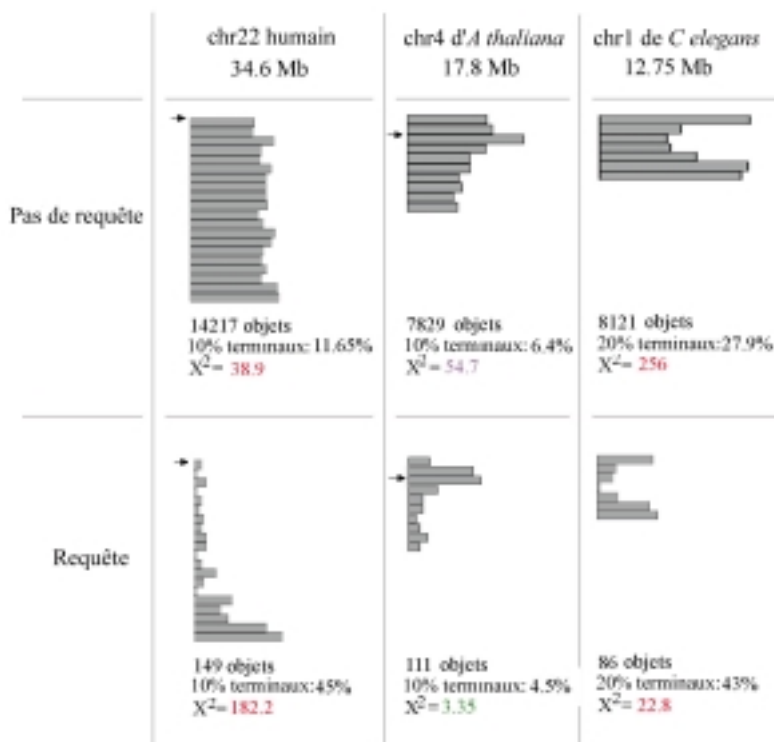


FIG. 2 – Comparaison de la répartition des répétitions en tandem dans trois organismes

La figure 2 présente une étude comparative entre le chromosome 22 humain (34.6 Mb, acrocentrique), le chromosome 4 d’*A thaliana* (17.8 Mb, acrocentrique) et le chromosome

1 de *C elegans* (12.75 Mb). Les flèches montrent la position du centromère (inconnue pour le chromosome 1 de *C elegans*). La requête est la suivante :  $U \geq 10$ ,  $N \geq 3$ ,  $L \geq 100$ ,  $0.3 \geq B_{GC} \geq 0.55$ ,  $\%M \geq 70\%$ . Les chromosomes ont été fragmentés en segments de longueur comparable : 1.73 Mb ( $\times 20$ ) pour le chromosome 22 humain, 1.78Mb ( $\times 10$ ) pour le chromosome 4 d'*A thaliana* et 1.82 Mb( $\times 7$ ) pour le chromosome 1 de *C elegans* respectivement. Les tailles relatives des chromosomes sont respectées dans la figure. Les valeurs de  $\chi^2$  sont en rouge quand on a un excès d'objets significatif vers les télomères, en violet quand on a un défaut d'objets significatif vers les télomères, et en vert quand il n'y a pas de différence significative entre les télomères et le reste du chromosome.

Le nombre total de répétitions en tandem identifiées par le logiciel TRF ("Pas de requête") n'est pas proportionnel à la longueur du chromosome : il est nettement plus élevé chez le nématode (637/Mb) que chez l'homme (415/Mb) ou la plante (445/Mb). La figure 2, ligne du bas, montre le résultat d'une requête représentative. Le nombre de minisatellites isolés est similaire dans les trois espèces, si on prend en compte la longueur relative des chromosomes.

Un fort biais télomérique est observé pour le chromosome 1 de *C elegans*, comme pour le chromosome 22 humain. Par contre, la répartition des minisatellites chez *A thaliana* est très différente de celle des deux autres génomes : les répétitions en tandem sont réparties principalement autour du centromère. Cette différence peut s'expliquer par le fait que chez l'homme, les extrémités de chromosomes ont un défaut en nodules de recombinaison par rapport à leur fort taux de recombinaison, tandis que chez les plantes, la distribution des nodules de recombinaison est corrélée à la fréquence de recombinaison le long des chromosomes (Ashley, 1994). Les minisatellites télomériques sont des candidats pour expliquer le paradoxe observé chez l'homme : des mécanismes spécifiques seraient activés en méiose mâle et les minisatellites seraient impliqués dans l'appariement des chromosomes, soit directement, soit par l'intermédiaire de protéines d'appariement (Sybenga, 1999). La comparaison de ces différents génomes a donc permis de confirmer la particularité des plantes par rapport à l'homme (et au nématode), qui pourrait correspondre à un rôle différent des minisatellites dans les mécanismes de recombinaison (absence de rôle chez la plante...).

Nous avons aussi étudié en détail la variation du biais de répartition télomérique (rapport entre les nombres de répétitions en tandem télomériques et non télomériques) en fonction de la longueur du motif répété et de la longueur totale pour les trois espèces : voir figure 3. (En ordonnée :  $\frac{\text{effectif } 10\% \text{ terminaux ("T")}-\text{effectif autres regions ramene a } 10\% \text{ ("A": autres)}}{A} \approx$  Z-score. En abscisse : plages de longueur d'unité répétée ou de longueur totale).

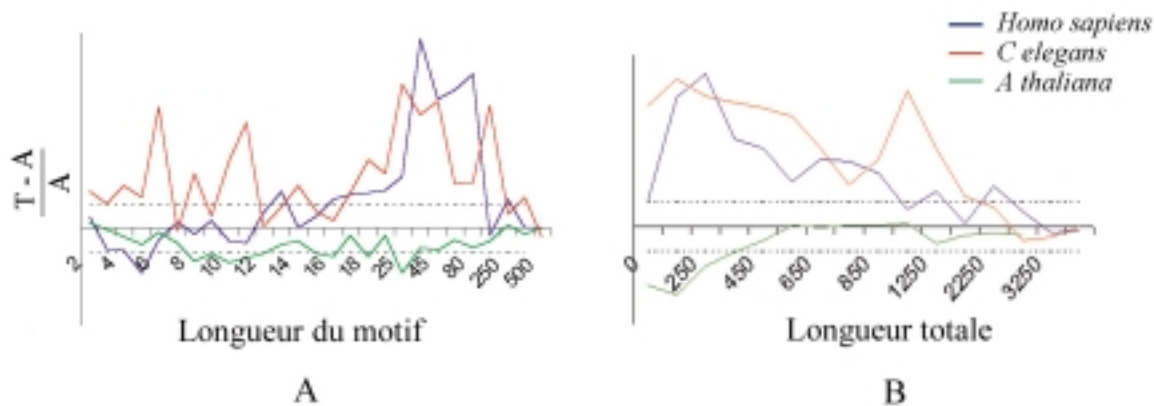


FIG. 3 – Mesure du biais selon la longueur du motif ou la longueur totale

*C. elegans* montre un biais télomérique pour toutes les longueurs d'unités répétées, même courtes : en particulier 6 et 12 pb, ce qui est dû à la présence de nombreuses répétitions en tandem de type télomérique (TTAGGC)<sub>n</sub> (The *C. elegans* Sequencing Consortium, 1998). En revanche, le biais télomérique n'apparaît chez l'homme que pour des unités répétées de longueur supérieure ou égale à 17 pb. De façon intéressante, 16 pb est justement la limite au delà de laquelle les mécanismes de réparation des mésappariements ne peuvent plus réparer les boucles d'ADN chez la levure (Sia, 1997).

Contrairement à *C. elegans*, le biais télomérique n'apparaît pour le chromosome 22 humain que pour des longueurs totales supérieures à 120-140 pb : sur le graphique 3B, on remarque que cette transition se situe entre 100 et 200 pb (premier point : 0 à 100 pb, deuxième point : 100 à 200 pb), mais une étude plus fine a permis de la déterminer avec précision (données non montrées). Ce seuil évoque l'instabilité des triplets observée au delà de 40-50 répétitions dans de nombreuses maladies neurodégénératives.

Pour le chromosome 4 d'*Arabidopsis thaliana*, ces diagrammes confirment bien un défaut de répétitions en tandem dans les régions télomériques.

### 3.2 Tests de polymorphisme

Afin d'avoir une idée du nombre de minisatellites polymorphes parmi l'ensemble des minisatellites, nous avons typé par PCR (voir paragraphe 2.2) un certain nombre de minisatellites du chromosome 22, choisis plus ou moins arbitrairement : certains correspondent à des requêtes mimant la répartition des minisatellites traditionnellement étudiés, et d'autres nous ont été suggérés par Gary Benson (auteur du logiciel *Tandem Repeats Finder*), avec qui nous collaborons, afin de tester son hypothèse sur la prédiction du polymorphisme.

Parmi les minisatellites suggérés par Gary Benson, certains n'étaient pas uniques : plusieurs minisatellites partageaient des séquences flanquantes et un motif répété similaires, mais le nombre de copies de ce motif différait. Il s'est avéré que l'une de ces "familles" appartenait à un rétroposon (SVA) contenant un nombre variable de copies d'une répétition en tandem de 40 pb : les amorces PCR ont été choisies à l'extérieur de la séquence du rétroposon. Une autre correspondait probablement à une région LCR (low copy repeats) du chromosome 22 : il a été impossible de choisir des amorces spécifiques pour ces minisatellites.

On peut dire qu'il existe des "familles" de minisatellites, si on considère simplement que ce type séquences a autant de chances d'être dupliqué et remanié au sein du génome que les autres. Cependant, ce terme a souvent été mal utilisé à cause des premiers moyens d'isolement des minisatellites : des sondes constituées de répétitions en tandem d'une quinzaine de paires de bases permettaient de détecter de nombreux minisatellites (Vergnaud, 1989), qu'on avait qualifiés comme appartenant à des "familles". En fait, ils n'avaient pas de véritable homologie de séquence, mais seulement quelques caractéristiques communes (richesse en GC...). Il n'existe pas à ce jour de "familles" de minisatellites comptant beaucoup de membres, mais l'étude systématique du génome humain rendue possible par le séquençage pourrait permettre d'en découvrir.

Les résultats des typages par PCR sont récapitulés dans le tableau 1.

Les PCR ont tout d'abord permis de vérifier que les tailles des allèles attendues d'après la séquence du chromosome 22 correspondent dans la grande majorité des cas (31/33) à celles de l'un des allèles amplifiés par PCR (les allèles correspondants sont marqués d'une \* dans le tableau 1). On peut en conclure que le séquençage a été efficace dans cette gamme de taille (ce qui n'était pas sûr, étant donné qu'une trop forte homologie entre les motifs peut générer des problèmes d'assemblage, c'est à dire la perte de certains motifs).

Dans l'un des cas d'incohérence entre l'allèle amplifié par PCR et la séquence du chromosome 22, l'explication la plus probable est que nous avons amplifié une autre séquence que celle du chromosome 22. En effet, par un **Blast** dans la base **htgs**, nous avons identifié dans le chromosome 18 une séquence homologue au minisatellite du chromosome 22 (qui comporte 28 répétitions d'un motif de 62 pb), mais ne comptant qu'un exemplaire de l'unité répétée suivi des 8 premiers nucléotides de l'unité, c'est à dire une région unique de 54 pb flanquée d'une répétition directe de 8 pb (voir annexe 5). Cette séquence correspond à la taille effectivement amplifiée par PCR. Les régions homologues du chromosome 18 et du chromosome 22 résultent certainement d'une duplication ancestrale, mais on peut se demander si celle du chromosome 18 provient d'une délétion des motifs répétés ou si celle

TAB. 1 – Résultats obtenus pour 30 minisatellites typés

Nombre d'allèles	Nombre de minisatellites	Détails	Proba 2 allèles différents
1	8	6/8 : tailles correspondant à la séquence du chr22 2/8 : tailles ne correspondant pas à la séquence du chr22	0
2 à 8	16	<b>Fréquences alléliques :</b> 7%, 93%* 5.5%, 94.5%* 35%, 65%* 50%, 50%* 75%, 25%* 95%*, 4%, 1% 89%*, 10%, 1% 11%, 4%, 85%* 96.5%*, 3%, 0.5% 35.5%, 0.5%, 64%* 13%, 65%*, 21%, 1% 13%, 47%*, 1%, 39% 39%, 56.5%*, 4%, 0.5% 15.5%, 14%, 57.5%*, 10%, 3% 4%, 40%, 0.5%, 4%, 30.5%*, 0.5% 0.75%, 67.5%*, 8%, 10%, 1.5%, 0.75%, 10%, 1.5%	0.13 0.104 0.455 0.5 0.375 0.096 0.198 0.264 0.068 0.464 0.516 0.61 0.527 0.615 0.688 0.517
>20	2	fréquences alléliques entre 0.5% et 13%	0.95~
échecs d'amplification	7	2/7 : primers non spécifiques 5/7 : pas d'amplification (problèmes de mise au point ?)	

du chromosome 22 provient d'une amplification du motif initial : ce cas semble le plus vraisemblable car une hypothèse sur l'origine des minisatellites est que de courtes (jusqu'à 10 pb) répétitions directes apparues par hasard (événement bien plus probable que l'apparition d'une répétition en tandem de plusieurs dizaines de paires de bases) seraient à l'origine de la duplication de la région qu'elles entourent, par glissement de réplication ou échanges non réciproques. Ensuite, l'expansion du minisatellite pourrait se produire par conversion génique, crossing over non réciproque ou glissement de réplication (Haber, 1997) : voir annexe 5. Ce modèle nécessite toutefois qu'après l'amplification, le minisatellite généré soit toujours flanqué de la répétition directe de quelques pb, ce qui n'est pas le cas sur le chromosome 22 : nous avons 28 motifs stricts. Une étude plus approfondie doit encore être menée pour expliquer ces observations : nous allons tenter de reconstruire l'histoire du minisatellite du chromosome 22 (Benson & Dong , 1999) afin de vérifier si l'hypothèse d'une amplification à partir de la séquence homologue du chromosome 18 est acceptable.

Pour le deuxième minisatellite dont l'allèle amplifié ne correspondait pas à la séquence du chromosome 22, aucune séquence homologue n'a été mise en évidence par **Blast** dans la base htgs : soit il existe une région homologue qui n'est pas encore séquencée, soit le

séquençage a été imparfait et a conduit à la perte de certaines unités répétées (l'allèle déduit de la séquence du chromosome 22 fait 300 pb de moins que l'allèle amplifié par PCR).

Environ 20% (7/33) des minisatellites testés ont résisté à l'amplification par PCR, soit parce que les primers ont été mal choisis (non spécifiques : hybridation croisée...), soit parce que les conditions optimales de PCR n'ont pas été trouvées. 69% (18/26) des minisatellites amplifiés ont plus d'un allèle : tous vérifient l'équilibre de Hardy Weinberg (si les fréquences alléliques de deux allèles sont  $p$  et  $q$ , les fréquences des homozygotes seront  $p^2$  et  $q^2$  et la fréquence d'hétérozygotes sera  $2pq$ ). Parmi les minisatellites amplifiés, 8% (2/26) sont très polymorphes et candidats pour être hypermutables. Le rendement global n'est que de 6% (2/33) : le besoin de détecter ces minisatellites dès l'analyse des séquences, sans effectuer les PCR "inutiles" est donc justifié.

La société CELERA ([www.celera.com](http://www.celera.com)), qui mène un projet de séquençage du génome humain, devrait fournir d'ici quelques mois une nouvelle séquence du chromosome 22, ce qui permettra d'améliorer nettement le crible pour la recherche de minisatellites très polymorphes. Dans le tableau 1, figurent les probabilités d'obtenir deux allèles différents, tirées des fréquences alléliques de chacun des 26 minisatellites amplifiés. Si nous avons disposé de deux séquences du chromosome 22 et utilisé un crible très simple visant à ne tester par PCR que les minisatellites dont les longueurs totales étaient différentes entre ces deux séquences, nous aurions eu 95% de chances de détecter chaque minisatellite très polymorphe. Parmi les 26 minisatellites amplifiés, nous en aurions typé 8 (somme des probas d'obtenir deux allèles différents) en moyenne, dont  $2 \times 0.95$  très polymorphes (24%). Il semble donc que cette méthode soit susceptible d'améliorer sensiblement le rendement d'obtention de minisatellites fortement polymorphes, c'est à dire potentiellement hypermutables.

Toutefois, d'autres méthodes sont envisageables pour identifier les minisatellites polymorphes, comme l'utilisation du pouvoir prédictif d'un seul allèle sur le polymorphisme : Il s'agirait d'étudier la succession des variants (positions divergentes) entre les motifs répétés. En effet, les minisatellites les plus instables ne sont pas forcément caractérisés par une homogénéité de séquence entre leurs motifs, ce qui les différencie des microsatellites (qui peuvent muter par glissement des fragments d'Okasaki lors de la réplication). Cependant, ils ont probablement des caractéristiques communes : par exemple, on devrait trouver moins de types différents d'unités répétées au sein d'un minisatellite hypermutable, soumis à des remaniements constants qu'au sein d'un minisatellite inactif, dont tous les motifs auront pu diverger indépendamment : ces pistes restent à exploiter, et font l'objet d'une collaboration avec Gary Benson.



## 4 Conclusion et perspectives

Précédemment, le nombre de minisatellites polymorphes classiques était estimé à quelques milliers dans tout le génome humain, soit quelques dizaines pour le chromosome 22. Des objets si rares ne semblaient pas pouvoir jouer un rôle dans le métabolisme du génome. L'analyse des séquences de chromosomes entiers a révélé un nombre bien plus important de minisatellites avec des unités répétées et une répartition biaisée vers les extrémités télomériques, comme pour les minisatellites classiques, mais avec des longueurs totales parfois bien plus courtes que celles qui étaient imposées par les contraintes expérimentales. Ces observations donnent bien plus de crédibilité aux minisatellites. Il sera intéressant de comparer les premières données de répartition avec celles de chromosomes humains plus grands, quand le séquençage sera achevé.

Pour identifier efficacement des minisatellites fortement polymorphes (c'est à dire potentiellement hypermutables) parmi le vaste ensemble de minisatellites du génome humain, des méthodologies sont à mettre en œuvre. Nous avons pu montrer que le fait de disposer d'un deuxième allèle du chromosome 22 et de ne considérer que les minisatellites de longueur différente dans ces deux séquences s'avérerait très efficace (malgré une probabilité non négligeable de manquer des minisatellites d'intérêt). Il reste aussi à exploiter la piste consistant à comparer la distribution des positions variantes au sein des motifs qui, même si elle n'améliorait pas très fortement la détection de minisatellites polymorphes, pourrait être informative sur l'origine et l'évolution des minisatellites ainsi que leurs mécanismes d'instabilité.

La génotoxicité est un domaine de recherche prometteur, lié aux minisatellites. En effet, de nombreuses études montrent que les minisatellites hypermutables pourraient servir de biomarqueurs pour l'exposition à des agents génotoxiques, comme les radiations ionisantes. Chez la souris, on a montré une sensibilité des minisatellites à ces radiations (Dubrova, 1998), ainsi que chez l'homme grâce à une étude menée sur les habitants de régions contaminés par l'accident de Chernobyl (Dubrova, 1997). Des produits chimiques (pollution...) sont aussi suspectés d'augmenter le taux de mutation (en méiose) des minisatellites hypermutables.

Des modèles ont été récemment développés chez la levure (insertion de minisatellites hypermutables humains dans différents fonds génétiques), et pourraient permettre (à l'occasion d'une thèse...) de développer des tests de génotoxicité identifiant spécifiquement des activités recombinogènes. De telles études pourraient également aider à comprendre quelle est la cible de l'agent génotoxique : la répétition en tandem elle-même, le point

chaud de recombinaison flanquant, ou la machinerie de réplication... Dans la deuxième hypothèse, les minisatellites agiraient comme des marqueurs de l'activité du point chaud de recombinaison associé, tandis que dans la première (et peut-être aussi la troisième), n'importe quel minisatellite (même non hypermutable) pourrait servir de biomarqueur pour l'agent génotoxique.

# Bibliographie

Amarger, V., D. Gauguier, M. Yerle, F. Apiou, P. Piton, F. Giraudeau, S. Monfouilloux, M. Lathrop, B. Dutrillaux, J. Buard and G. Vergnaud. 1998. Analysis of the human, pig, and rat genomes supports a universal telomeric origin of minisatellite sequences. *Genomics* **52** : 62-71.

Ashley, T. 1994. Mammalian meiotic recombination : a reexamination. *Hum Genet* **94** : 587-93.

Benson, G. and L. Dong. 1999. Reconstructing the duplication history of a tandem repeat. *Ismb*

Benson, G. 1999. Tandem repeats finder : a program to analyze DNA sequences. *Nucleic Acids Res* **27** : 573-80.

Bois, P. and A. J. Jeffreys. 1998. Minisatellite instability and germline mutation. *Cell Mol Life Sci* **55** : 1636-48.

Buard, J. and G. Vergnaud. 1994. Complex recombination events at the hypermutable minisatellite CEB1 (D2S90). *EMBO J.* **13** : 3203-3210.

Debrauwère, H., J. Buard, J. Tessier, D. Aubert, G. Vergnaud and A. Nicolas. 1999. Meiotic instability of human minisatellite CEB1 in yeast requires DNA double-strand breaks. *Nat Genet* **23** : 367-71.

Dubrova, Y. E., V. N. Nesterov, N. G. Krouchinsky, V. A. Ostapenko, G. Vergnaud, F. Giraudeau, J. Buard and A. J. Jeffreys. 1997. Further evidence for elevated human minisatellite mutation rate in Belarus eight years after the Chernobyl accident. *Mut. Res.* **381** : 267-278.

Dubrova, Y. E., M. Plumb, J. Brown, J. Fennelly, P. Bois, D. Goodhead and A. J. Jeffreys. 1998. Stage specificity, dose response, and doubling dose for mouse minisatellite germ-line mutation induced by acute radiation. *Proc Natl Acad Sci U S A* **95** : 6251-5.

Dunham, I., N. Shimizu, B. A. Roe, S. Chissole, A. R. Hunt, J. E. Collins, R. Bruskiwich, D. M. Beare, M. Clamp, L. J. Smink, R. Ainscough, J. P. Almeida, A. Babbage, C. Bagguley, J. Bailey, K. Barlow, K. N. Bates, O. Beasley, C. P. Bird, S. Blakey, A. M. Bridgeman, D. Buck, J. Burgess, W. D. Burrill, K. P. O'Brien et al. 1999. The DNA sequence of human chromosome 22. *Nature* **402** : 489-95.

Haber, J. E. and E. J. Louis. 1997. Minisatellite origins in yeast and humans. *Genomics* **48** : 132-5.

Jeffreys, A. J., Murray J. and Neumann R. 1998. High resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. *Molecular Cell.* **2** : 267-73.

Jeffreys, A. J. and R. Neumann. 1997. Somatic mutation processes at a human minisatellite.

Jeffreys, A. J., K. Tamaki, A. MacLeod, D. G. Monkton, D. L. Neil and J. A. L. Armour. 1994. Complex gene conversion events in germline mutation at human minisatellites. *Nat. Genet.* **6** : 136-145.

Mayer, K., C. Schuller, R. Wambutt, G. Murphy, G. Volckaert, T. Pohl, A. Dusterhoft, W. Stiekema, K. D. Entian, N. Terry, B. Harris, W. Ansorge, P. Brandt, L. Grivell, M. Rieger, M. Weichselgartner, V. de Simone, B. Obermaier, R. Mache, M. Muller, M. Kreis, M. Delseny, P. Puigdomenech, M. Watson, W. R. McCombie et al. 1999. Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature* **402** : 769-77.

Monkton, D. G., R. Neumann, T. Guram, N. Fretwell, K. Tamaki, A. MacLeod and A. J. Jeffreys. 1994. Minisatellite mutation rate variation associated with a flanking DNA sequence polymorphism. *Nat genetics* **8** : 162-170.

Sia, E. A., R. J. Kokoska, M. Dominska, P. Greenwell and T. D. Petes. 1997. Microsatellite instability in yeast : dependence on repeat unit size and DNA mismatch repair genes. *Mol Cell Biol* **17** : 2851-8.

Sybenga, J. 1999. What makes homologous chromosomes find each other in meiosis? A review and an hypothesis. *Chromosoma* **108** : 209-19.

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans* : a platform for investigating biology. *Science* **282** : 2012-8.

Vergnaud G. and F. Denœud. 2000. Minisatellites : Mutability and Genome Architecture. *Genome Research* july 2000.

Vergnaud G. 1989. Polymers of random short oligonucleotides detect polymorphic loci in the human genome. *Nucleic Acides Res.* **17** : 7623-30.

Vergnaud, G., D. Mariat, F. Apiou, A. Aurias, M. Lathrop and V. Lauthier. 1991. The use of synthetic tandem repeats to isolate new VNTR loci : cloning of a human hypermutable sequence. *Genomics* **11** : 135-144.

## ANNEXE 1: MINISATELLITES HYPERMUTABLES HUMAINS CONNUS

Nom du minisatellite	Locus (ou D-number)	Localisation	Accession. number	Séquence du motif consensus	Longueur du consensus	Nb de positions variantes	Homologie entre motifs (%)	Insertions -délétions (%)	%GC (biais)	%Pur (biais)	instabilité	proportion d'événements paternels
CEB1	D2S90	2q37.3 (T)	AF048727	GAGGGGGGAGGGAGGGTGGCCTGCAG AGGTCCCTGGGCT	39	8	90	3	72 (0.53)	68 (0.38)	6.7%	97%
CEB15	D1S172	1p36.33 (T)	AL096805	ACGGAGCAGAGACACGGG	18	8	91	3	68	77 (0.56)	1.5%	100%
CEB25	D10S180	10q26.3 (T)	AL096806	TGGGTGTGAGTGTAAAGTGTGGAGGTA GATGTAAGGGTGGGTGTAAGTGTGGG	52	>20	82	5	53 (0.73)	67 (0.36)	2.5%	65%
CEB36	D10S473	10q26.3 (T)	AL096810	TGTCTGGTGGGATTAGGAGGGAGGTG CCCGTGGTGGGCAGGT	42	18	93	0	64 (0.65)	66 (0.32)	1.8%	50%
CEB42	D8S358	8q24.3 (T)	AL096807	CCTCTGTGTCCCAGGCCGGCGGTAGG ACCCACAGAAGTGTC	41	4	95	0	63 (0.11)	53 (0.08)	0.5%	100%
CEB72	D17S888	17q25 (T)	AL096808	GCGAAGGGAGTGGGCACAGGG	21	0	100	0	71 (0.60)	80 (0.62)	1.8%	65%
MS1	D1S7	1p33-35 (I)	X14856	ATAGGGTGG	9	3	85	0	58 (0.93)	79 (0.60)	5%	50%
MS32	D1S8	1q42 (I)	AF048729	GAGCAGGTGGCCAGGGTGA CT CAGA ATG	29	3	92	2	61 (0.37)	67 (0.35)	0.4%	50%
B6.7		20q13.3 (T)	AF081787	TGAGGGTGGACAGTGAGGGGGTCTCT ACAGGCCA	34	6	91	2	60 (0.5)	68 (0.37)	5.5%	64%
CSTB	EPM1	21q22.3 (T)	U46692	CGCGGGGCGGGG	12	1	96	0	100 (0.5)	100 (0.5)	47%*	76%

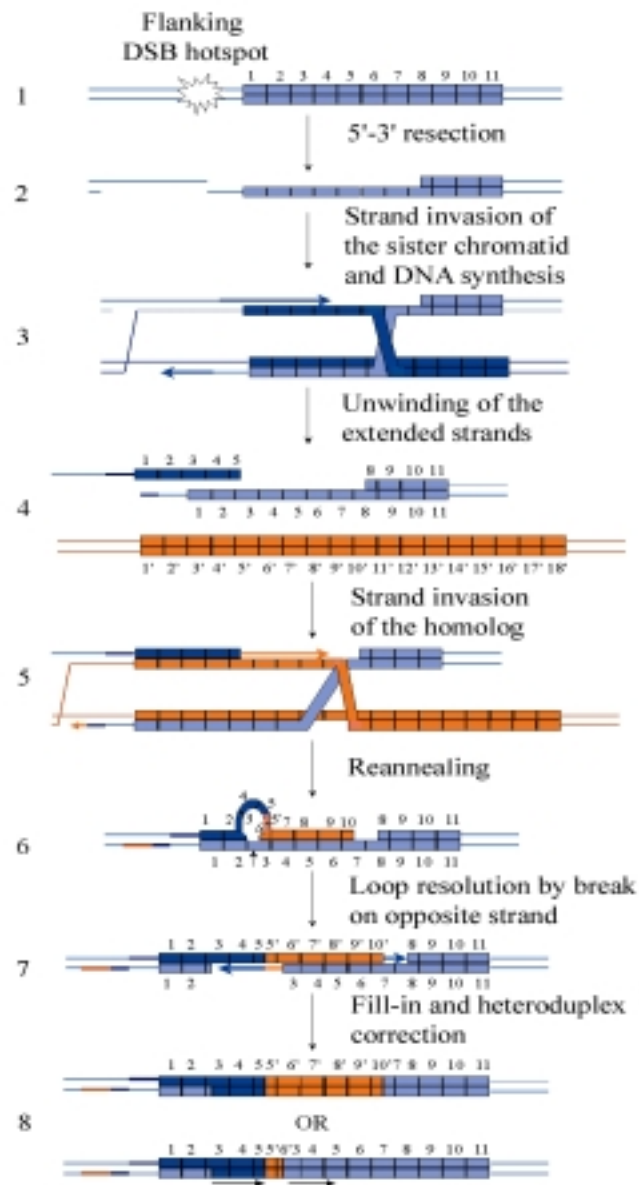
Les valeurs d'instabilité sont des valeurs moyennes sauf pour CSTB (\*) où l'instabilité mesurée correspond à celle d'allèles étendus, pathogènes.

Le pourcentage d'homologie entre les motifs correspond à l'homologie moyenne entre les motifs et le consensus.

Le biais GC est la valeur absolue de  $(\%G - \%C) / (\%G + \%C)$ , le biais purine est la valeur absolue de  $(\%pur - \%pyr)$ .

## ANNEXE 2:

Modèle d'instabilité méiotique des minisatellites  
(Vergnaud & Denoed, 2000)



# ANNEXE 3-1: page d'interrogation de la base de données (<http://minisatellites.u-psud.fr>)

## Consultation de la base de données des séquences répétées dans des chromosomes entièrement séquencés (Query in the tandem repeats data base)

### 1. Sélectionner un chromosome (Select a chromosome) :

- Chromosome 22 humain : 34.6 Mb (Human chromosome 22)
- Chromosome 2 de P falciparum: 925 kb (P falciparum chr 2)
- Chromosome 1 de C elegans: 12.75 Mb (C elegans chr 1)
- Chromosome 4 d'A thaliana: 17.8 Mb (A thaliana chr 4)
- Chromosome 3 de P falciparum: 1060 kb (P falciparum chr 3)
- Chromosome 3 de S cerevisiae: 316 kb (S cerevisiae chr 3)

### 2. Choisir un critère (Select a criterion):

L = Longueur totale de la répétition en tandem (total length)

U = Longueur du motif répété (unit length)

N = Nombre de répétitions du motif (copy number)

V = Variation : % conservation (percent matches)

Pos = Position physique : kb (physical position)

%GC = Pourcentage en G + C (percent G+C)

B = Biais entre les brins (bias between strands)

L	U	N	V	Pos	%GC	B
min : <input type="text" value="500"/>	min : <input type="text" value="17"/>	min : <input type="text" value="5"/>	min : <input type="text" value="80"/>	min : <input type="text" value="0"/>	min : <input type="text" value="60"/>	min : <input type="text" value="0"/>
max : <input type="text" value="50000"/>	max : <input type="text" value="500"/>	max : <input type="text" value="6000"/>	max : <input type="text" value="100"/>	max : <input type="text" value="34600"/>	max : <input type="text" value="100"/>	max : <input type="text" value="1"/>

*Pour une valeur fixe, mettre deux fois la même valeur (To ask for a strict value, write it twice)*

### Mesure de biais (bias)

ex: GC -> B =  $|\%G - \%C| / \%(G+C)$

### 3. Choisir le type de requête (Select Query type) :

- Répartition des répétitions (Tandem repeats distribution)
- Isolement de séquences répétées (Tandem repeats characteristics)

Répartition selon (distribution according to):

Intervalle de distance physique pour la répartition selon Pos (step for physical distribution):

### 4. Filtrer (filter) :

Éliminer (eliminate) :

### 5. Resultat (result) :

- Français
- English

Lancer la requête (Submit)

## ANNEXE 3-2: Résultat d'une requête « répartition »

### Homo sapiens chr 22

#### Répartition selon Pos des répétitions en tandem répondant aux critères :

1. Longueur totale du minisatellite entre 500 et 50000
2. Taille du motif entre 17 et 500
3. Nombre de répétitions du motif entre 5 et 6000
4. Conservation du motif entre 80 et 100
5. Pourcentage en G+C entre 60 et 100

Position (kb)	Histogramme	Nb seq	Synténie	Recomb	G+C
0 -> 1000	▮	1			GC -
1000 -> 2000	▮ ▮	2	MMu_6		GC -
2000 -> 3000	▮	1			
3000 -> 4000	▮	1	MMu_16		
4000 -> 5000	▮	1	MMu_16		
5000 -> 6000	▮ ▮	2	MMu_16		
6000 -> 7000		0	MMu_16		
7000 -> 8000	▮ ▮	2	MMu_10		GC+
8000 -> 9000		0			
9000 -> 10000	▮	1	MMu_5	recomb	GC -
10000 -> 11000		0	MMu_5	recomb	GC -
11000 -> 12000		0	MMu_5		
12000 -> 13000		0			
13000 -> 14000	▮	1	MMu_11		
14000 -> 15000		0	MMu_11		
15000 -> 16000		0	MMu_11		
16000 -> 17000	▮	1			GC -
17000 -> 18000		0			GC -
18000 -> 19000		0	MMu_8		GC -
19000 -> 20000	▮	1	MMu_15	recomb	
20000 -> 21000	▮	1	MMu_15	recomb	GC+
21000 -> 22000	▮	1	MMu_15		GC+
23000 -> 24000		0	MMu_15		
24000 -> 25000	▮	1	MMu_15		GC -
25000 -> 26000		0	MMu_15		GC+
26000 -> 27000	▮	1	MMu_15		GC+
27000 -> 28000	▮ ▮	2	MMu_15		
28000 -> 29000	▮	1	MMu_15	recomb	
29000 -> 30000	▮	1	MMu_15		
30000 -> 31000	▮ ▮ ▮ ▮	4	MMu_15		
31000 -> 32000	▮	1	MMu_15		
32000 -> 33000	▮ ▮ ▮ ▮ ▮	5	MMu_15	recomb	
33000 -> 34000	▮ ▮ ▮ ▮ ▮ ▮ ▮ ▮ ▮	9	MMu_15	?	
34000 -> <b>34600</b>	▮ ▮ ▮ ▮ ▮	5	MMu_15	?	

**Nombre total de séquences répondant aux critères : 47**



# ANNEXE 3-3: résultat d'une requête « isolement »

## Homo sapiens chr 22

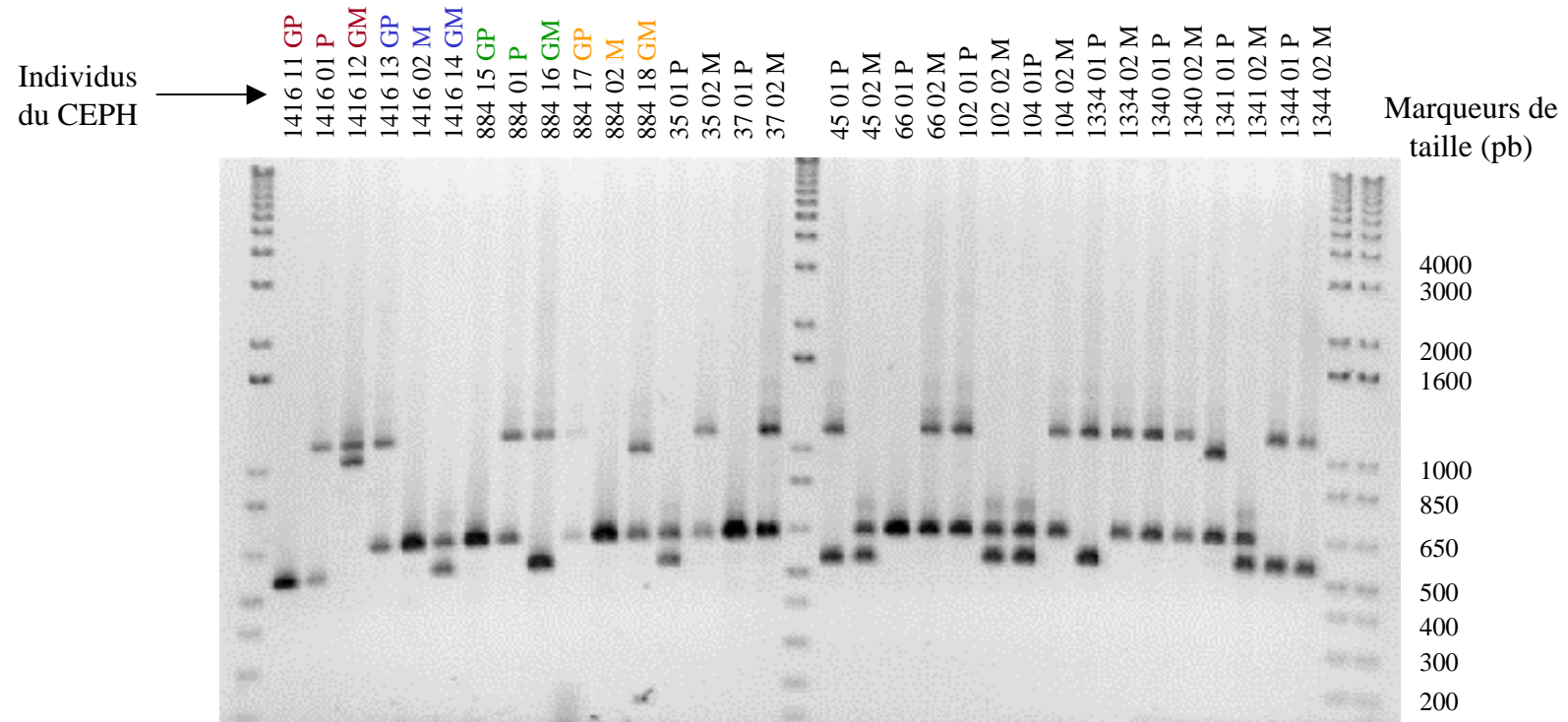
### Isolement des répétitions en tandem répondant aux critères :

1. Longueur totale du minisatellite entre 500 et 50000
2. Taille du motif entre 17 et 500
3. Nombre de répétitions du motif entre 5 et 6000
4. Conservation du motif entre 80 et 100
5. Pourcentage en G+C entre 60 et 100

Position physique (kb)	Nom du contig	Taille du motif	Nombre de répétitions	Longueur totale	Conservation	%C	%G	%T	%A	Biais GC	Séquence
46	contig1A	194	5	1016	84%	33%	29%	22%	14%	6,45161290218522E-02	<a href="#">Lien vers l'alignement</a>
1528	contig3D	41	21	855	91%	9%	54%	17%	18%	0,714285714172336	<a href="#">Lien vers l'alignement</a>
1571	contig3D	62	25	1548	94%	43%	19%	18%	18%	0,387096774131113	<a href="#">Lien vers l'alignement</a>
2783	contig4B	27	28	745	83%	30%	35%	25%	8%	7,69230769112426E-02	<a href="#">Lien vers l'alignement</a>
3817	contig4G	39	41	1586	91%	22%	41%	18%	17%	0,301587301539431	<a href="#">Lien vers l'alignement</a>
4285	contig5_1A	27	291	7850	83%	35%	32%	10%	20%	4,47761193963021E-02	<a href="#">Lien vers l'alignement</a>
5111	contig5_1E	27	47	1266	86%	37%	30%	7%	24%	0,104477611924705	<a href="#">Lien vers l'alignement</a>
5236	contig5_1F	27	35	934	85%	30%	36%	25%	7%	9,09090908953168E-02	<a href="#">Lien vers l'alignement</a>
7364	contig5_2E	33	20	669	90%	32%	43%	2%	21%	0,146666666647111	<a href="#">Lien vers l'alignement</a>
7517	contig5_2F	44	12	547	99%	29%	34%	15%	20%	7,93650793524817E-02	<a href="#">Lien vers l'alignement</a>
9314	contig5_3E	44	16	723	93%	34%	29%	20%	16%	7,93650793524817E-02	<a href="#">Lien vers l'alignement</a>
16658	contig5_6J	21	28	602	85%	58%	13%	2%	25%	0,63380281681214	<a href="#">Lien vers l'alignement</a>
19338	contig5_8B	36	26	930	86%	42%	25%	25%	7%	0,253731343245712	<a href="#">Lien vers l'alignement</a>
20915	contig5_8J	32	16	510	94%	25%	43%	10%	20%	0,264705882314014	<a href="#">Lien vers l'alignement</a>
21044	contig5_9A	18	47	851	80%	58%	5%	13%	21%	0,841269841136306	<a href="#">Lien vers l'alignement</a>
22624	contig5_9I	176	6	1018	93%	40%	39%	10%	10%	0,012658227846499	<a href="#">Lien vers l'alignement</a>
24318	contig5_10F	176	10	1753	86%	39%	38%	10%	10%	1,29870129853264E-02	<a href="#">Lien vers l'alignement</a>
26884	contig5_11I	64	10	642	92%	44%	24%	6%	24%	0,294117647015571	<a href="#">Lien vers l'alignement</a>
27654	contig6B	28	33	918	91%	55%	22%	13%	7%	0,42857142851577	<a href="#">Lien vers l'alignement</a>
27656	contig6B	42	20	860	91%	36%	24%	16%	22%	0,199999999966667	<a href="#">Lien vers l'alignement</a>
28638	contig7C	33	32	1044	92%	37%	34%	21%	6%	4,22535211208094E-02	<a href="#">Lien vers l'alignement</a>
29664	contig8A	41	19	770	94%	25%	46%	12%	14%	0,295774647845666	<a href="#">Lien vers l'alignement</a>
30400	contig8C	49	13	643	98%	28%	34%	24%	12%	9,67741935327784E-02	<a href="#">Lien vers l'alignement</a>
30404	contig8C	46	16	755	90%	51%	23%	18%	6%	0,378378378327246	<a href="#">Lien vers l'alignement</a>
30790	contig8E	43	14	590	92%	33%	27%	13%	24%	9,99999998333333E-02	<a href="#">Lien vers l'alignement</a>
30945	contig8F	27	19	502	87%	30%	42%	23%	3%	0,166666666643519	<a href="#">Lien vers l'alignement</a>
31995	contig8L	46	11	507	86%	36%	27%	13%	23%	0,142857142834467	<a href="#">Lien vers l'alignement</a>
32021	contig8L	244	6	1425	80%	35%	27%	20%	16%	0,129032258043704	<a href="#">Lien vers l'alignement</a>
32447	contig9A	41	26	1064	92%	36%	26%	19%	17%	0,161290322554631	<a href="#">Lien vers l'alignement</a>
32458	contig9A	23	43	996	80%	57%	13%	23%	4%	0,628571428481633	<a href="#">Lien vers l'alignement</a>
32751	contig10A	96	13	1254	89%	37%	24%	25%	12%	0,213114754063424	<a href="#">Lien vers l'alignement</a>
33057	contig10C	33	33	1086	96%	22%	49%	9%	18%	0,380281690087284	<a href="#">Lien vers l'alignement</a>
33197	contig10D	68	12	840	85%	27%	33%	17%	20%	9,99999998333333E-02	<a href="#">Lien vers l'alignement</a>
33279	contig10E	37	14	515	89%	36%	24%	21%	18%	0,199999999966667	<a href="#">Lien vers l'alignement</a>
33318	contig10E	45	22	1024	88%	17%	47%	16%	18%	0,468749999926758	<a href="#">Lien vers l'alignement</a>
33419	contig10F	32	18	563	91%	14%	53%	16%	16%	0,582089552151927	<a href="#">Lien vers l'alignement</a>
33457	contig10F	49	12	604	91%	34%	37%	12%	15%	4,22535211208094E-02	<a href="#">Lien vers l'alignement</a>
33592	contig10G	47	13	611	98%	42%	18%	20%	17%	0,399999999933333	<a href="#">Lien vers l'alignement</a>
33854	contig11A	18	42	751	81%	11%	50%	27%	9%	0,639344262190271	<a href="#">Lien vers l'alignement</a>
33963	contig11B	24	40	954	87%	42%	23%	16%	17%	0,292307692262722	<a href="#">Lien vers l'alignement</a>
34017	contig11B	81	9	703	84%	31%	30%	20%	16%	1,63934426202634E-02	<a href="#">Lien vers l'alignement</a>
34182	contig12A	62	28	1732	93%	21%	50%	14%	14%	0,408450704167824	<a href="#">Lien vers l'alignement</a>
34357	contig12B	57	11	625	85%	23%	47%	8%	20%	0,342857142808163	<a href="#">Lien vers l'alignement</a>
34458	contig12B	56	9	506	85%	24%	38%	23%	13%	0,225806451576483	<a href="#">Lien vers l'alignement</a>
34495	contig12C	25	32	807	92%	30%	31%	16%	22%	1,63934426202634E-02	<a href="#">Lien vers l'alignement</a>

Nombre total de séquences répondant aux critères : 47

## ANNEXE 4: Photo d'un gel après migration des produits de PCR



Ce minisatellite compte 4 allèles de 1200, 1000, 650, et 550 pb environ, qui ségrègent convenablement au sein des familles.

# ANNEXE 5

## Minisatellite chr 22 (28 x 62pb)



## Homologue chr18



## Modèle proposé par Haber (1997) pour l'origine des minisatellites

