

MUSEUM NATIONAL D'HISTOIRE NATURELLE



Sciences de la Nature et de l'Homme : Evolution et Ecologie - ED 227

Année 2016

N° attribué par la bibliothèque

| - | - | - | - | - | - | - | - | - | - | - |

THESE

Pour obtenir le grade de

DOCTEUR DU MUSEUM NATIONAL D'HISTOIRE NATURELLE

Spécialité : Biologie moléculaire - Génomique évolutive

Présentée et soutenue publiquement par

Lauriane CACHEUX

le 15 Novembre 2016

Diversité et histoire évolutive de l'ADN alpha satellite chez les Cercopithèques

Sous la direction de : **Dr. ESCUDE Christophe, Chargé de recherche, CNRS**

Pr. DENYS Christiane, Professeur, MNHN

JURY :

Dr. YERLE Martine
Pr. GIULOTTO Elena
Dr. RIVALS Eric
Dr. KOSZUL Romain
Pr. DENYS Christiane
Dr. ESCUDE Christophe

Directrice de recherche, INRA, Toulouse
Professeur, Université de Pavie
Directeur de recherche, CNRS, Montpellier
Directeur de recherche, CNRS, Paris
Professeur, MNHN, Paris
Chargé de recherche, CNRS, Paris

Présidente du jury
Rapportrice
Rapporteur
Examinateur
Directrice de Thèse
Directeur de Thèse



Merci à mes directeurs, aux membres de mon jury de thèse, et aux nombreuses personnes qui, au-delà de m'avoir accompagnée dans ce voyage, ont illuminé chacun de mes jours ; avec une pensée toute particulière pour Loïc Ponger, Astrid Lancrey, Bernard et Anne Marie Dutrillaux, Michèle Gerbault-Seureau, Violaine Nicolas, Jean-Baptiste Boulé, Emmanuelle Delagoutte, Alexandre Boutorine, Judith Lopes, François Strauss, Christophe Lavelle, François Loll, Julien Cochennec, Evelyne Duvernois, Valentine Arlot, Mathieu Valade, Peggy Motsch, Nicolas Buisine, Amélie Bonnet-Garnier, Sylvie Aulard-Widemann, Régis Debruyne, Delphine Gey, José Utge et Jérôme Fuchs.

"Il y a entre les êtres et les choses des relations de prodige ; dans cet inépuisable ensemble, de soleil à puceron, on ne se méprise pas ; on a besoin les uns des autres. La lumière n'emporte pas dans l'azur les parfums terrestres sans savoir ce qu'elle en fait ; la nuit fait des distributions d'essence stellaire aux fleurs endormies. Tous les oiseaux qui volent ont à la patte le fil de l'infini. La germination se complique de l'élosion d'un météore et du coup de bec de l'hirondelle brisant l'oeuf. Où finit le télescope, le microscope commence. Lequel des deux a la vue la plus grande ? Choisissez."

Les Misérables, Victor Hugo

*À Florence Richard, qui a fait naître ce magnifique projet
et avec qui ç'aurait été un bonheur de le voir grandir*

Résumé

Les régions centromériques reposent, chez les Primates, sur une famille de séquences répétées en tandem appelée l'ADN alpha satellite. Les monomères de cet ADN (≈ 170 pb) se sont diversifiés au cours de l'évolution, formant des familles de séquences aux profils d'organisation et distribution variés. La diversité des alpha satellites chez les primates non-humains reste cependant peu caractérisée, et la compréhension de la dynamique évolutive de cet ADN nécessite son intégration dans de plus larges analyses comparatives. Les cercopithèques, qui présentent une évolution chromosomique originale par fissions et émergences de nouveaux centromères, apparaissent comme des modèles d'étude prometteurs.

Nous avons appliqué une nouvelle technologie de séquençage à des monomères et dimères d'alpha satellites, isolés à partir des génomes de *Cercopithecus solatus* ($2n = 60$) et *C. pogonias* ($2n = 72$). Ces deux espèces appartiennent à des lignées primaires distinctes au sein des cercopithèques, et ont divergé l'une de l'autre il y a plusieurs millions d'années. L'analyse computationnelle des séquences collectées a permis la caractérisation de six familles d'alpha satellites, dont quatre sont partagées entre espèces et deux ne sont retrouvées que chez *C. pogonias*. Au moins trois familles seraient impliquées dans des répétitions d'ordre supérieur, profil d'organisation jusque là inconnu dans l'ADN alpha satellite des cercopithèques. L'hybridation in situ en fluorescence des familles identifiées, réalisée grâce à des sondes oligonucléotidiques hautement discriminantes, a permis de visualiser leur distribution sur les chromosomes de *C. solatus* et *C. pogonias*. Certaines de ces familles se distribuent différemment entre chromosomes, révélant l'existence d'une diversité interchromosomique de l'ADN alpha satellite chez les singes de l'Ancien Monde. Leurs positions sur les régions centromériques vont en faveur de l'hypothèse du gradient d'âge des alpha satellites, selon laquelle les familles se forment aux centromères en déplaçant les familles préexistantes vers les péricentromères. L'extension de cette analyse cytogénétique à quinze espèces et l'interprétation de ses résultats à la lumière d'une phylogénie moléculaire, nouvellement reconstruite, nous ont permis de proposer un scénario évolutif pour l'ADN alpha satellite chez les cercopithèques. Celui-ci apparaît évoluer de manière concertée avec les chromosomes, se diversifiant et se déplaçant sur les régions centromériques à mesure que ces derniers se fissinent et voient l'émergence de nouveaux centromères. Ces travaux ont enfin apporté des informations nouvelles quant aux relations de parenté entre cercopithèques, invitant à une intégration de l'ADN alpha satellite dans l'étude de l'histoire évolutive des Primates.

L'approche méthodologique mise au point a permis de caractériser la diversité et de comprendre l'évolution de l'ADN alpha satellite chez les cercopithèques. Elle pourra être appliquée à l'étude de ces séquences particulières chez d'autres primates, ainsi qu'à l'étude de différents satellites chez des espèces primates comme non-primates.

Mots-clés : ADN alpha satellite, Cercopithèques, Centromère, Evolution des génomes

Abstract

Alpha satellite DNA is the main family of tandemly repeated sequences lying in primate centromere regions. Alpha satellite monomers (≈ 170 bp) diversified during the course of evolution, forming distinct families of alpha satellite sequences that exhibit specific organizational and distribution patterns. The limited amount of studies concerning non-human primates is a restriction to the understanding of alpha satellite evolutionary dynamics, which calls for the integration of this element into comparative studies. Cercopithecini, which display an unusual chromosomal evolution by multiple fissions and new centromere formations, constitute a promising study model.

We carried out next generation sequencing of alpha satellite monomers and dimers isolated from the *Cercopithecus solatus* ($2n = 60$) and *C. pogonias* ($2n = 72$) genomes. These species belong to different primary lineages within the Cercopithecini tribe and diverged from each other several million years ago. Computational tools were used to analyze the collected sequences and characterize six alpha satellite families, four of them being shared between species and two being limited to *C. pogonias*. At least three families belong to higher order repeats, an organizational pattern that had never been observed in Cercopithecini. The fluorescence *in situ* hybridization of each family, performed with highly discriminant oligonucleotide probes, showed their distribution on *C. solatus* and *C. pogonias* chromosomes. Some of them distribute on distinct sets of chromosomes, disclosing the existence of alpha satellite interchromosomal diversity in Old World monkeys. Their position along centromeric regions is largely in accordance with the age-gradient hypothesis, according to which new families expand at centromere, thereby splitting and displacing older families toward pericentromeres. The extension of this analysis to fifteen species, combined to a newly reconstructed molecular phylogeny, allowed us to propose an evolutionary scenario for alpha satellite DNA in Cercopithecini. Alpha satellite DNA diversification and displacement on centromere regions appear intimately connected to chromosome rearrangement dynamics, including new centromere formations, which suggests that centromeres and chromosomes evolve in a concerted manner. Finally, this work provided information about Cercopithecini relationships and thus encourages the integration of alpha satellite DNA into the study of primate evolutionary history.

Our new methodological approach allowed deciphering alpha satellite diversity and dynamics in Cercopithecini. This framework could be used to study alpha satellite DNA in other primates, and be applied to different satellites in primates as in non-primate species.

Key words : Alpha satellite DNA, Cercopithecini, Centromere, Genome evolution

Table des matières

1	Introduction	19
1.1	Diversité et fonction des séquences d'ADN répétées chez les Eucaryotes	19
1.1.1	Séquences d'ADN répétées dispersées	19
1.1.2	Séquences d'ADN répétées en tandem	21
1.1.3	Séquences répétées et régulation des génomes	23
1.2	Organisation des régions centromériques chez les Mammifères	26
1.2.1	Structure épigénétique des régions centromériques	27
1.2.2	ADN des régions centromériques	29
1.3	Dynamique évolutive de l'ADN alpha satellite chez les Primates	38
1.3.1	Diversité d'organisations et de séquences	38
1.3.2	Diversité et distribution chromosomique	45
1.3.3	Mécanismes évolutifs	48
1.4	Diversification et évolution chromosomique chez les Cercopithèques	55
1.4.1	Diversité écologique et comportementale	55
1.4.2	Relations de parenté entre espèces	63
1.4.3	Cytogénétique évolutive des cercopithèques	67
1.5	Projet de thèse	82

2 Résultats	83
2.1 Nouvelle approche d'étude de l'ADN alpha satellite appliquée à <i>C. solatus</i>	83
2.1.1 Abstract	84
2.1.2 Introduction	84
2.1.3 Results	87
2.1.4 Discussion	97
2.1.5 Conclusion	100
2.1.6 Methods	101
2.1.7 Additional files	105
2.2 Diversité et distribution de l'ADN alpha satellite chez <i>C. pogonias</i>	120
2.2.1 Abstract	120
2.2.2 Introduction	121
2.2.3 Results	123
2.2.4 Discussion	133
2.2.5 Conclusion	135
2.2.6 Methods	136
2.2.7 Additional files	140
2.3 Histoire évolutive de l'ADN alpha satellite chez les Cercopithèques	141
2.3.1 Abstract	141
2.3.2 Introduction	142
2.3.3 Results	144
2.3.4 Discussion	157

2.3.5	Conclusion	160
2.3.6	Methods	160
2.3.7	Additional files	163
3	Discussion et conclusion	165
3.1	Méthodologie d'étude de l'ADN alpha satellite	165
3.1.1	Isolement, séquençage et analyse bioinformatique	165
3.1.2	Analyse cytogénétique	166
3.2	Diversité et évolution de l'ADN alpha satellite chez les Cercopithèques	167
3.2.1	Pluralité de familles et d'organisations	167
3.2.2	Dynamique d'émergence et d'extinction	168
3.2.3	Histoire évolutive des Cercopithèques	170
3.3	Conclusion	170
4	Matériel supplémentaire	173
4.1	Matériel supplémentaire : premier article	173
4.1.1	Figures supplémentaires	173
4.1.2	Tables supplémentaires	180
4.2	Matériel supplémentaire : deuxième article	181
4.2.1	Figures supplémentaires	181
4.2.2	Tables supplémentaires	193
4.3	Matériel supplémentaire : troisième article	194
4.3.1	Figures supplémentaires	194
4.3.2	Tables supplémentaires	214

Chapitre 1

Introduction

1.1 Diversité et fonction des séquences d'ADN répétées chez les Eucaryotes

Les génomes eucaryotes contiennent des millions de copies d'éléments répétés d'ADN présentant une variété d'origines et d'organisations. Tandis que certains sont déjà largement caractérisés, comme l'ADN ribosomal, de transfert ou télomérique, l'obscurité persiste sur la plupart de ces séquences répétées. Successivement envisagées comme des éléments génomiques fonctionnels et comme de simples parasites à la recherche de véhicules à leur reproduction, elles sont aujourd'hui largement étudiées dans l'optique de comprendre leur influence sur l'organisation et l'expression des génomes. Nous discuterons ici des éléments transposables et de l'ADN satellite, qui font respectivement partie des séquences répétées dispersées et des séquences répétées en tandem (Figure 1.1).

1.1.1 Séquences d'ADN répétées dispersées

L'impact des éléments transposables sur l'expression des gènes et les phénotypes qui en découlent a mené à leur découverte en 1950 chez le maïs (McClintock, 1950), où ils influencent la pigmentation de certains grains. La mise en évidence de leur rôle dans la dysgénésie des hybrides chez la drosophile (Picard et al., 1978; Rubin and Spradling, 1982; Bingham et al., 1982) a stimulé l'intérêt porté par la communauté scientifique sur ces séquences originales, qui remettaient en question la stabilité admise de l'ADN. Capables de se répliquer et de se déplacer d'un locus chromosomique à un autre, les éléments transposables ont été depuis décrits dans l'ensemble du vivant (Hua-Van et al., 2011) et peuvent représenter jusqu'à 80 % des génomes (Schnable et al., 2009). Le système de classification des éléments transposables le plus utilisé permet de

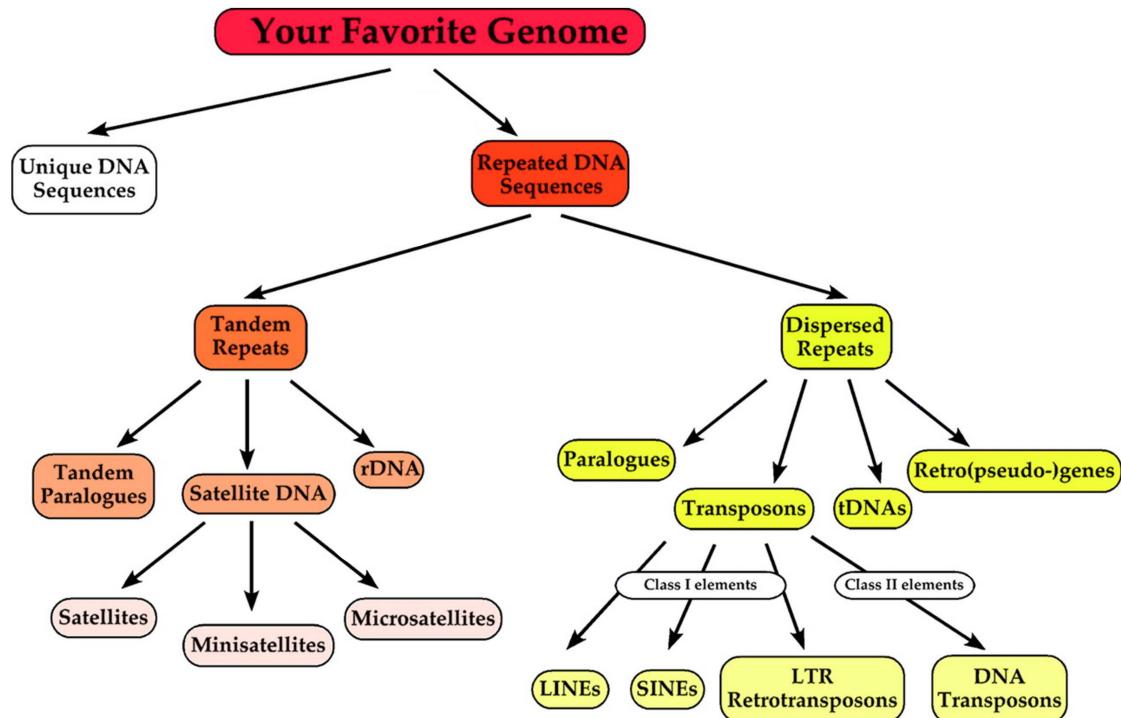


FIGURE 1.1 – Séquences d’ADN répétées dans les génomes eucaryotes (d’après Richard et al. (2008)).

distinguer parmi eux deux grandes classes de séquences utilisant différents intermédiaires de transposition (Finnegan, 1989).

Eléments transposables de classe I

Les éléments de classe I, appelés également rétrotransposons, utilisent une molécule d’ARN comme intermédiaire de transposition. Leur séquence est en effet transcrive en ARN, puis rétrotranscrit par une transcriptase inverse en un ADN voué à s’insérer dans le génome. Ils sont généralement de grande taille, allant de quelques centaines à quelques milliers de paires de bases. Deux types d’éléments différant par leur structure et mécanisme d’intégration sont distingués au sein de cette classe.

Les rétrotransposons à LTR (Long terminal repeats) ont une structure et un fonctionnement proches de ceux des virus et tiennent leur nom des longues répétitions d’ADN qu’ils portent à leurs extrémités. Ces rétrotransposons codent pour une protéine Gag (Group-specific antigen) qui s’associe avec leur intermédiaire ARN en formant des particules pseudo-virales, ainsi que pour une réverse transcriptase et une RNase qui prennent en charge la rétrotranscription de ces transcrits dans le cytosol, et enfin pour une intégrase qui assurera l’intégration génomique ultérieure de l’ADNc généré (Eickbush and Jamburuthugoda, 2008; Finnegan, 2012).

Les rétrotransposons sans LTR voient leur intermédiaire ARN pris en charge par une réverse transcriptase et une nucléase dans le cytosol, puis ce complexe s'associer à un site d'insertion sur l'ADN génomique. Une coupure simple brin induite par la nucléase permet l'appariement de l'extrémité du transcrit à cet ADN puis sa rétrotranscription par la réverse transcriptase. Ce sont les enzymes de réparation de la cellule hôte qui achèvent l'intégration du rétrotransposon dans le génome. Les rétrotransposons sans LTR comprennent les LINE (Long interspersed nuclear elements), qui codent pour la nucléase et la réverse transcriptase nécessaires à leur copie, et les SINE (Short interspersed nuclear elements) qui se multiplient en détournant la machinerie de transposition des LINE ; on les dit pour cela non autonomes ([Dewannieux et al., 2003](#); [Finnegan, 2012](#)).

Eléments transposables de classe II

Les éléments de classe II, également appelés transposons à ADN, utilisent une molécule d'ADN comme intermédiaire de transposition. Plus courts que les éléments de classe I, ils peuvent être délimités par des TIR (Terminal inverted repeats) de taille variable, et codent généralement pour une transposase. Une fois traduite, cette protéine se fixe sur une copie de l'élément transposable d'origine afin de provoquer des cassures de l'ADN, de l'exciser puis de l'insérer à un autre endroit du génome. Ce mécanisme de transposition implique généralement un intermédiaire ADN double brin et ne permet donc pas une augmentation du nombre de copies de l'élément considéré, à moins que la transposition ne se déroule au moment de la réPLICATION, depuis un locus nouvellement répliqué vers un locus non encore répliqué ([Feschotte and Pritham, 2007](#); [Cerveau et al., 2011](#)). Cependant, des éléments de classe II utilisant un intermédiaire ADN simple brin ont récemment été découverts, et transposeraient via un processus qui serait cette fois réPLICATIF ([Kapitonov and Jurka, 2001](#); [Pritham and Feschotte, 2007](#)). On trouve également chez les transposons à ADN des éléments non autonomes, comme les MIKE (Miniature inverted-repeat transposable elements), qui ne codent pour aucune protéine mais utilisent la machinerie de transposition d'autres éléments de classe II pour se déplacer dans les génomes ([Feschotte and Mouches, 2000](#)).

1.1.2 Séquences d'ADN répétées en tandem

La présence d'ADN satellite dans les génomes eucaryotes a été mise en évidence en 1961 chez la souris et le cochon d'Inde ([Kit, 1961](#)) : présentant un biais de composition nucléotidique par rapport au reste du génome, cet ADN répété en tandem peut apparaître, suite à un fractionnement par centrifugation en gradient de chlorure de césium, sous la forme de bandes additionnelles "satellites" aux côtés de la bande principale d'ADN génomique ([Richard et al., 2008](#); [Biscotti et al., 2015](#)). De nombreuses familles d'ADN satellite ont depuis été caractérisées dans l'ensemble du vivant. Elles présentent une diversité dans la taille et la composition de leurs unités

de répétition, mais également dans la longueur occupée par leurs répétitions. Trois catégories peuvent se distinguer au sein de l'ADN satellite sur la base de ces caractères : les satellites, minisatellites et microsatellites.

Satellites

Les satellites sont organisés en longues répétitions allant de plusieurs centaines de milliers à plusieurs millions de paires de bases ([Charlesworth et al., 1994](#)). Ils sont constitués d'unités de répétition d'une taille généralement supérieure à 100 pb, et allant le plus souvent de 140 à 180 pb ou de 300 à 360 pb, tout comme l'ADN des mono- et dinucléosomes ([Macas et al., 2002](#)). Des expériences de cartographie cytogénétique ont permis de montrer que les satellites étaient préférentiellement localisés dans l'hétérochromatine des régions centromériques et subtélomériques des chromosomes eucaryotes. Très peu représentés chez de rares organismes, comme la levure *Schizosaccharomyces pombe* ([Wood et al., 2002; Richard et al., 2008](#)), ils peuvent chez d'autres constituer plus de la moitié des génomes ([Hacch and Mazrimas, 1974; Petitpierre et al., 1995](#)). Différentes familles de satellites peuvent coexister et être retrouvées entre espèces proches ; leur nombre de copies et leurs localisations chromosomiques peuvent en revanche varier drastiquement d'une espèce à l'autre ([Vershinin et al., 1996; Ross et al., 1997; Mestrovic et al., 1998; Nijman and Lenstra, 2001](#)).

Minisatellites

Les minisatellites, ou VNTR (Variable number tandem repeats), sont organisés en courtes répétitions de l'ordre du millier de paires de bases. Ils sont constitués d'unités adoptant une taille allant généralement de 7 à 100 pb ([Vergnaud and Denoeud, 2000](#)). Leur nombre de copies à un locus donné peut varier entre individus via des délétions ou duplications en tandem, ce qui a permis leur découverte suite à l'identification d'un locus humain présentant un polymorphisme interindividuel de longueur ([Wyman and White, 1980](#)). Généralement riches en GC, ces minisatellites polymorphes sont préférentiellement associés aux régions subtélomériques des chromosomes eucaryotes ([Piazza et al., 2012](#)).

Microsatellites

Les microsatellites, également appelés STR (Short tandem repeats), sont organisés en très courtes répétitions d'une taille inférieure à 150 pb et dont les unités mesurent entre 1 et 6 pb seulement ([Schlötterer, 2000](#)). Tout comme les VNTR, ils présentent une variabilité interindividuelle en nombre de copies à un locus donné, mais se distribuent en revanche dans l'ensemble du génome. Depuis le début des années 1990, la possibilité d'amplifier l'ADN par PCR a permis

que les microsatellites, comme les minisatellites, soient largement utilisés en biologie évolutive comme marqueurs génétiques pour étudier les relations généalogiques entre individus, la structure et l'évolution de populations ou encore l'origine d'animaux et plantes domestiqués (Litt and Luty, 1989; Armour et al., 1996; Jarne and Lagoda, 1996; Bowers et al., 1999; Vilà et al., 2001).

1.1.3 Séquences répétées et régulation des génomes

Bien qu'aucune fonction cellulaire des éléments transposables et de l'ADN satellite en tant que tels ne soit avérée, il est aisément envisager une adaptation évolutive structurale et fonctionnelle des génomes eucaryotes à la prolifération en leur sein de ces séquences répétées. Bien connues pour leur rôle clé dans certaines innovations évolutives, comme la gestation placentaire ou l'immunité adaptative (Agrawal et al., 1998; Lynch et al., 2011), les séquences répétées se voient à présent attribuer des rôles essentiels dans l'architecture et l'expression des génomes en général.

Architecture des génomes

Les séquences répétées peuvent impacter, de manière active, l'état de la chromatine le long des chromosomes et l'organisation de ces chromosomes dans l'espace du noyau.

L'état de la chromatine dépendrait en effet en partie de sa composition en séquences répétées. L'hétérochromatine constitutive présentant une forte densité en séquences satellites et éléments transposables, leur rôle dans l'établissement et le maintien de cette conformation chromatinnienne est depuis longtemps proposé (Yunis and Yasmineh, 1971). Dans les années 1990, l'élaboration de systèmes expérimentaux présentant de plus ou moins longues répétitions de transgènes génomiques a permis d'établir une corrélation positive entre le nombre de copies d'une séquence et son hétérochromatisation, ce qui a inspiré l'acronyme RIGS (Repeat-induced gene silencing) (Assaad et al., 1993; Dorer and Henikoff, 1994; Garrick et al., 1998). Plus récemment, la découverte d'ARN transcrits depuis les répétitions inversées des régions centromériques de *S. pombe* a apporté de premiers éléments en faveur d'un rôle actif des séquences répétées dans l'établissement de l'état hétérochromatique. Ces transcrits seraient en effet pris en charge par un complexe RITS (RNA-induced transcriptional silencing), et le guideraient aux répétitions centromériques où il induirait le recrutement de protéines et l'apposition de marques épigénétiques répressives (Figure 1.2). La mutation d'acteurs impliqués dans ce système entraîne par ailleurs une accumulation aberrante de transcrits centromériques, une perte des marques répressives au niveau des séquences centromériques et une détérioration de la fonction du centromère (Volpe et al., 2002; Verdel et al., 2004; Bühler et al., 2006). Un tel silencing a par ailleurs été décrit chez le coléoptère *Tribolium castaneum*; les satellites centromériques impliqués sont également retrouvés dispersés dans les régions euchromatiques de son génome, où leurs gènes adjacents

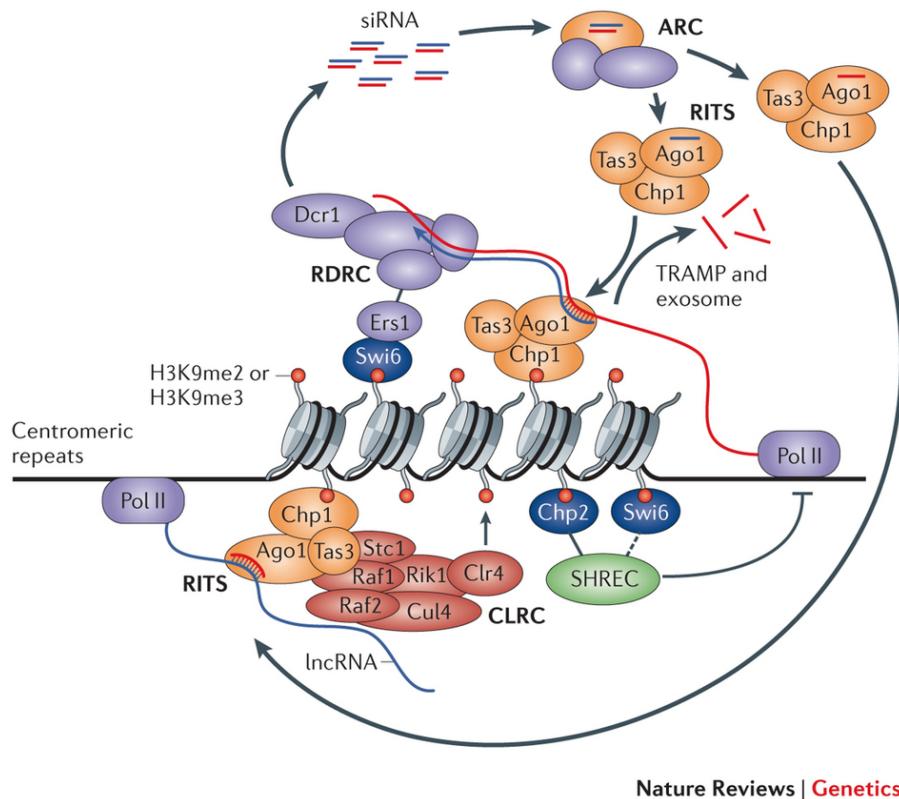


FIGURE 1.2 – Modèle du "RNA-induced transcriptional silencing" (RITS) des répétitions centromériques chez *S. pombe* (d'après Holoch and Moazed (2015)). Le complexe RITS cible les transcrits des répétitions centromériques et induit leur transformation en ARNsi (siRNAs, small interfering RNAs). Il les utilisera comme guides pour retrouver de nouveaux transcrits naissants depuis les régions centromériques et y recruter des protéines responsables de la condensation chromatinienne.

sont eux aussi impactés par la répression induite (Feliciello et al., 2015). Un autre système similaire de silencing pouvant induire la compaction de la chromatine au niveau des éléments transposables, la voie PIWI, a depuis été décrit chez la drosophile, le poisson zèbre et la souris (Brennecke et al., 2007; Houwing et al., 2007; Carmell et al., 2007; Sienski et al., 2012).

L’organisation du génome dans le noyau en interphase serait également influencée par les séquences répétées. Le séquençage et l’assemblage des génomes humains et murins, déjà très avancés au début des années 2000, ont permis d’observer la conservation de LINE et de SINE dans des régions synténiques entre ces deux modèles ; une sélection positive s’appliquerait à la position de ces éléments dans les génomes (Silva et al., 2003). Le développement récent des techniques de capture de conformation chromosomique (Dekker et al., 2002; de Laat and Dekker, 2012) a par ailleurs permis de montrer que l’organisation tridimensionnelle de ces régions synténiques était elle aussi conservée, et que la fréquence des contacts entre des régions distantes du génome était positivement corrélée avec leur enrichissement en rétrotransposons (Cournac et al., 2016). Ces résultats suggèrent un rôle des éléments transposables dans le repliement du génome et le maintien de sa structure tridimensionnelle. De plus, des rétrotransposons, comme *gypsy* chez la drosophile, sont impliqués dans la formation de boucles chromatiniennes (Byrd

and Corces, 2003). Leur insertion ciblée dans un génome peut mener au repositionnement, en périphérie du noyau, de séquences d'ADN, ou encore au rapprochement dans l'espace du noyau de séquences initialement localisées dans des régions différentes (Gerasimova et al., 2000). Enfin, des domaines riches en séquences satellites, comme les régions centromériques de l'homme et de la souris, sont eux aussi connus pour s'associer et se localiser préférentiellement en certains endroits du noyau (Weierich et al., 2003; Ollion et al., 2015). Bien que cette idée manque encore d'arguments expérimentaux, les séquences répétées en tandem pourraient donc elles aussi jouer un rôle dans le positionnement spatial des chromosomes.

Expression des génomes

L'expression des génomes, liée aux éléments répétés via leur incidence sur l'état chromatinien et l'organisation nucléaire, peut également être impactée par des répétitions régulatrices modifiant localement la transcription des gènes, mais également la maturation des ARN ou encore leur traduction en protéines.

La transcription des gènes peut se voir influencée par des séquences répétées insérées et conservées dans des régions spécifiques du génome. Pour exemples, les régions enhancers du génome humain apparaissent particulièrement enrichies en SINE capables de fixer des facteurs de transcription et d'interagir avec les régions promotrices (Norris et al., 1995; Vasant and Reynolds, 1995; Zhou et al., 2000, 2002; Jjingo et al., 2014; Su et al., 2014). Un tel enhancer, dérivé de SINE, a également été décrit chez le coelacanthe où il régule le gène *ISL1* responsable de la différentiation des neurones moteurs (Bejerano et al., 2006). Par ailleurs, 25 % des régions promotrices humaines contiendraient des séquences dérivées d'éléments transposables (Jordan et al., 2003), et un SINE contenant un promoteur actif pour l'ARN polymérase II, et contrôlant ainsi l'expression d'un des gènes codant pour la laminine 5, a été identifié chez la souris (Ferrigno et al., 2001). En revanche, certains éléments transposables régulent négativement la transcription des gènes, comme les LINE introniques qui peuvent inhiber l'elongation de la transcription de leurs gènes hôtes chez l'homme (Han et al., 2004). La transcription des gènes peut également être influencée par la transcription des éléments transposables ; chez la souris, les transcrits des rétrotransposons VL30 sont capables de relâcher l'inhibition de l'expression de certains gènes en fixant les protéines responsables de cette répression (Song et al., 2004), et la transcription de rétrotransposons B2 permet de relâcher la répression de régions avoisinantes et donc d'activer l'expression de certains gènes (Lunyak et al., 2007). De plus, c'est la voie PIWI qui régulerait, chez cet organisme ainsi que chez le rat, la méthylation répressive de novo du gène soumis à empreinte parentale *Rasgrf1* (Watanabe et al., 2011).

Les mini- et microsatellites jouent eux aussi un rôle régulateur sur la transcription génique, notamment chez la drosophile où la présence de répétitions GA est indispensable au niveau de certains loci homéotiques pour leur fixation et silencing par des protéines du groupe Polycomb, et donc pour le développement de l'organisme (Hodgson et al., 2001). Ces répétitions sont

présentes dans les introns de nombreux autres gènes et y permettraient la fixation du facteur de transcription GAGA, l'initiation et l'elongation de la transcription (van Steensel et al., 2003). De telles répétitions GA dans des promoteurs géniques et leur fixation par des facteurs de transcription GAGA-like ont également été observées chez les plantes (Sangwan and O'Brian, 2002). Enfin, le rôle essentiel de répétitions en tandem sur la transcription de loci a également été montré chez les Mammifères, et notamment chez la souris où des répétitions CTAT sont impliquées dans le silencing type cellulaire-dépendant du gène codant pour la perforine, protéine exclusivement exprimée dans les cellules du système immunitaire et leur permettant d'éliminer les cellules non fonctionnelles de l'organisme (Youn et al., 2002).

La maturation et la traduction des ARN peuvent également être influencées par des éléments répétés. Chez le nématode *Caenorhabditis elegans*, les introns des gènes dont les transcrits sont soumis à splicing alternatif se voient enrichis en minisatellites qui réguleraient ce splicing (Glauser, 2014), tandis que certains SINE seraient capables de stimuler la traduction des ARNm chez l'homme et la souris (Rubin et al., 2002).

Conclusion

Les éléments répétés présents dans les génomes eucaryotes, bien que largement diversifiés en séquences et organisations, y adoptent certaines fonctions architecturales et régulatrices communes. L'implication des séquences satellites dans l'organisation spatiale et l'expression des génomes apparaît cependant moins avérée que celle des éléments transposables ; une caractérisation plus approfondie de régions riches en satellites, comme par exemple les régions centromériques des Mammifères, permettrait d'éclairer les raisons, si raison il y a, de leur maintien dans les génomes.

1.2 Organisation des régions centromériques chez les Mammifères

Une région chromosomique requise pour l'attachement du fuseau mitotique et la ségrégation des chromosomes, durant la division cellulaire, a été mise en évidence en 1882 chez la salamandre (Flemming, 1882) et appelée le centromère. Tous les chromosomes eucaryotes portent un centromère ; il représente le voire les sites d'assemblage du kinétochore, complexe protéique le liant aux microtubules du fuseau. Différentes stratégies d'organisation centromérique ont été adoptées au cours de l'évolution (Figure 2.1). Chez les Mammifères, les centromères correspondent à la constriction primaire des chromosomes métaphasiques et reposent très généralement sur des séquences d'ADN satellites (Steiner and Henikoff, 2015). Celles-ci s'étendent jusque dans les compartiments adjacents, appelés péricentromères ; ils participent également à la fonction

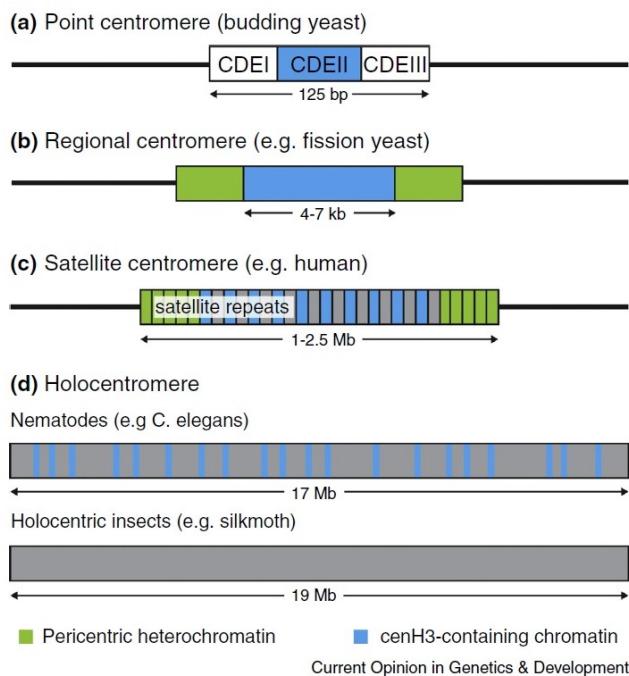


FIGURE 2.1 – Diversité des organisations centromériques (d’après Steiner and Henikoff (2015)).

centromérique en facilitant la cohésion des chromatides sœurs au cours du cycle cellulaire (Sumner, 1991; Bernard et al., 2001; Nonaka et al., 2002; Kiburz et al., 2005; Yamagishi et al., 2008; Sakuno and Watanabe, 2009). Les régions centromériques, i. e. les centromères et péricentromères, sont constituées d’éléments épigénétiques et génétiques spécifiques qui les distinguent du reste du génome.

1.2.1 Structure épigénétique des régions centromériques

La structure épigénétique qui définit les régions centromériques des Mammifères est très conservée à l’échelle de l’évolution. Deux domaines chromatiniens peuvent y être distingués : la chromatine dite centromérique (ou centrochromatine), et l’hétérochromatine des péricentromères.

Protéines centromériques

Les centromères sont définis par la présence de nucléosomes dans lesquels un variant de l’histone H3 canonique, appelé CenH3 ou CENP-A, s’associe avec les histones H2A, H2B et H4 (Sullivan et al., 1994). Un nucléosome sur 6 à 8 au centromère serait CenH3, et entouré de nucléosomes canoniques (Joglekar et al., 2008; Johnston et al., 2010; Fukagawa and Earnshaw, 2014). Le repliement de la chromatine centromérique et le positionnement des nucléosomes CenH3 par rapport au kinétochore restent controversés, tout comme la structure octamérique ou tétramé-

rique de ces nucléosomes *in vivo* (Blower et al., 2002; Dimitriadi et al., 2010; Ribeiro et al., 2010; Hasson et al., 2013; Fukagawa and Earnshaw, 2014). Il est intéressant de constater que la majorité des histones CenH3 peuvent être incorporées en dehors de la chromatine centromérique, comme 74 % d'entre elles dans les cellules RPE1 humaines (Bodor et al., 2014) ; mais leur concentration reste bien plus élevée aux centromères. Plusieurs autres protéines sont présentes au centromère et contribueraient à la formation de sa chromatine particulière, comme les protéines CENP-T, CENP-W, CENP-S et CENP-X qui forment ensemble un complexe tétramérique nucléosome-like capable de fixer et de replier l'ADN centromérique (Okada et al., 2006; Hori et al., 2008; Amano et al., 2009; Nishino et al., 2012; Takeuchi et al., 2014). La plupart des protéines centromériques n'auraient pas de motifs de fixation spécifiques sur l'ADN, à l'exception des protéines CENP-B et pJalpha dont les fonctions restent obscures (Masumoto et al., 1989; Gaff et al., 1994). CENP-B, très conservée chez les Mammifères, formerait un lien structural entre l'ADN centromérique et le kinétochore (Ohzeki et al., 2002; Suzuki et al., 2004; Foltz et al., 2006) ; toutefois, des kinétochères fonctionnels sont maintenus dans des souris ayant subi le knockout de CENP-B (Hudson et al., 1998; Kapoor et al., 1998; Perez-Castro et al., 1998).

L'hétérochromatine constitutive des péricentromères est caractérisée par la présence de la protéine HP1 (Heterochromatin protein 1) et de plusieurs histone méthyltransférases comme SUV39H ou SUV420H (Saksouk et al., 2015). Par ailleurs, les péricentromères seraient également enrichis en cohésine, ce qui les distinguerait des autres régions hétérochromatiques du génome (Gonzalez et al., 1991; Weber et al., 2004; Stephens et al., 2011; Bloom, 2014). La protéine shugoshine est également présente aux péricentromères, où elle aide au maintien de la cohésine au début de la mitose (Watanabe, 2005).

Marques chromatinianennes

La méthylation répressive de l'ADN est dense dans l'ensemble des régions centromériques, tandis que les marques épigénétiques des histones diffèrent entre centromères et péricentromères (Gopalakrishnan et al., 2009).

Les histones de la chromatine centromérique portent des marques épigénétiques qualifiées de permissives. D'une part la marque H4K20me1 y est apposée spécifiquement et est essentielle à la localisation de certaines protéines centromériques, comme CENP-H et CENP-T (Hori et al., 2014) ; cette marque fait partie des facteurs rendant possible l'assemblage du kinétochore sur la chromatine centromérique. La marque H3K4me2 est également caractéristique du centromère et lui permettrait d'assurer sa fonction ; en effet, la déplétion d'H3K4me2 au centromère l'inactive progressivement, en y prévenant l'incorporation de l'histone CenH3 (Bergmann et al., 2011).

L'hétérochromatine des péricentromères porte en revanche des marques épigénétiques dites répressives. Ses histones sont globalement hypoacétylées, ce qui permet sa forte compaction. Elle

est également enrichie en H3K9me2 et H3K9me3, marques que SUV39H y dépose spécifiquement. H3K9me3 permet la liaison de la protéine HP1 aux péricentromères, qui va à son tour y recruter SUV39H (Jenuwein, 2001; Peters et al., 2001). H3K9me3 et HP1 permettent par ailleurs l'apposition aux péricentromères des marques H4K20me2 et H4K20me3 par SUV420H, qui ne seraient cependant pas essentielles à l'état hétérochromatique de ces régions (Saksouk et al., 2015). La marque H3K27me3 peut également être retrouvée dans l'hétérochromatine des péricentromères, mais est généralement plus dense dans l'hétérochromatine facultative (Sullivan and Karpen, 2004; Saksouk et al., 2015).

1.2.2 ADN des régions centromériques

Les Mammifères se séparent en deux sous-classes : les Theria, qui comprennent les Eutheria (mammifères placentaires) et les Metatheria (marsupiaux), et les Prototheria, qui comprennent les monotrèmes (échidnés et ornithorynques). La majorité des centromères des mammifères Theria étudiés jusqu'ici sont constitués de séquences d'ADN satellites ; nous discuterons ici de la diversité des séquences satellites centromériques au sein de ce clade et de leur rôle potentiel, les tentatives d'isolement de telles séquences chez les mammifères Prototheria étant jusqu'ici restées sans succès (Alkan et al., 2011).

Diversité des séquences satellites centromériques

Bien que la fonction des régions centromériques soit très conservée, l'ADN de ces régions peut varier considérablement entre espèces, tant au niveau de sa quantité génomique, de la taille de ses unités de répétition, de leur séquence ou encore de leur organisation. Cette singularité est connue sous le nom de paradoxe du centromère et pourrait se voir expliquée par un phénomène appelé la dérive centromérique (Figure 2.2) (Henikoff et al., 2001; Malik and Henikoff, 2002) ; pour exemple, la taille des unités de répétition de l'ADN centromérique varie de 7 pb, chez le wallaby à cou rouge, à 3,2 kb, chez la vache (Giannuzzi et al., 2012). Néanmoins, le motif de fixation de la protéine CENP-B (\approx 17 pb), appelé la CENP-B box, est largement retrouvé dans les régions centromériques des Mammifères. Nous comparerons succinctement les satellites centromériques de cinq espèces plus ou moins éloignées phylogénétiquement : le wallaby, le mouton, la vache, la souris et l'homme (Figure 2.3).

Au sein des Metatheria, les espèces de la famille des Macropodidae (kangourous, wallabies et wallaroos) présentent une grande diversité caryotypique (Rofe, 1978; Hayman, 1990; Eldridge and Close, 1993). Le centromère a joué un rôle essentiel dans cette diversification, puisqu'il a été le point de cassure des réarrangements chromosomiques majeurs ayant eu lieu au sein de ce clade (Rens et al., 2004; O'Neill et al., 2004). De manière exceptionnelle, les péricentromères du wallaby à cou rouge (*Macropus rufogriseus*) sont particulièrement étendus et peuvent

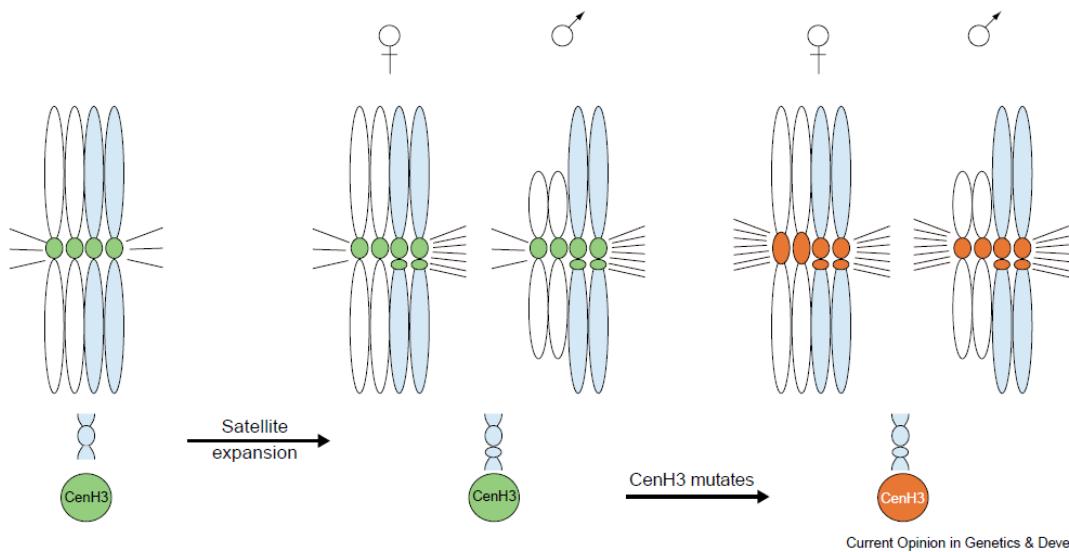


FIGURE 2.2 – Modèle de la dérive centromérique (d’après Malik and Henikoff (2002)). L’amplification de nouvelles séquences satellites sur un centromère pourrait faciliter l’incorporation d’un plus grand nombre d’histones CenH3, et la liaison à plus de microtubules durant la division cellulaire. Lors de la méiose asymétrique femelle, le chromosome portant ce caractère pourrait être préférentiellement transmis à la cellule œuf, et donc les nouvelles séquences satellites positivement sélectionnées. Cependant, la disparité de tension engendrée entre les chromosomes homologues lors de la division cellulaire peut s’avérer délétère pour la méiose mâle, plus sensible à celle-ci que la méiose femelle. De nouvelles histones CenH3 capables de supprimer cette disparité et de restaurer la complète fertilité des mâles seraient donc à leur tour positivement sélectionnées, etc.

comprendre jusqu’à la moitié des chromosomes ; les régions centromériques correspondent ainsi à 30 % de son génome (Bulazel et al., 2006). Trois familles de séquences satellites y ont été caractérisées : Mrb-sat1, Mrb-sat23 et Mrb-B29 (Figure 2.4). Elles présentent différentes distributions chromosomiques, puisque Mrb-sat23 est localisée aux centromères et péricentromères de tous les chromosomes, tandis que Mrb-sat1 et Mrb-B29 ne sont majoritairement visibles que sur les chromosomes sexuels. Contrairement à Mrb-sat1 et Mrb-B29, les séquences Mrb-sat23 contiennent une CENP-B box. Par ailleurs, le rétrotransposon KERV (Kangaroo Endogenous Retrovirus) se trouve également être un constituant majeur des régions centromériques des Macropodidae. Chez le wallaby, cet élément est présent dans les centromères et péricentromères de tous les autosomes (Ferreri et al., 2004, 2011).

La vache (*Bos taurus*) et le mouton (*Ovis aries*) sont tous deux des Eutheria de la famille des Bovidae. Ils partageraient deux familles de séquences satellites, les satellites I et II, qui n’adoptent cependant pas des caractéristiques identiques chez les deux espèces. Le satellite I par exemple divergerait d’une séquence de 31 pb qui s’est amplifiée en unités de 1,4 kb chez la vache et de 820 pb chez le mouton (Kurnit et al., 1973; Buckland, 1983; Burkin et al., 1996). Deux familles de séquences satellites supplémentaires ont été identifiées chez la vache, les satellites III et IV. Le satellite III est formé d’unités de répétitions comprenant deux séquences de 23 pb, appelées motifs Pvu et Sau, qui auraient elles aussi divergé de la séquence ancestrale du satellite I (Pech et al., 1979) ; les satellites I et III se regrouperaient donc dans une même

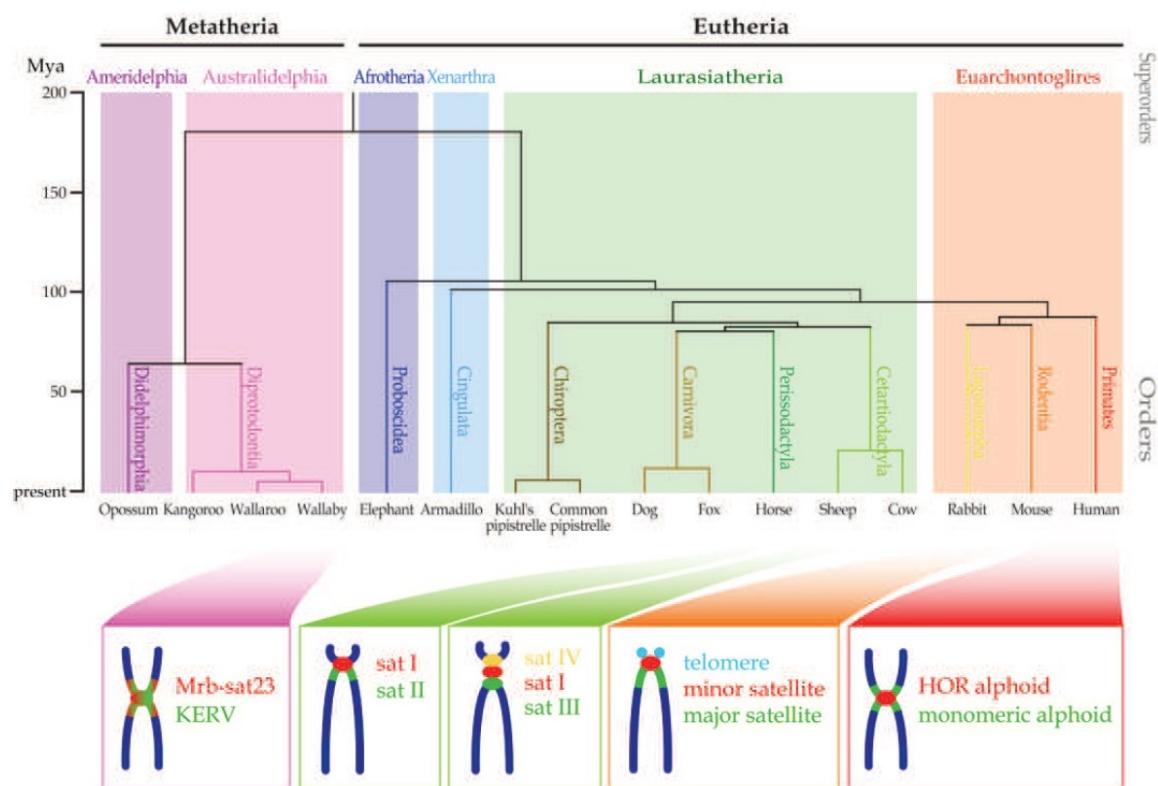


FIGURE 2.3 – Exemples de satellites centromériques présents chez les Mammifères (d'après Gianuzzi et al. (2012)). En haut : Relations phylogénétiques entre mammifères. En bas : Distribution de différentes familles de satellites sur les régions centromériques des chromosomes du wallaby, du mouton, de la vache, de la souris et de l'homme.

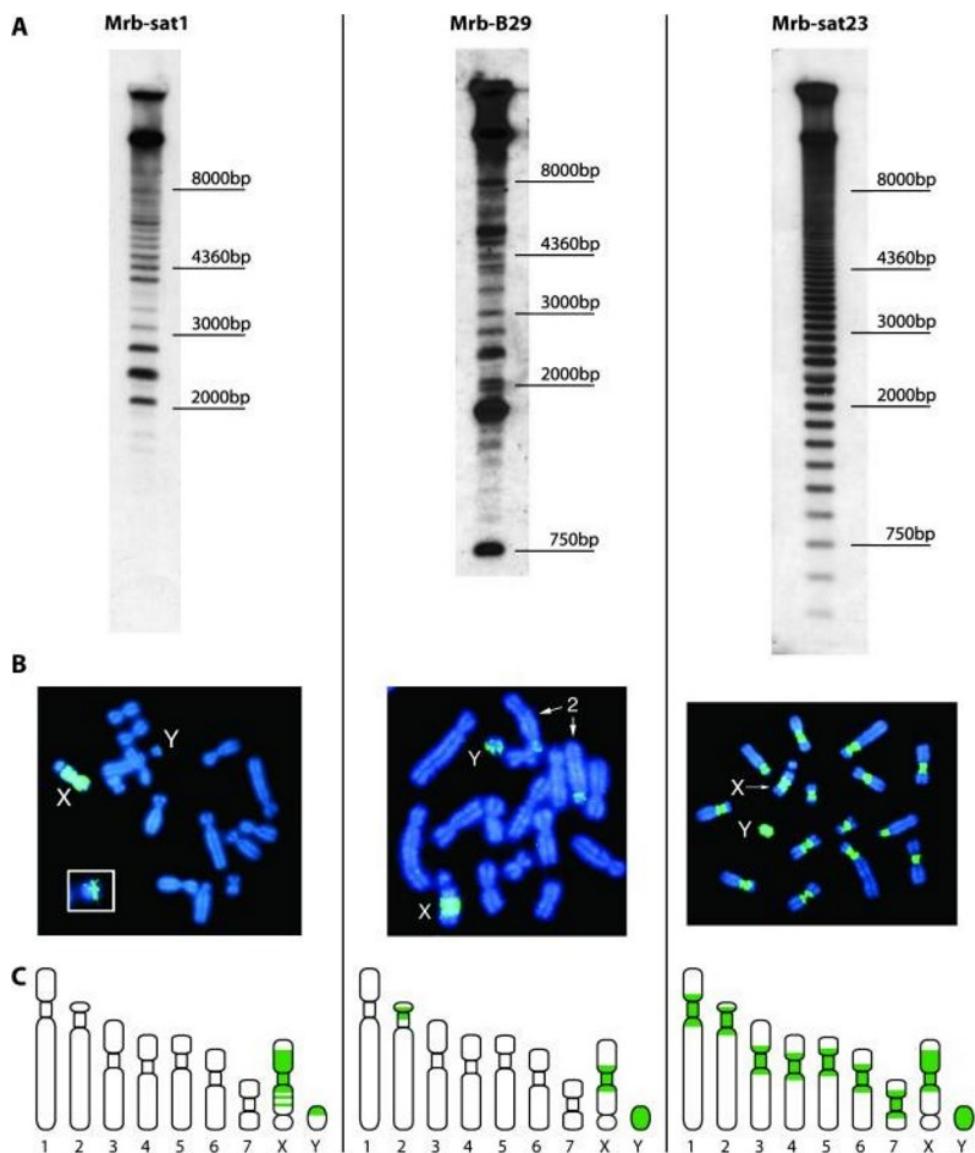


FIGURE 2.4 – Caractérisation des satellites centromériques du wallaby à cou rouge (d'après [Bulazel et al. \(2006\)](#)). (A) Southern-blot : hybridation de clones Mrb-sat1, Mrb-B29, et Mrb-sat23 sur de l'ADN génomique digéré par l'enzyme BglII. (B) Hybridation in situ en fluorescence de ces mêmes clones sur les chromosomes métaphasiques du wallaby. Fenêtre Mrb-sat1 : chromosome Y (exposition prolongée). (C) Représentation schématique de la distribution chromosomique des trois satellites.

famille de séquences homologues. Chez la vache, les satellites I à IV sont localisés aux régions centromériques des autosomes et adoptent différentes distributions ; en revanche, ils sont absents des chromosomes sexuels (Kopecka et al., 1978; Kurnit et al., 1973). Lorsque les satellites I, III et IV sont présents sur la même région centromérique, ils s'organisent en suivant toujours un certain ordre : p -satIV- satI-satIII- q (Chaves et al., 2003). Chez le mouton, les satellites I et II se localisent respectivement aux centromères et péricentromères (Burkin et al., 1996; D'aiuto et al., 1997). Le satellite I se distribue sur tous les autosomes, en quantité variant légèrement d'un chromosome à l'autre. Le satellite II, qui se distribue également sur les autosomes mais aussi sur le chromosome X, présente des variations de quantité plus marquées puisqu'il est absent sur certains acrocentriques et particulièrement abondant sur les métacentriques et le chromosome X (Burkin et al., 1996). Aucune CENP-B box n'a encore été détectée chez les Bovidae.

Bien que les séquences satellites centromériques de plusieurs rongeurs aient été caractérisées, ce sont celles de la souris (*Mus musculus*, famille des Muridae) qui ont été les plus largement étudiées. Son caryotype est composé exclusivement de chromosomes acrocentriques dont les régions centromériques portent deux familles de séquences satellites, appelées satellite majeur et satellite mineur (Pardue and Gall, 1970; Wong and Rattner, 1988)(Figure 2.5). Le satellite mineur est formé d'unités de répétition de 120 pb, AT-riches et possédant la CENP-B box (Yoda et al., 1996). Il représente 0,5 à 1 % du génome de la souris, et occupe 300 à 600 kb sur la région terminale des bras courts des autosomes et du chromosome X ; cette région correspond au centromère (Kipling et al., 1991). Le satellite majeur, plus abondant, représente 5 à 10 % du génome. Il présente une organisation dite en répétitions d'ordre supérieur ou HORs (Higher order repeats) puisque ses répétitions constitutives de 234 pb sont, elles-mêmes, formées de répétitions internes (Figure 2.6). Lui aussi AT-riche, il occupe les péricentromères des autosomes et du chromosome X, sur des régions allant de 240 à 2 000 kb (Vissel and Choo, 1989). Les satellites mineur et majeur sont considérés comme des séquences très homogènes, montrant peu de divergence interchromosomique avec à peine 5 % de variabilité globale (Vissel and Choo, 1989). Toutefois, plusieurs variants chromosomiques du satellite mineur ont été caractérisés par FISH (Hybridation in situ en fluorescence) dans les années 1990 (Broccoli et al., 1991; Hayashi et al., 1993; Kipling et al., 1994) ; de plus, l'analyse informatique récente de données de séquençage de génomes entiers suggère une diversité des séquences satellites majeures bien plus élevée qu'il n'était pensé jusqu'alors, puisque celles-ci présenteraient en moyenne 15 % de variabilité entre elles (Komissarov et al., 2011). En plus des satellites mineur et majeur, les régions centromériques de la souris portent deux familles de séquences satellites, cette fois GC-riches : MS3 (Mouse satellite 3 ; 2,2 % du génome) et MS4 (Mouse satellite 4 ; 1,6 % du génome). Leurs unités de répétitions sont respectivement de 150 et 300 pb, et possèdent la CENP-B box. Les séquences MS3 seraient extrêmement homogènes, avec 0,7 % de variabilité seulement (Kuznetsova et al., 2005). Enfin, le chromosome Y possède une famille de séquences satellites appelée Ymin ; elles sont homologues et similaires à 77 % aux séquences satellites mineures, et s'organisent en HORs sur 90 kb (Pertile et al., 2009).

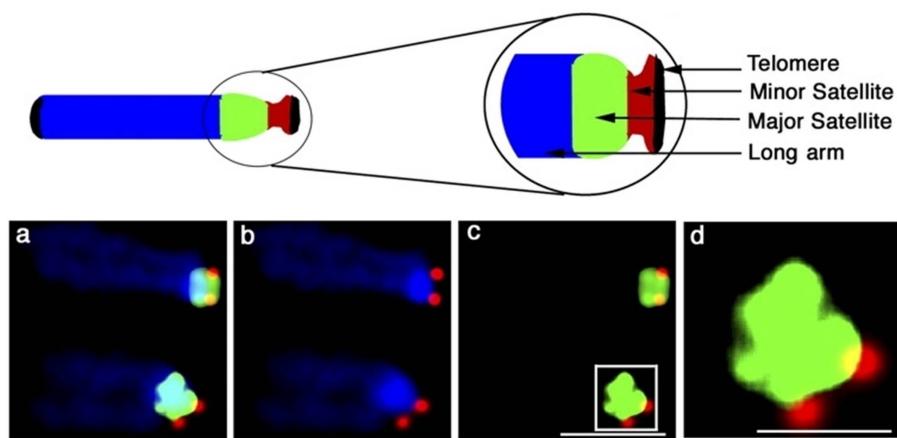


FIGURE 2.5 – Distribution des satellites majeur et mineur sur les régions centromériques de la souris (d’après Guenatri et al. (2004)). En haut : schéma d’un chromosome acrocentrique de souris. En bas : hybridation in situ en fluorescence du satellite mineur (rouge) et du satellite majeur (vert) sur deux chromosomes métaphasiques (bleu). Barre d’échelle : 2 µm.

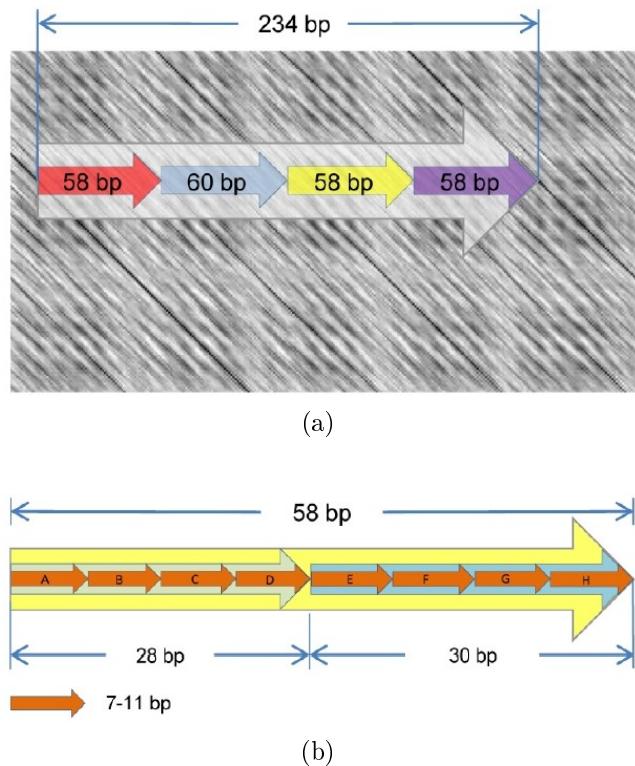


FIGURE 2.6 – Organisation en HORs du satellite majeur de la souris (d’après Komissarov et al. (2011)). (A) Structure de l’unité tétramérique de 234 pb du satellite majeur. (B) Structure des monomères de 58 pb de cette unité tétramérique.

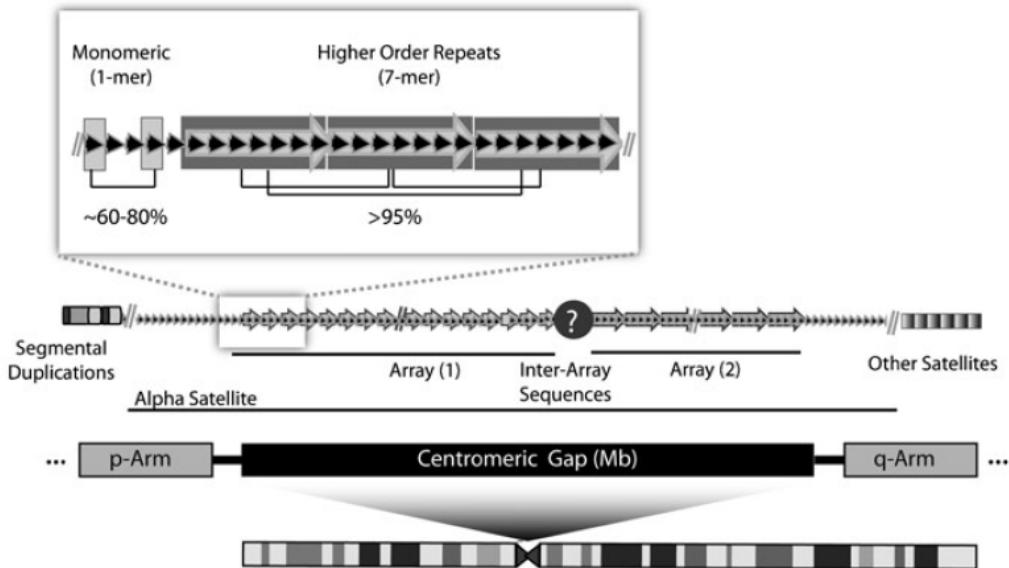


FIGURE 2.7 – Modèle d’organisation des régions centromériques humaines (d’après Hayden (2012)).

L’homme est un Eutheria de la famille des Hominidae (homme, chimpanzés, gorilles et orangs-outans). Ses régions centromériques sont constituées d’une famille prédominante de séquences satellites, appelée ADN alpha satellite, qui comprend 3 % du génome humain (Manuelidis, 1978; Hayden et al., 2013). Cette famille est formée d’unités AT-riches d’environ 171 pb, également appelées monomères d’alpha satellites, qui se répètent sur 250 à 5 000 kb (Wevrick and Willard, 1989; Willard, 1990). Au centromère, l’ADN alpha satellite s’organise en HORs dont les unités de répétition peuvent comprendre plusieurs dizaines de monomères ; aux péricentromères, les monomères d’alpha satellites ne seraient généralement pas impliqués dans de telles organisations (Figure 2.7). Les HORs centromériques humaines diffèrent entre chromosomes, mais chacune d’entre elles est homogène ; les unités de répétitions multimériques d’une même HOR sont réputées très conservées, avec en moyenne 98 % d’identité de séquence entre elles ; elles peuvent tout de même être parfois plus divergentes, et n’adopter qu’en moyenne 94 % d’identité de séquence (Rudd and Willard, 2004). Ces unités peuvent également, bien que rarement, différer entre elles dans le nombre de leurs monomères constitutifs ou encore être interrompues par des éléments transposables (Figure 2.8). Les monomères constitutifs de chaque unité sont plus divergents entre eux que les unités entre elles ; ils ne partagent que 72 % d’identité de séquence moyenne, tout comme, en général, les monomères des péricentromères dans lesquels sont de plus dispersés de nombreux éléments transposables (Guy et al., 2003; Kazakov et al., 2003; Schueler et al., 2001; Rudd and Willard, 2004). Enfin, il peut être noté que les régions centromériques humaines sont entourées d’autres familles de séquences satellites que les alpha, comme les beta et gamma satellites, ainsi que les satellites 1, 2, 3 et 4 humains (Lee et al., 1997; Cardone et al., 2004; Schueler et al., 2005; Kim et al., 2009; Aldrup-MacDonald and Sullivan, 2014; Altemose et al., 2014). Des copies de ces séquences pourraient toutefois se retrouver intégrées au sein des régions centromériques, entre des blocs de séquences alpha satellites (Malik and Henikoff, 2002; Hayden, 2012).

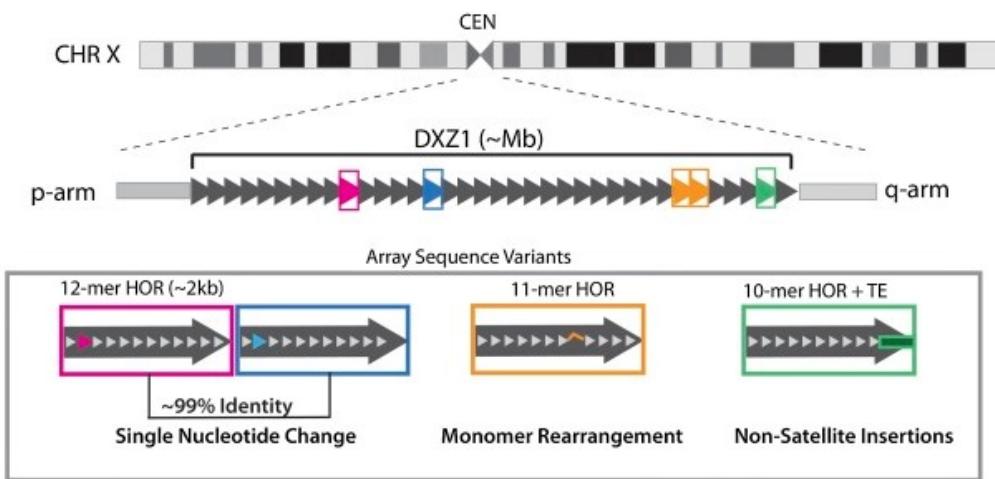


FIGURE 2.8 – Organisation en HORs du chromosome X humain (d’après Miga et al. (2014)).

Séquences satellites et fonction centromérique

Au cours de l’évolution se produisent des phénomènes de relocalisation centromérique, où un centromère s’inactive au profit de l’émergence d’un autre dans une région chromosomique ectopique (Montefalcone et al., 1999; Ventura et al., 2007). Les nouveaux centromères évolutifs issus de tels phénomènes se formeraient dans des régions dénuées de séquences satellites, et pourraient en rester dépourvus pendant des millions d’années (Amor and Choo, 2002; Piras et al., 2010). Ainsi, plusieurs nouveaux centromères évolutifs n’ayant pas encore acquis de séquences satellites ont été caractérisés chez les équidés, ainsi que chez l’orang-outan (Wade et al., 2009; Piras et al., 2010; Locke et al., 2011). De nombreux néocentromères se formant d’une telle manière sur des chromosomes humains ont été observés lors d’études cliniques ; ils sont eux aussi dépourvus de séquences satellites (Marshall et al., 2008).

Les séquences satellites n’apparaissent donc pas comme des éléments indispensables à la fonction du centromère ; cependant, la colonisation progressive des centromères par des séquences satellites semble être inéluctable, et pourrait donc se révéler essentielle au maintien à long terme de ces structures. En effet, comme révélé par l’étude d’un même nouveau centromère évolutif chez différents individus d’une espèce d’équidé (Purgato et al., 2015), la position des centromères ne serait pas fixe et pourrait varier légèrement autour d’un locus. Or le déplacement, même léger, d’un centromère dans une région portant des gènes pourrait affecter leur expression, et nuire ainsi à la survie des individus ; en revanche, le déplacement d’un centromère dans une région pourvue de séquences satellites aurait moins de risques d’affecter un gène. Les séquences satellites pourraient donc être un environnement propice à la mobilité des centromères, et l’hétérochromatine péricentromérique être une barrière entre les centromères et les régions euchromatiques adjacentes (Figure 2.9).

Par ailleurs, les séquences satellites centromériques pourraient participer à la fonction du cen-

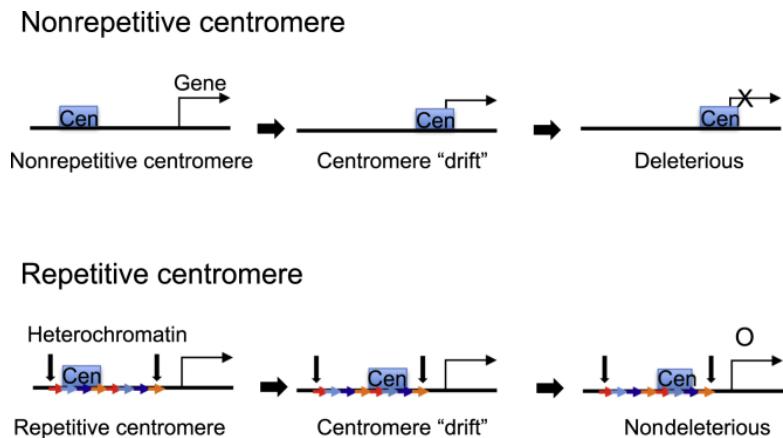


FIGURE 2.9 – Hypothèse quant au rôle des séquences répétées au centromère (d’après Fukagawa and Earnshaw (2014)).

tromère via leurs transcrits, dont l’existence a été reportée dès les années 1960 chez la souris (Harel et al., 1968; Cohen et al., 1973). En effet, ces ARN satellites sont capables de s’associer chez l’homme à des protéines telles que CENP-C, et permettraient le recrutement direct ou indirect de plusieurs autres protéines au centromère, comme INCENP ou Aurora B, assurant ainsi la ségrégation correcte des chromosomes lors de la division cellulaire (Wong et al., 2007; Chan et al., 2012). Chez la souris, les transcrits centromériques permettraient même le recrutement de l’histone CenH3 (Ferri et al., 2009).

Enfin, il a été montré que la topoisomérase II, protéine influençant la topologie de l’ADN en y créant des cassures double brin, reconnaissait et clivait spécifiquement des structures en épingle à cheveux formées dans les satellites centromériques humains (Jonstrup et al., 2008). Cette protéine ayant un rôle clef dans la séparation des chromatides sœurs au cours de l’anaphase (Clarke et al., 1993; Dawlaty et al., 2008), de telles structures secondaires pourraient être impliquées dans son recrutement et permettre ainsi l’achèvement de la division cellulaire. On peut également envisager que le repliement des séquences satellites permette le recrutement de protéines constitutives de la chromatine des régions centromériques (Benfante et al., 1990; Bigot et al., 1990).

Conclusion

Les régions centromériques des Mammifères sont, dans leur grande majorité, enrichies en séquences d’ADN satellites qui nourrissent différentes hypothèses fonctionnelles. Celles-ci se distribuent en de nombreuses familles de séquences homologues présentant des organisations variées ; cependant, la diversité au sein même de ces familles reste généralement peu regardée. L’ADN alpha satellite des régions centromériques humaines, également retrouvé chez les autres primates, représente un modèle d’étude intéressant pour mieux comprendre la dynamique évolutive de ces séquences particulières.

1.3 Dynamique évolutive de l'ADN alpha satellite chez les Primates

L'ordre des Primates comprend deux sous-ordres, les Strepsirrhini et les Haplorrhini (Figure 3.1). Les Strepsirrhini se divisent en trois infra-ordres comprenant les lémuriens et microcèbes, les loris et galagos, et l'aye-aye ([Wilson and Reeder, 2005](#)). Les Haplorrhini comprennent quant à eux deux infra-ordres : les tarsiers ou Tarsiiformes d'une part, et les singes ou Simiiformes d'autre part. Les Simiiformes se divisent en deux micro-ordres, les Platyrrhini et les Catarrhini. Chez les Platyrrhini, également appelés singes du Nouveau Monde, on retrouve par exemple les saïmiris, les capucins, les atèles, les marmosets ou encore les sakis. Chez les Catarrhini, on retrouve deux superfamilles : les Cercopithecoidea et les Hominoidea. Les Cercopithecoidea ne comprennent qu'une famille actuelle, les Cercopithecidae ou singes de l'Ancien Monde, qui se divise en deux sous-familles : les Cercopithecinae où l'on retrouve les cercopithèques, les macaques et les babouins, et les Colobinae où l'on retrouve les colobes, les nasiques et les rhinopithèques. Les Hominoidea comprennent quant à eux les gibbons ou Hylobatidae, et les grands singes ou Hominidae. La famille des Hominidae se divise en deux sous-familles : les Ponginae où l'on retrouve les orangs-outans, et les Homininae où l'on retrouve les gorilles, les chimpanzés et l'homme.

L'ADN alpha satellite, abordé dans la section précédente, ne se retrouve que chez les Primates. Il a été découvert dans les années 1970, grâce à la centrifugation en gradient de chlorure de césium du génome de *Chlorocebus aethiops*, une espèce de la famille des Cercopithecidae et de la tribu des Cercopithecini ([Maio, 1971; Kurnit and Maio, 1974](#)). Ce composant a par la suite été identifié dans les régions centromériques de plusieurs autres Catarrhini, Cercopithecoidea comme Hominoidea, dont l'homme. Il a également été retrouvé chez les Platyrrhini ; en revanche, les quelques études menées sur les Strepsirrhini n'ont pas révélé la présence de séquences alpha satellites dans leurs centromères, mais plutôt celle d'autres familles de satellites ([Musich et al., 1980; Maio et al., 1981; Lee et al., 2011](#)). Les régions centromériques des Tarsiiformes n'ont, quant à elles, été jusqu'ici l'objet d'aucune étude.

1.3.1 Diversité d'organisations et de séquences

La digestion enzymatique, le clonage, le séquençage et l'étude par hybridation *in situ* de nombreux fragments d'ADN alpha satellite, voire plus occasionnellement l'analyse de données de séquençage à haut débit ([Hayden et al., 2013; Miga et al., 2014](#)), ont permis d'apprécier la diversité des séquences alpha satellites dans les génomes de certains primates. En effet, l'amplification des séquences alpha satellites peut mener à différentes organisations dans l'ADN alpha satellite, selon que cette amplification soit monomérique ou multimérique. Par ailleurs, les unités de répétition de l'ADN alpha satellite, homologues et similaires, ne sont toutefois pas

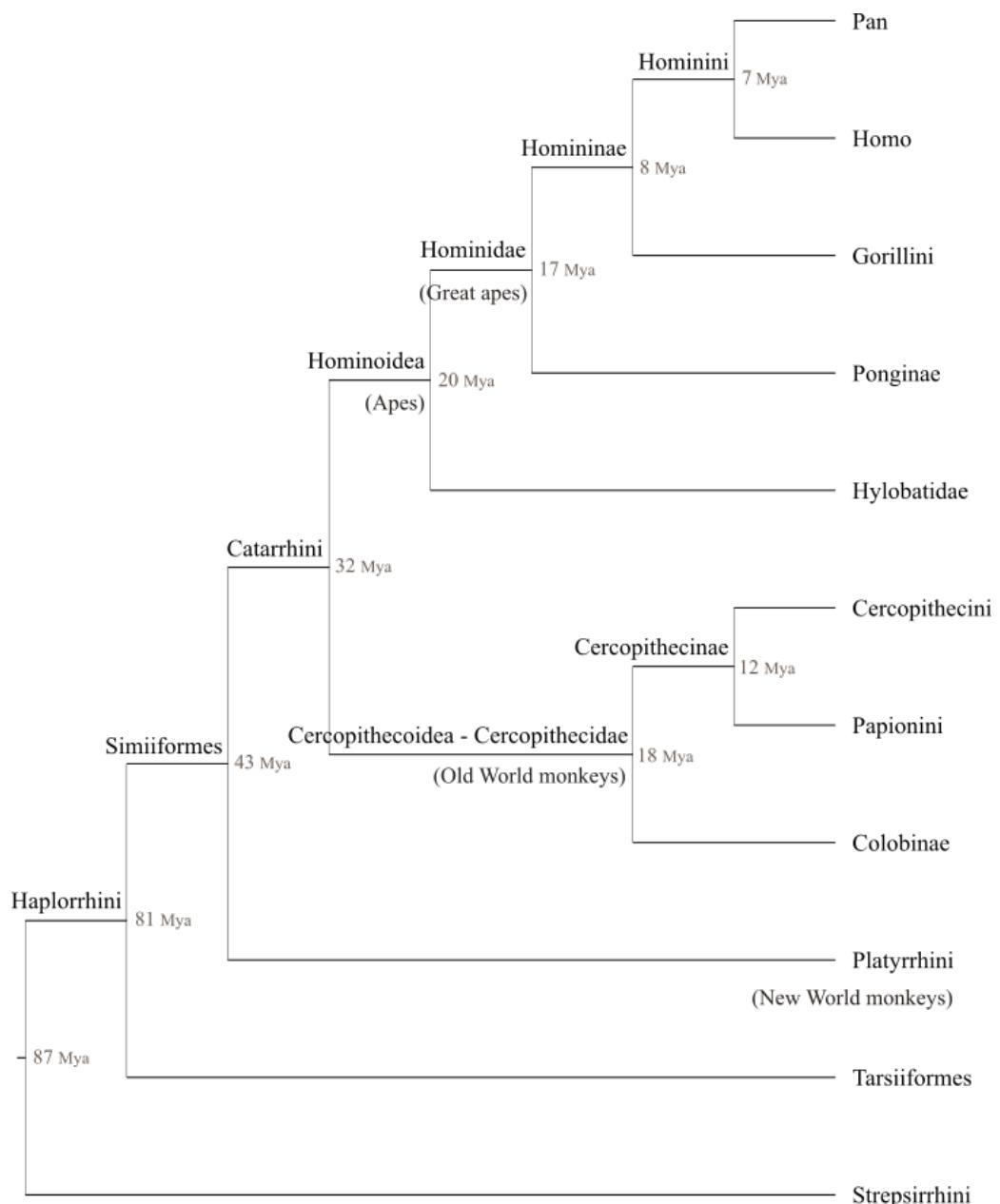


FIGURE 3.1 – Phylogénie des Primates (d'après Arnold et al. (2010) et Perelman et al. (2011)).

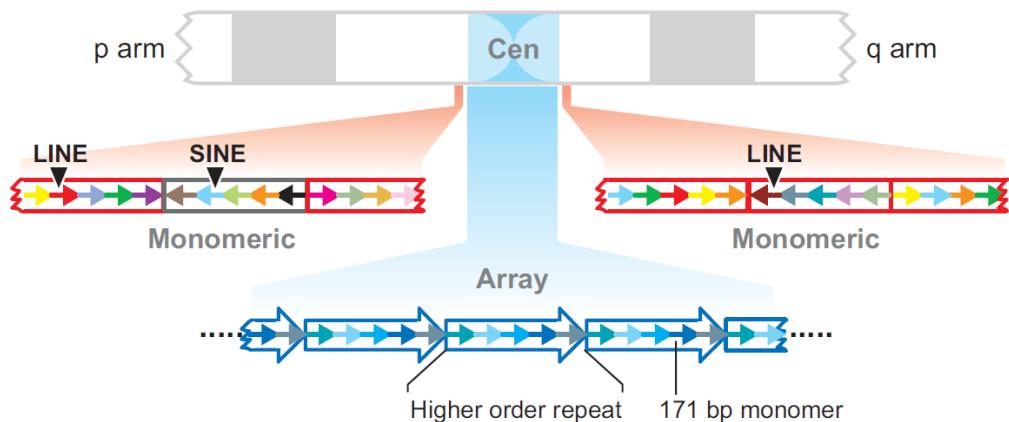


FIGURE 3.2 – Organisation de l’ADN alpha satellite des régions centromériques humaines (d’après Schueler and Sullivan (2006)).

identiques entre elles ; différentes familles de séquences alpha satellites peuvent se distinguer dans et entre les génomes.

Organisation monomérique ou en HORs

Comme mentionné précédemment, deux types d’organisation peuvent être observés dans l’ADN alpha satellite : une organisation dite en HORs, issue de l’amplification d’une unité comprenant plusieurs monomères, et une organisation dite monomérique, issue de l’amplification d’un monomère unique (Figure 3.2). Tandis que des HORs d’ordres 2 et 3, i. e. dont l’unité d’amplification comprend 2 ou 3 monomères, ont été identifiées chez les singes du Nouveau et de l’Ancien Monde dans les années 1980 et 1990 (Donehower et al., 1980; Rubin et al., 1980; Pike et al., 1986; Alves et al., 1998), des HORs d’ordres allant de 5 à 36 étaient caractérisées dans le même temps chez l’homme (Tyler-Smith and Brown, 1987; Alexandrov et al., 2001). De telles HORs complexes ont pendant longtemps été considérées comme l’apanage des grands singes (Waye and Willard, 1986; Willard and Waye, 1987; Wevrick and Willard, 1989; Warburton et al., 1996; Alexandrov et al., 2001; Cellamare et al., 2009; Catacchio et al., 2015) ; cependant, des HORs d’ordres 4, 6 et 13 ont récemment été caractérisées chez les Hylobatidae (Terada et al., 2013; Koga et al., 2014), ainsi que des HORs d’ordres 9 et 12 chez les singes du Nouveau Monde (Sujiwattanarat et al., 2015). L’amplification de l’ADN alpha satellite en HORs complexes serait donc un évènement commun chez les Primates, bien que de telles organisations n’aient pour le moment été mises en évidence chez aucun singe de l’Ancien Monde ; certaines espèces de ce clade, comme le cercopithèque *C. aethiops*, n’auraient même aucune organisation en HORs, pas même d’ordre 2 ou 3 (Goldberg et al., 1996; Yoda et al., 1996; Alexandrov et al., 2001) (Figure 3.3).

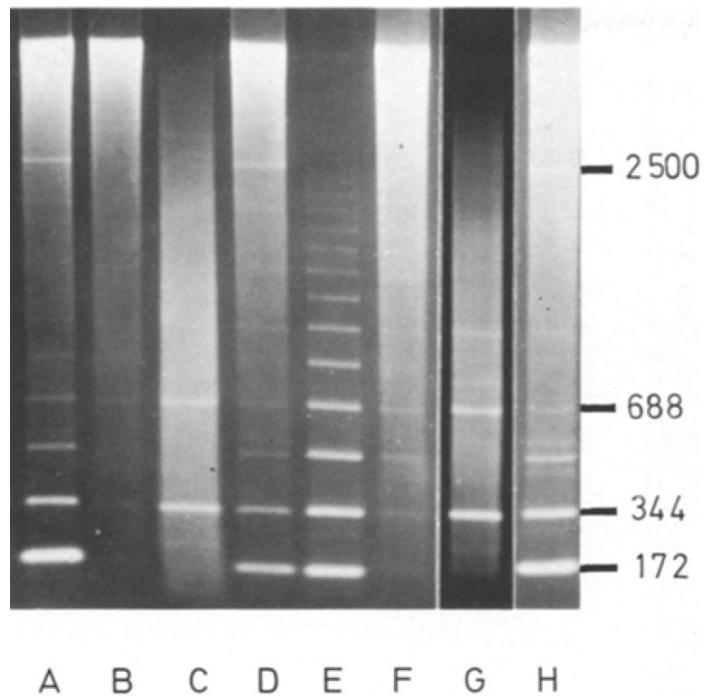


FIGURE 3.3 – Profil de migration de l'ADN alpha satellite de singes de l'Ancien Monde (d'après Musich et al. (1980)). De l'ADN génomique de cercopithèque (*C. aethiops*, A, E), de macaque (*Macaca mulatta*, B, C, D) ou de babouin (*Papio sp.*, F, G, H) est digéré par l'enzyme de restriction HindIII (A, D, H) ou EcoRI (B, F et C, E, G - tampon enrichi en magnésium). La migration sur gel des produits de digestion montre un profil en échelle caractéristique de l'ADN alpha satellite, avec la révélation de différentes bandes à environ $n \times 170$ pb. Lorsque digéré par EcoRI, l'ADN du macaque et du babouin (C, G) peut présenter des bandes majoritaires à environ $n \times 340$ pb, témoignant d'une organisation en HORs d'ordre au moins 2 dans leurs génomes. Ce profil n'est pas retrouvé chez le cercopithèque (E).

TABLE 3.1 – Diversité des familles d’alpha satellites chez les Primates (d’après Alexandrov et al. (1993), Alves et al. (1994, 1998), Romanova et al. (1996), Alexandrov et al. (2001), Shepelev et al. (2009), Prakhongcheep et al. (2013) et Catacchio et al. (2015)).

Famille	NWM	OWM	Hominoidae	Hominidae	Homininae	Taille (pb)	Type	Qualificatif	SF
S1		x				172	A	-	-
S2		x				171	A	-	-
S3	x					172	A	-	-
S4	x					171	A	-	-
S5	x					196	A	-	-
OwlAlp1	x					186	Nd	-	-
J1				x		171	A	New	1
J2					x	169	B	New	1
D1					x	168-189	B	New	2
D2					x	171	A	New	2
W1					x	167	B	New	3
W2					x	171	B	New	3
W3					x	171	B	New	3
W4					x	169-171	A	New	3
W5					x	171	A	New	3
R1			x		x	171	B	Old	5
R2			x		x	171	A	Old	5
M1		x		x	x	171	A	Old	4
V1		x		x	x	172	A	Old	-
H1	x		x	x	x	172	A	Ancient	-
H2	x		x	x	x	172	A	Ancient	-
H3	x		x	x	x	172	A	Ancient	-
H4	x	x	x	x	x	172	A	Ancient	-

NOTE - NWM = New World monkeys, singes du Nouveau Monde; OWM = Old World monkeys, singes de l’Ancien Monde; Type A = pJalpha box; Type B = CENP-B box; SF = Famille suprachromosomique; Nd = Non déterminé; New = Nouvelle; Old = Agée; Ancient = Ancienne.

Familles de séquences alpha satellites

Au cours de l’évolution, des mutations font diverger les monomères d’alpha satellites entre eux quant à leur taille et leur séquence. L’amplification subséquente de certains variants mène à l’émergence de différentes familles de séquences au sein de l’ADN alpha satellite, qui portent chacune leurs spécificités (Table 3.1).

Chez les singes de l’Ancien Monde, deux familles d’alpha satellites ont été initialement identifiées et nommées S1 et S2 (Donehower et al., 1980; Rubin et al., 1980; Pike et al., 1986; Prassolov et al., 1986; Alexandrov et al., 2001; Alkan et al., 2007); leur caractérisation repose cependant sur la comparaison de quelques dizaines de monomères seulement, et doit donc être considérée avec précaution. Les monomères de la famille S1 ont une taille moyenne de 172 pb et adoptent une organisation monomérique chez le cercopithèque *C. aethiops*, dont le génome

ne porterait pas d'autres familles d'alpha satellites. Ses séquences alpha satellites n'auraient pas plus de 1 à 5 % de variabilité entre elles (Goldberg et al., 1996; Yoda et al., 1996). Chez les macaques, babouins et colobes (*Macaca fuscata*, *M. radiata*, *M. fascicularis*, *M. mulatta*, *Papio hamadryas* et *Colobus guereza*), les monomères de la famille S1 s'organisent en HORs d'ordre 2 avec les monomères de la famille S2, qui ont quant à eux une taille moyenne de 171 pb. Les monomères S1 et S2 présentent 30 à 45 % de variabilité entre eux, tandis que les unités d'amplification dimériques S1-S2 ne varient entre elles que de 2 à 5 % (Alkan et al., 2007).

Chez les singes du Nouveau Monde, trois familles d'alpha satellites ont été initialement identifiées et nommées S3, S4 et S5, toujours d'après l'analyse de quelques monomères (Alves et al., 1994, 1998; Alexandrov et al., 2001). Les monomères des familles S3 et S4, d'une taille moyenne respective de 172 et 171 pb, s'associent en HORs d'ordre 2 dans le génome des capucins, titis, ouistiti, ouakaris et sakis (*Cebus apella*, *Callicebus moloch*, *Cebuella pygmaea*, *Cacajao melanocephalus*, *Chiropotes satanas* et *Pithecia irrorata*). Ces deux familles s'associent par ailleurs en HORs d'ordre 3 avec la famille S5 chez les sakis. Les monomères de la famille S5 sont des chimères entre monomères S3 et S4 ; ils résulteraient d'un crossing over inégal entre deux dimères S3-S4. Ils sont plus longs qu'usuellement puisque présentant une taille moyenne de 196 pb. Plus récemment, une quatrième famille d'alpha satellites a été spécifiquement caractérisée dans le génome du douroucouli d'Azara (*Aotus azarae*), et nommée OwlAlp1 (Prakhongcheep et al., 2013). La taille moyenne de ses monomères constitutifs est de 186 pb.

Des efforts particuliers de séquençage et d'analyse de l'ADN alpha satellite des grands singes ont permis de mettre en évidence de manière fiable plus d'une dizaine de familles d'alpha satellites, dont certaines seraient retrouvées jusque chez les singes du Nouveau Monde. Ainsi, 9 familles d'alpha satellites qualifiées de nouvelles (new) s'associent en HORs chez l'homme : J1 avec J2, D1 avec D2, et W1 avec W2, W3, W4 et W5 (Alexandrov et al., 1988, 1993). On ne les retrouverait que chez les Homininae. Chez le gorille, la famille D1 se divise en 5 sous-familles (gD1.0 à gD1.4) et la famille D2 se divise en 3 sous-familles (gD2.0 à gD2.2) (Catacchio et al., 2015). 4 familles d'alpha satellites qualifiées d'âgées (old) et ne s'associant généralement pas en HORs sont ensuite retrouvées dans le génome humain : R1 et R2, s'arrangeant de manière irrégulière dans les mêmes domaines génomiques et retrouvées chez les Hominidae, ainsi que M1 et V1, que l'on retrouve chez les Hominoidea (Figure 3.4) (Romanova et al., 1996; Shepelev et al., 2009). Enfin, 4 familles d'alpha satellites dites anciennes (ancient) sont retrouvées chez l'homme : HI et H2, associées en HORs d'ordre 2 et que l'on pourrait retrouver chez les Catarrhini, tout comme H3 (observées chez *M. mulatta*) ; et pour terminer H4, que l'on pourrait retrouver chez les Simiiformes (observée chez le marmoset *Callithrix jacchus*) (Shepelev et al., 2009).

Les différentes familles d'alpha satellites caractérisées chez les Primates ont par ailleurs été classées historiquement en deux types : celles dont les séquences portent une pJalpha box, appartenant au type A, et celles dont les séquences portent une CENP-B box, appartenant au type B. Ces deux séquences sont détectables à la même position sur les monomères d'alpha

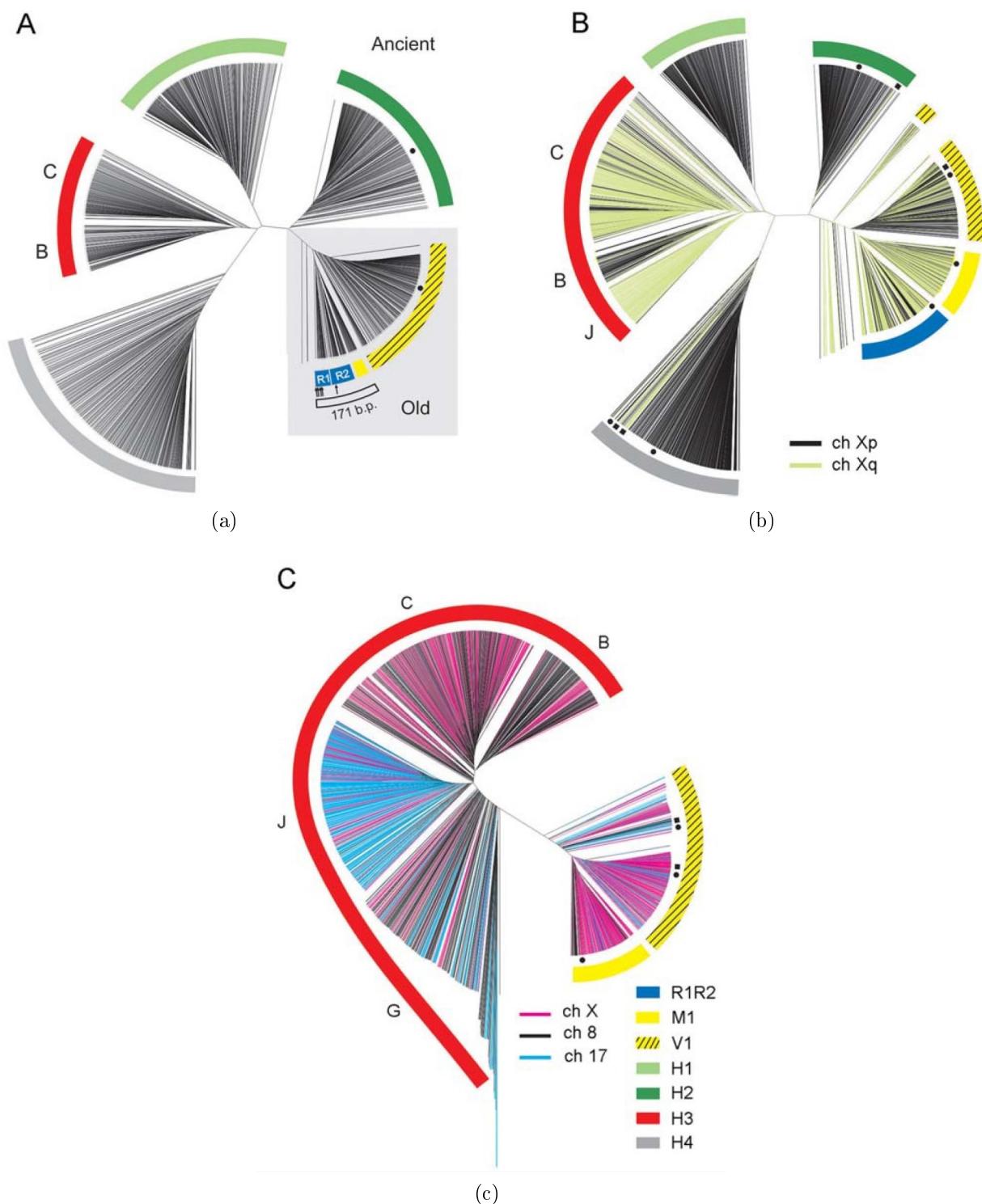


FIGURE 3.4 – Analyses phylogénétiques des monomères d’alpha satellites des péricentromères des chromosomes X, 8 et 17 humains (d’après (Shepelev et al., 2009)). Méthode de distance. Les couleurs des arcs correspondent aux différentes familles d’alpha satellites représentées. (A) 1 434 monomères du péricentromère de l’Xp. (B) 2 516 monomères des péricentromères de l’Xp et de l’Xq. (C) 2 588 monomères des familles H3, V1 et M1 des péricentromères des chromosomes 8, 17 et X. Les lettres autour des arcs rouges indiquent des sous-familles de séquences H3.

satellites (Romanova et al., 1996). Ainsi, les familles S1 à S5 appartiennent au type A, tout comme les familles J1, D2, W4, W5, R2, M1, V1 et H1 à H4, tandis que les familles J2, D1, W1, W2 et R1 appartiennent au type B (Romanova et al., 1996; Shepelev et al., 2009).

1.3.2 Diversité et distribution chromosomique

Dès les années 1980, la distribution chromosomique des différentes HORs humaines a été déterminée via le clonage et séquençage d'alpha satellites depuis des chromosomes spécifiques, isolés dans des noyaux de cellules de rongeurs (Willard, 1985; Waye and Willard, 1986). Dans le même temps, des sondes étaient élaborées depuis des vecteurs de clonage contenant des fragments de HORs, en vue d'être utilisées par FISH pour étudier leur distribution chromosomique (Jabs et al., 1984; Wevrick and Willard, 1989). Ces méthodes ont par extension permis d'étudier la distribution des différentes familles d'alpha satellites contenues dans ces HORs. L'assemblage plus récent de quelques péricentromères humains a également permis d'étudier la distribution des familles d'alpha satellites dans ces régions particulières (Shepelev et al., 2009). La distribution des alpha satellites chez les primates non-humains n'a en revanche été étudiée que dans une bien moindre mesure.

Distribution interchromosomique

Chez l'homme, les différentes HORs sont pour la plupart spécifiques d'un chromosome, tandis que les familles les composant peuvent se retrouver sur des chromosomes différents (Alexandrov et al., 1988, 2001; Hayden, 2012)(Figure 3.5). Ainsi, les familles W1 à W5 s'associent par exemple en HORs d'ordre 11 sur le chromosome 1, d'ordre 5 sur le chromosome 11, d'ordre 16 sur le chromosome 17 ou encore d'ordre 12 sur le chromosome X (Willard and Waye, 1987; Alexandrov et al., 2001). Plusieurs HORs distinctes peuvent par ailleurs être retrouvées sur un même chromosome, comme par exemple sur le chromosome 18 qui porte une HOR d'ordre 8 et une HOR d'ordre 10, chacune composée d'une association de monomères D1 et D2 (Alexandrov et al., 1991). Les chromosomes acrocentriques 13, 14 et 21, au contraire des autres chromosomes, partageraient au moins certaines de leurs HORs (Vissel and Choo, 1991; Hayden, 2012). Les familles impliquées dans des organisations monomériques, telles que M1, R1, R2, V1 ou encore H3, se retrouvent elles aussi sur différents chromosomes (Shepelev et al., 2009; Hayden, 2012)(Figure 3.6).

Les familles d'alpha satellites nouvelles et âgées ont été historiquement regroupées en familles suprachromosomiques (SF, Table 3.1), en fonction de leurs associations dans le génome humain ; ces familles suprachromosomiques n'ont cependant pas de signification phylogénétique. Les familles J1 et J2 appartiennent ainsi à SF1, D1 et D2 à SF2, W1 à W5 à SF3, M1 à SF4 et enfin R1 et R2 à SF5 (Alexandrov et al., 1986, 1988; Willard and Waye, 1987; Romanova et al., 1996).

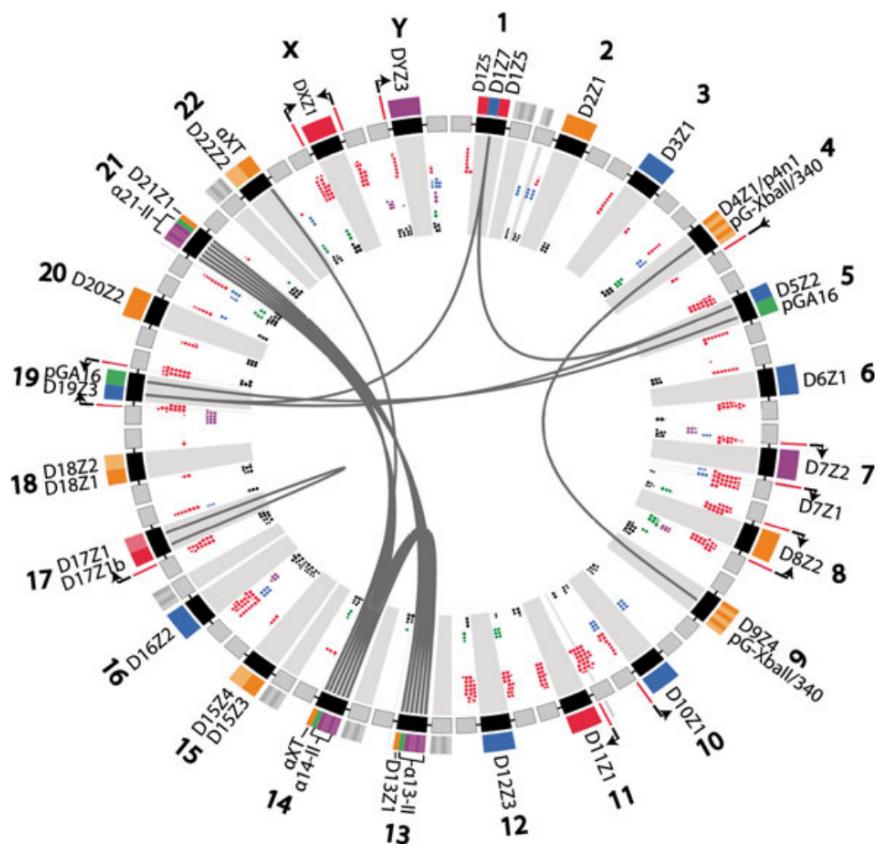


FIGURE 3.5 – Distribution interchromosomique de familles d’alpha satellites nouvelles et âgées chez l’homme (d’après Hayden (2012)). J1/J2 = bleu; D1/D2 = orange; W1-W5 = rouge; M1 = violet; R1/R2 = vert. D = organisation en HORs.

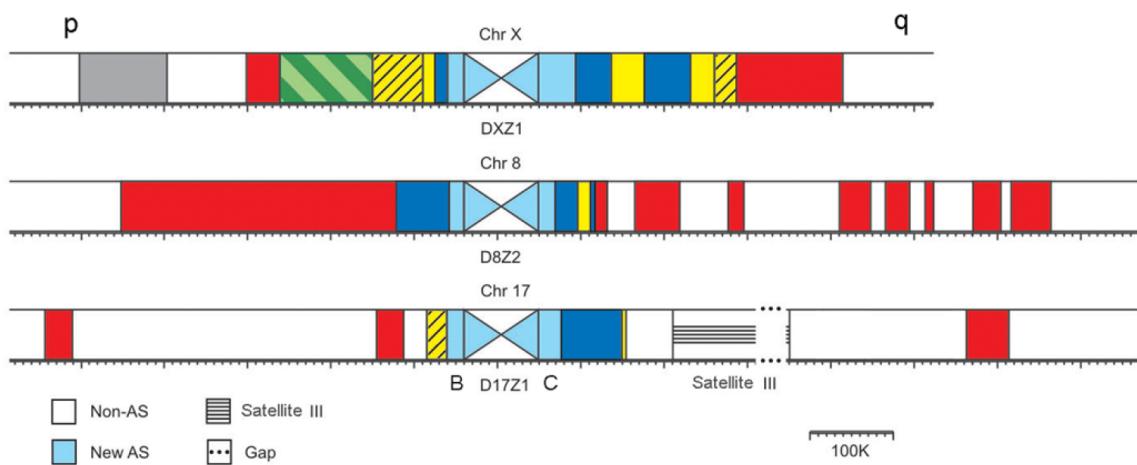


FIGURE 3.6 – Distribution chromosomique des familles d’alpha satellites présentes sur les péri-centromères des chromosomes X, 8 et 17 humains (d’après Shepelev et al. (2009)). R1/R2 = bleu foncé; M1 = jaune; V1 = jaune rayé; H1 = olive (vert pâle); H2 = vert; H3 = rouge; H4 = gris.

La majorité des chromosomes humains ne portent qu'une de ces familles suprachromosomiques, tandis que certains peuvent porter jusqu'à trois familles suprachromosomiques différentes (Figure 3.5). Leur recherche chez d'autres grands singes a permis de montrer que la distribution de SF3 était plutôt conservée, tandis que SF1 et SF2 se distribuaient sur des chromosomes non homologues (Baldini et al., 1992; D'Aiuto et al., 1993; Rocchi et al., 1994; Archidiacono et al., 1995; Haaf and Willard, 1997, 1998). Des HORs humaines sont par ailleurs conservées sur des chromosomes uniques chez d'autres grands singes, mais là encore sur des chromosomes non homologues (Durfy and Willard, 1990; Baldini et al., 1991; Jørgensen et al., 1987, 1992; Warburton et al., 1996; Haaf and Willard, 1997). Par exemple, l'HOR du chromosome 17 humain serait retrouvée sur le chromosome 12 du chimpanzé, homologue du chromosome 2p humain (Warburton et al., 1996).

Enfin, chez les singes du Nouveau et de l'Ancien Monde, l'ADN alpha satellite ne présenterait pas de variabilité interchromosomique ; les sondes alpha satellites utilisées en FISH sur les chromosomes de ces primates s'hybrident plutôt indifféremment sur tous les chromosomes, à l'exception de leur chromosome Y où, pour le moment, la présence de séquences alpha satellites n'a pas été démontrée (Alkan et al., 2007; Prakhongcheep et al., 2013).

Distribution intrachromosomique

L'étude des régions centromériques des grands singes a permis de montrer que, dans les génomes où organisations en HORs et organisations monomériques se côtoient, les premières sont généralement assignées aux centromères et les secondes aux péricentromères (Figure 3.2)(Schueler et al., 2001; Rudd et al., 2006; Hayden, 2012) ; cependant, lorsque deux HORs différentes sont présentes sur un même chromosome, seule l'une des deux serait le site de formation de la centrochromatine (Maloney et al., 2012). Cette règle n'est toutefois pas absolue puisque de courtes HORs humaines (1-10 kb) peuvent être retrouvées au milieu de séquences alpha satellites monomériques (Rudd and Willard, 2004) ; par ailleurs, les familles anciennes H1 et H2 sont associées en HORs dans les péricentromères humains (Shepelev et al., 2009). Par extension, on peut considérer que, globalement, les familles nouvelles se distribuent aux centromères et les familles âgées et anciennes aux péricentromères (Figure 3.6)(Schueler et al., 2001, 2005; Shepelev et al., 2009).

Par ailleurs, les chromosomes de certains primates portent de larges blocs d'hétérochromatine dans leurs régions subtélomériques, comme les chimpanzés et les gorilles où ceux-ci sont constitués de séquences répétées de 32 pb appelées StSat (Subterminal satellites)(Royle et al., 1994; Koga et al., 2011). En revanche, chez les gibbons, ces larges blocs subtélomériques seraient composés de séquences alpha satellites (Koga et al., 2012; Baicharoen et al., 2012) ; celles-ci pourraient donc remplir une fonction dans d'autres compartiments génomiques que les régions centromériques. Plusieurs dizaines de blocs d'alpha satellites ont été de plus identifiés en dehors de ces régions dans le génome humain, possiblement déplacés en ces endroits ectopiques

via des phénomènes de recombinaison entre rétrotransposons (Deininger et al., 2003; Rudd and Willard, 2004).

1.3.3 Mécanismes évolutifs

Les régions centromériques semblent évoluer différemment de l'ensemble du génome. Leurs alpha satellites adopteraient en effet une évolution dite concertée, induisant l'homogénéisation intraspécifique de changements dans ces séquences et donc leur rapide divergence interspécifique. Les séquences alpha satellites s'amplifieraient par ailleurs de manière proximale aux centromères, présentant ainsi un gradient d'âge des centromères vers les péricentromères.

Evolution concertée

L'évolution concertée de séquences d'ADN répétées se définit comme l'évolution non-indépendante de ces séquences, résultant en une similarité réciproque plus élevée à l'intérieur d'une espèce qu'entre espèces différentes (Dover, 1982) ; on parle également d'homogénéisation intraspécifique des séquences répétées. L'ADN alpha satellite adopterait ce type d'évolution par le biais de différents mécanismes d'amplification de séquences au sein voire entre les régions centromériques (Waye and Willard, 1989) ; ces processus sont regroupés sous le terme de dérive moléculaire (Dover, 1982). La transposition de séquences satellites à divers loci pourrait ainsi jouer un rôle dans leur évolution concertée ; par ailleurs, de nombreux modèles mathématiques et simulations informatiques ont montré que l'amplification intra- et interchromosomique des séquences répétées pouvait avoir lieu par crossovers inégaux entre chromatides sœurs, chromosomes homologues et non homologues (Krüger and Vogel, 1975; Smith, 1976; Perelson and Bell, 1977; Ohta and Kimura, 1981; Takahata, 1981; Ohta, 1983; Ohta and Dover, 1984; Alkan et al., 2004)(Figure 3.7). Cependant, ce type de recombinaison homologue pourrait aussi induire des délétions et contractions de séquences répétées, tout comme la recombinaison entre séquences d'un même chromosome (Stephan, 1986; Walsh, 1987; Mirkin, 2007). La conversion génique, un autre type de recombinaison homologue employé par la cellule pour réparer les cassures double brin (Figure 3.8), apparaît donc comme un processus parallèle et essentiel à l'amplification et au maintien des séquences répétées (Ohta and Dover, 1983; Elder Jr and Turner, 1995; Schindelhauer and Schwarz, 2002). Par ailleurs, la découverte chez l'homme d'ADN circulaires portant des séquences satellites a permis d'envisager un autre mécanisme d'amplification pour ces séquences ; elles pourraient en effet être excisées de leur locus chromosomal par recombinaison intrachromosomique, puis amplifiées par réPLICATION EN CERCLE ROULANT ET ENFIN RÉINTÉGRÉES DANS LE GÉNOME (Assum et al., 1993; Feliciello et al., 2006)(Figure 3.9).

L'évolution concertée des séquences alpha satellites peut donc induire leur rapide divergence entre espèces, voire même entre différentes populations d'une même espèce (Willard and Waye,

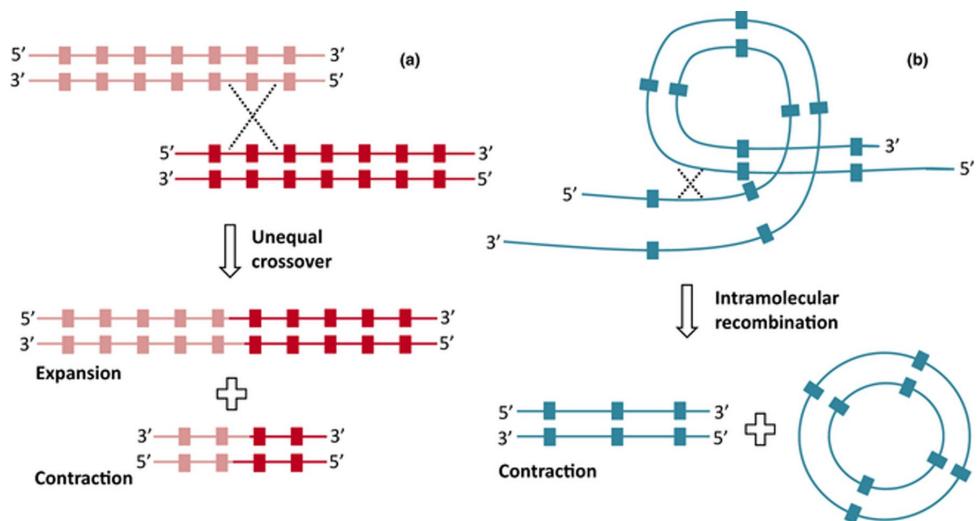


FIGURE 3.7 – Expansion et contraction de séquences répétées en tandem par crossover inégal ou recombinaison intrachromosomique (d'après Zhou et al. (2014)).

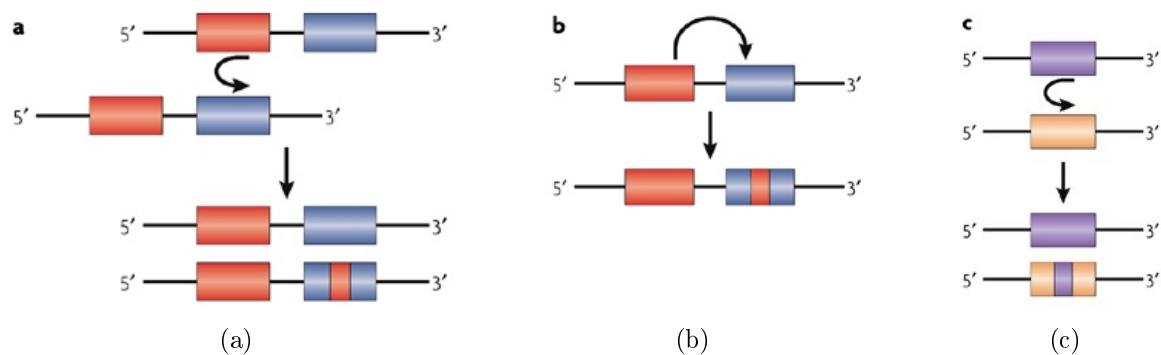


FIGURE 3.8 – Différents types de conversion génique (d'après Chen et al. (2007)). (A) Conversion génique non-allélique en *trans*. (B) Conversion génique non-allélique en *cis*. (C) Conversion génique inter-allélique.

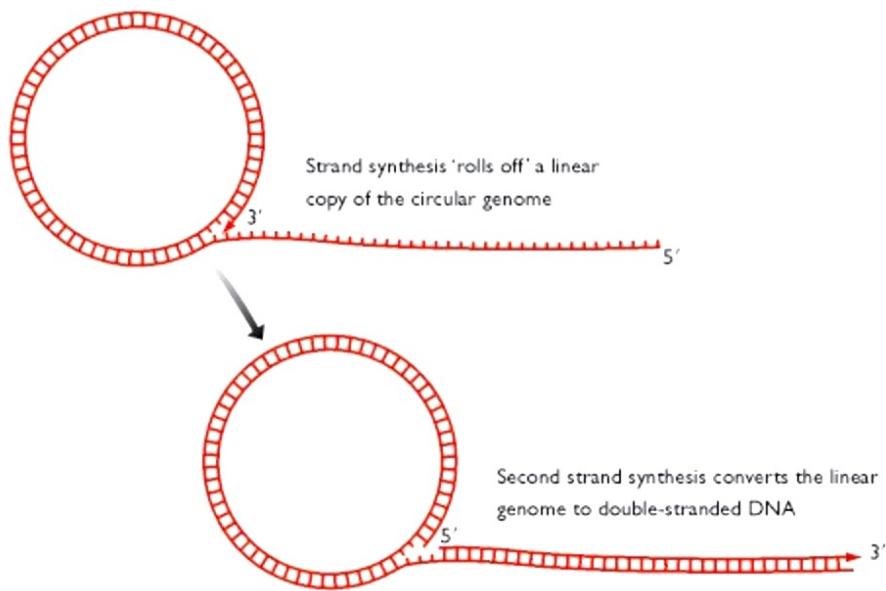


FIGURE 3.9 – Réplication en cercle roulant (d’après Brown (2002)).

1987; Warburton and Willard, 1995). Puisque les unités de répétition des péricentromères sont parsemées d’éléments transposables et apparaissent plus divergentes entre elles que les unités des centromères, l’homogénéisation des séquences alpha satellites s’appliquerait essentiellement sur ces dernières (Schueler et al., 2001; Rudd and Willard, 2004). La comparaison interspécifique des alpha satellites des péricentromères entre deux chromosomes homologues d’homme et de chimpanzé, respectivement le 17 et le 19, a d’ailleurs montré que leur identité de séquence était de 98 % en moyenne, et donc équivalente à l’identité de séquence entre les deux génomes dans leur globalité (Rudd and Willard, 2004; Watanabe et al., 2004; Sequencing et al., 2005). En revanche, la comparaison des HORs centromériques de ces deux chromosomes a montré une identité de séquence interspécifique de seulement 92 à 95 % en moyenne, traduisant une homogénéisation rapide de chacune de ces organisations dans les lignées respectives de l’homme et du chimpanzé. L’évolution concertée des alpha satellites expliquerait donc pourquoi des chromosomes homologues ne porteraient pas, même entre espèces proches, les mêmes familles suprachromosomiques (Roizès, 2006).

De même, l’évolution concertée des séquences alpha satellites peut les faire rapidement diverger entre les chromosomes d’un même génome si les processus d’homogénéisation sont plus efficaces en intra- qu’en interchromosomique ; cela semble être le cas chez les grands singes, dont les HORs sont pour la plupart spécifiques d’un chromosome (Willard, 1985; Willard and Waye, 1987; Willard, 1991). Des homogénéisations intrachromosomiques distinctes peuvent par ailleurs expliquer pourquoi plusieurs HORs homogènes sont retrouvées sur un même chromosome (Willard, 1991). Chez les singes de l’Ancien Monde, l’homogénéisation interchromosomique serait en revanche tout aussi efficace que l’homogénéisation intrachromosomique (Warburton and Willard, 1996; Alkan et al., 2007) ; une transition vers l’homogénéisation intrachromosomique se serait produite chez l’ancêtre des Hominoidea, de manière concomitante avec

l'apparition d'HORs complexes au cœur des centromères ([Shepelev et al., 2009](#)).

Expansion proximale

Bien que, en raison du caractère répété de ses séquences, l'assemblage informatique du centromère soit difficile et qu'il ait été par conséquent largement omis du projet d'assemblage du génome humain ([Collins et al., 1998](#); [Eichler et al., 2004](#); [Rudd and Willard, 2004](#)), les régions péricentromériques de quelques chromosomes ont tout de même pu être reconstruites ([Schueler et al., 2001](#); [Abdellah et al., 2004](#); [Rudd and Willard, 2004](#); [She et al., 2004](#); [Ross et al., 2005](#)) ([Figure 3.10](#)). Cet assemblage a notamment permis une analyse fine des séquences du péricentromère situé sur le bras court du chromosome X, où même l'HOR a été atteinte ([Schueler et al., 2001, 2005](#)). Différents domaines physiques ont ainsi pu être identifiés dans cette région, contenant chacun des séquences alpha satellites particulièrement proches phylogénétiquement ([Figure 3.11](#)) ; cette observation va en faveur de la séparation des alpha satellites en différentes zones d'homogénéisation distinctes le long des régions centromériques. L'analyse du domaine le plus proximal, correspondant à l'HOR de l'X, a pu montrer que ses séquences y étaient de plus en plus en plus divergentes à mesure qu'elles s'éloignaient du cœur de l'HOR vers son extrémité distale ([Figure 3.12](#)). Par ailleurs, l'analyse des éléments transposables retrouvés dans chaque domaine a montré que le domaine le plus proximal ne contenait que des LINE actuellement actifs, tandis que les autres domaines portaient des LINE actifs il y a environ 25 voire 35 millions d'années pour le domaine le plus distal. Un gradient d'âge est ainsi observé le long du péricentromère de l'Xp, depuis de récentes régions proximales vers d'anciennes régions distales. L'ADN des régions centromériques évoluerait donc par amplifications successives de séquences alpha satellites en leur centre, repoussant ainsi de plus anciennes séquences vers les extrémités distales de ces régions ; on parle d'expansion proximale progressive des centromères ([Schueler et al., 2005](#)) ([Figure 3.13](#)). Les monomères les plus distaux des régions centromériques humaines seraient donc les descendants directs de la séquence alpha satellite ancestrale des Primates. De manière intéressante, une séquence LINE a été retrouvée insérée dans ces monomères distaux ; elle est également présente à cette même position chez six autres espèces de Catarrhini, suggérant le caractère ancestral de son insertion et de sa région d'insertion ([Schueler et al., 2005](#)).

Ce phénomène d'expansion proximale fait par ailleurs écho à la distribution, mentionnée plus avant, des familles d'alpha satellites nouvelles aux centromères et des familles âgées et anciennes aux péricentromères. L'analyse de la distribution relative des familles âgées et anciennes sur les péricentromères des chromosomes X, 8 et 17 humain a permis de conforter l'idée qu'au cours de l'évolution de l'homme, les familles d'alpha satellites anciennes auraient été repoussées des centromères par les familles âgées, elles même repoussées par les familles nouvelles ([Shepelev et al., 2009](#)). En effet, les familles anciennes, retrouvées chez des espèces éloignées de l'homme, se distribuent de manière distale et les familles âgées, retrouvées chez des espèces plus proches, se distribuent de manière proximale ; aussi bien sur les péricentromères des bras courts que

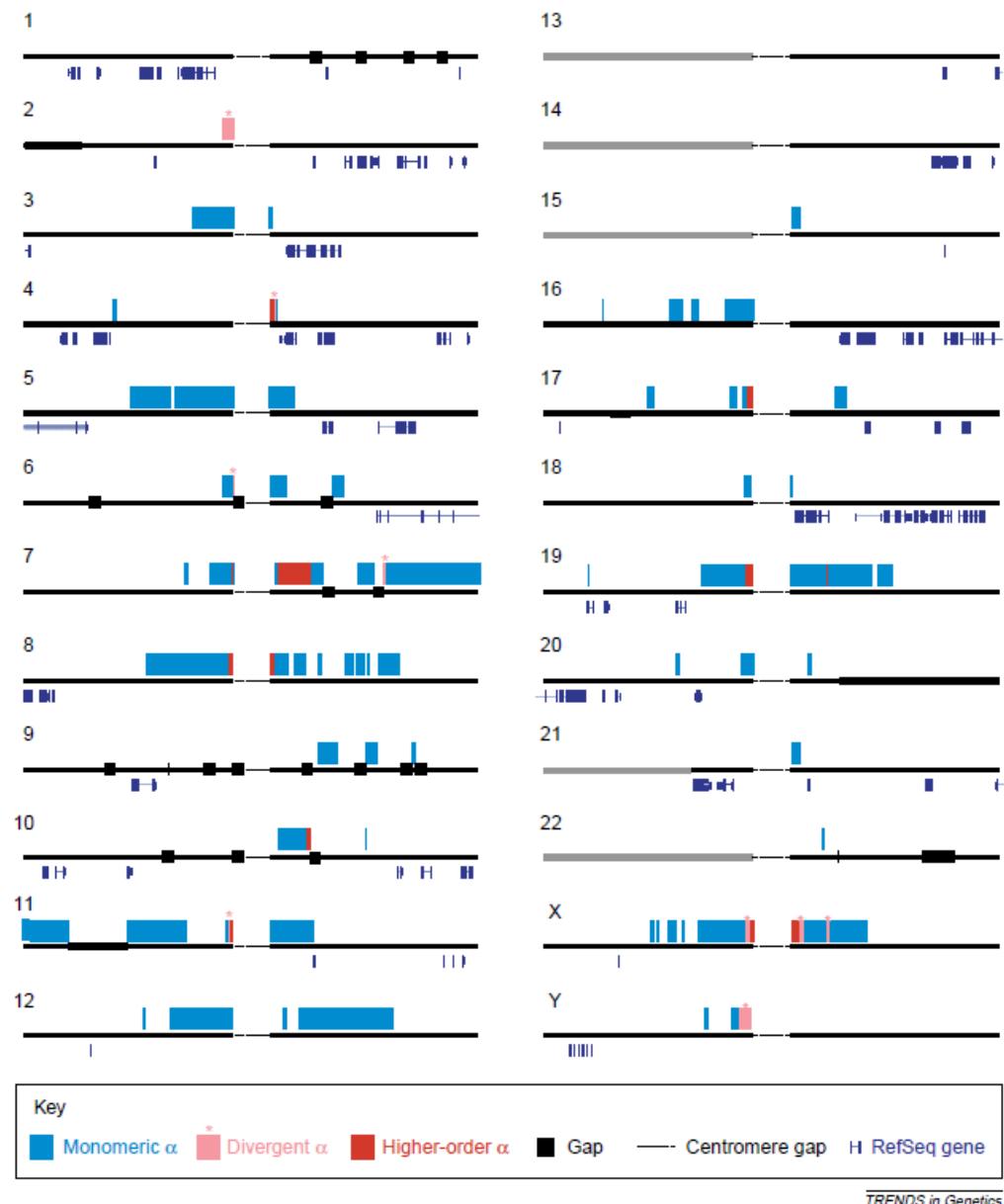


FIGURE 3.10 – ADN alpha satellite assemblé sur la carte du génome humain (d'après Rudd and Willard (2004)). Les péricentromères des bras courts sont représentés à gauche des centromères, et ceux des bras longs à droite.

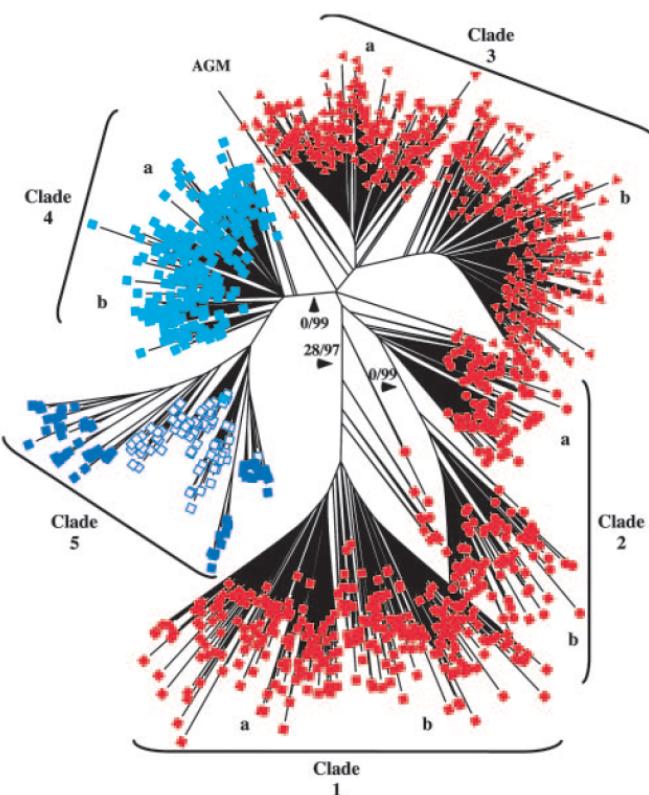


FIGURE 3.11 – Analyse phylogénétique des monomères d’alpha satellites du péricentromère du chromosome Xp humain (d’après Schueler et al. (2005)). Méthode de distance. Séquences de l’HOR du X = carrés bleu foncé, pleins. Séquences adjacentes à l’HOR = carrés bleu foncé, vides et carrés bleu clair. Séquences plus distales = rouges. AGM = African green monkey (*Chlorocebus* ; groupe externe).

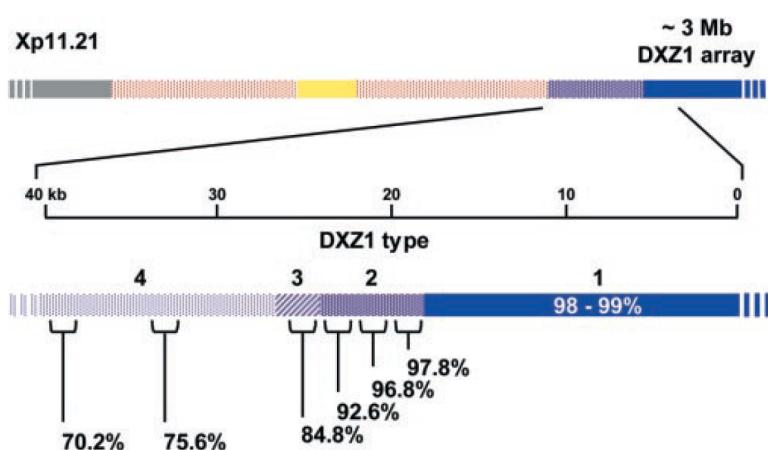


FIGURE 3.12 – Analyse de l’identité de séquence le long de l’HOR du chromosome X humain (d’après Schueler et al. (2001)). Les identités de séquence sont calculées par comparaison avec la séquence consensus de l’HOR DXZ1.

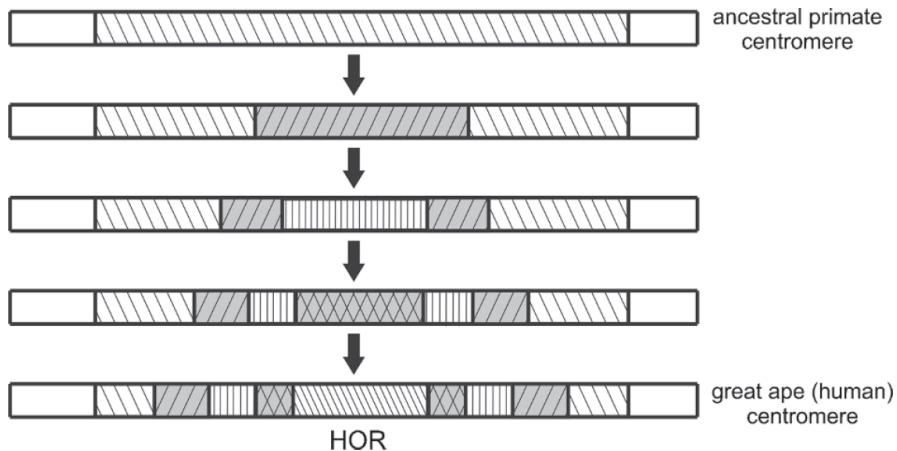


FIGURE 3.13 – Modèle d'évolution par expansion proximale des régions centromériques (d'après Ugarković (2009)). L'évolution de l'ancêtre des Primates jusqu'à l'homme est représentée ici. Des séries d'amplification permettent aux familles d'alpha satellites émergeant aux centromères de remplacer les plus anciennes, qui restent cependant préservées dans les péricentromères.

TABLE 3.2 – Pourcentage d'identité de séquence moyenne des familles d'alpha satellites présentes sur les péricentromères des chromosomes X, 8 et 17 humains (d'après Shepelev et al. (2009))

Chromosome	H4 (gris)	H3 (rouge)	H1 (olive)	H2 (vert)	V1 (jaune rayé)	M1 (jaune)
Xp	71	73	79	79	82	85
Xq		74			81	83
8p		73				
8q		73				
17p		73			80	
17q		72				85

des bras longs des chromosomes, donnant une impression de symétrie autour des centromères (Figure 3.6). L'analyse des LINE retrouvés dans ces familles a par ailleurs permis de confirmer ce gradient d'âge, tout comme le calcul de la divergence à l'intérieur de chacune des familles. On s'attend en effet à ce que plus une famille soit proche du centromère, plus elle soit récemment apparue et moins la divergence au sein de cette famille soit élevée (Table 3.2). Puisque l'apparition progressive des familles d'alpha satellites peut être ainsi inférée d'après l'ordre qu'elles adoptent sur les régions centromériques, et que celle-ci est corrélée avec l'apparition de différents taxons de primates, il a été proposé que l'origine des espèces soit écrite dans les centromères et qu'ils puissent ainsi être utilisés comme outils dans le cadre d'études phylogénétiques (Shepelev et al., 2009).

Conclusion

La diversité de séquences et d'organisations de l'ADN alpha satellite est largement explorée dans le génome humain, voire dans les génomes des autres grands singes ; les informations disponibles chez les singes de l'Ancien et du Nouveau Monde restent en revanche très limitées. La plupart d'entre elles reposent de surcroît sur des études menées il y a plusieurs décennies ; nos connaissances quant à l'ADN alpha satellite de ces clades n'ont pas encore bénéficié de l'utilisation des informatives nouvelles technologies de séquençage. L'analyse des alpha satellites humains a apporté des informations précieuses quant à la diversification de cet ADN ; cependant, son étude approfondie chez de nouveaux modèles primates serait la clé d'une compréhension plus exhaustive de sa dynamique évolutive. Les cercopithèques, bien qu'ayant permis la découverte de l'ADN alpha satellite, sont restés quelque peu oubliés depuis et apparaissent comme des candidats tout à fait propices.

1.4 Diversification et évolution chromosomique chez les Cercopithèques

Les cercopithèques, ou Cercopithecini, sont une tribu de singes de l'Ancien Monde appartenant à la sous-famille des Cercopithecinae (Figure 3.1). Ce clade africain a émergé il y a environ 10 millions d'années et s'est diversifié en 35 espèces actuelles présentant des relations de parenté ambiguës, ainsi que des phénotypes, modes et milieux de vie variés. Leurs génomes ont été largement étudiés via des méthodes de cytogénétique classique et moléculaire, mettant en évidence la survenue de remaniements chromosomiques tout à fait particuliers au cours de leur évolution ; les chromosomes des cercopithèques se seraient en effet régulièrement fissionnés, et ce de manière non centromérique, depuis la divergence de ce clade. De nombreux nouveaux centromères évolutifs se seraient ainsi formés au sein de leurs génomes.

1.4.1 Diversité écologique et comportementale

Les cercopithèques forment la plus grande tribu des Catarrhini actuels et se divisent en 5 genres : *Cercopithecus* (25 espèces), *Chlorocebus* (6 espèces), *Erythrocebus* (1 espèce), *Allenopithecus* (1 espèce) et *Miopithecus* (2 espèces) (Groves, 2001; Wilson and Reeder, 2005). Ils se distribuent largement en Afrique et adoptent des niches écologiques différentes, aux influences multiples sur l'évolution de leurs stratégies de survie. Les cercopithèques sont ainsi particulièrement diversifiés dans leurs couleurs, morphologies, régimes alimentaires ou encore organisations sociales.

Distribution géographique

Toutes les espèces de cercopithèques sont endémiques à l'Afrique subsaharienne ([Wolfheim, 1983](#); [Enstam and Isbell, 2007](#)) (Figure 4.1). Le genre *Chlorocebus* possède la plus grande aire de répartition, puisqu'on retrouve les espèces qui le composent dans toute l'Afrique subsaharienne, à l'exception du désert de Namibie en Afrique du Sud et des denses forêts tropicales d'Afrique de l'Ouest (Niger, Cameroun, Congo et République Démocratique du Congo). Les espèces du genre *Cercopithecus* se distribuent principalement en Afrique centrale et Afrique de l'Ouest ; la distribution de quelques espèces, comme *Cercopithecus lhoesti*, *C. mitis* et *C. neglectus*, s'étend à l'est jusqu'au Rwanda, en Uganda et au Kenya. D'autres espèces de *Cercopithecus* ont en revanche une distribution très restreinte, comme *C. solatus* (Figure 4.2) qui est endémique du Gabon ([Harrison, 1988](#)). *Allenopithecus nigroviridis* est retrouvé en Afrique centrale, depuis l'Angola au sud-ouest jusqu'à la République Démocratique du Congo au nord-est. La distribution d'*Erythrocebus patas* prend quant à elle la forme d'une bande étroite au sud du désert du Sahara, allant du Sénégal et de la Guinée à l'ouest jusqu'à l'Ethiopie, le Kenya et la Tanzanie à l'est. Enfin, les deux espèces de *Miopithecus* sont retrouvées en Afrique de l'Ouest, avec *Miopithecus talapoin* distribué plus au nord, au Cameroun et au Gabon, et *M. ogouensis* distribué plus au sud, en Angola et en République Démocratique du Congo.

Habitats et régimes alimentaires

Les espèces du genre *Cercopithecus* occupent des habitats forestiers variés mais généralement denses, comme les forêts tropicales humides, les forêts de bambous ou encore les forêts marécageuses ([Zieve, 1991](#); [Fleury and Gautier-Hion, 1997](#); [Struhsaker, 2000](#)). Les *Chlorocebus* préfèrent des habitats plus ouverts comme les savanes arborées, tout comme *E. patas* que l'on retrouve également dans des prairies ([Nakagawa, 1999](#); [Enstam and Isbell, 2002](#)). *Allenopithecus nigroviridis* vit quant à lui dans des forêts marécageuses, tout comme les *Miopithecus* que l'on peut cependant retrouver aussi dans des forêts plus tropicales ([Gautier-Hion, 1971, 1973](#); [Gautier, 1985](#)). Ces différents milieux de vie influent sur l'alimentation des cercopithèques ; les espèces vivant en milieux forestiers et celles vivant en milieux ouverts adoptent des régimes alimentaires distincts.

Le régime des cercopithèques forestiers se compose essentiellement de fruits mais inclut également de jeunes feuilles, des fleurs, des graines, des insectes ou encore de petits vertébrés ; en proportions variant entre espèces, mais également entre populations d'une même espèce. La proportion de fruits inclus dans le régime alimentaire de *C. mitis* peut ainsi varier de 15 à 91 % ([Butynski, 1990](#); [Lawes et al., 1990](#)), et la proportion de feuilles incluses dans le régime de *C. cephus* (Figure 4.3) aller jusqu'à 74 % pour les individus de la forêt de Budongo (Uganda) ([Sheppard, 2000](#)). Les graines, bien que difficiles à digérer, peuvent composer la majeure partie du régime de certaines populations vivant dans des niches pauvres en ressources, comme

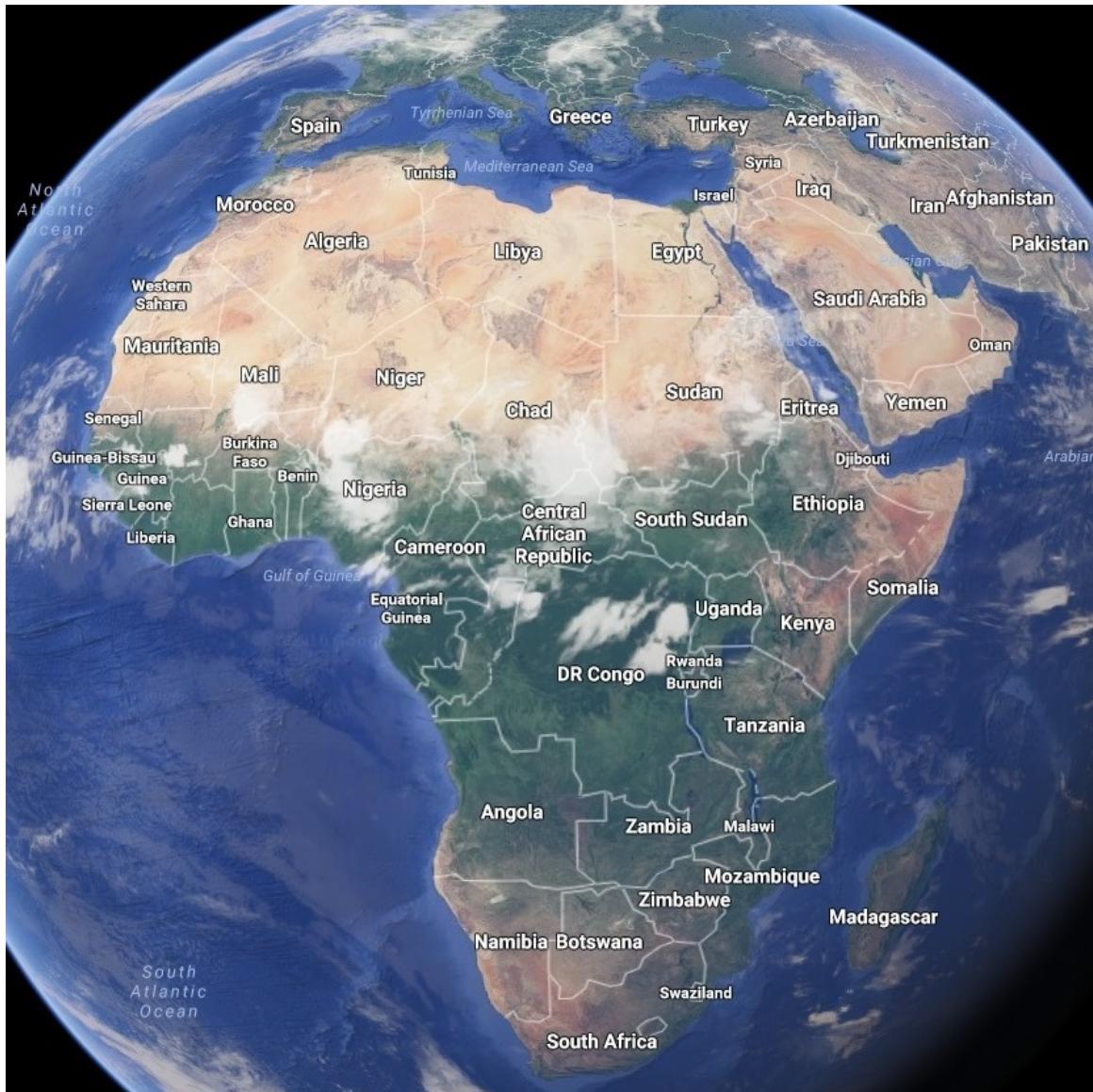


FIGURE 4.1 – Carte géographique du continent africain (Imagery ©2016 Data SIO, NOAA, U.S. Navy, NGA, GEBCO, Landsat, IBCAO, Map data ©2016 Google, ORION-ME). Le Congo se situe à l'ouest de la République Démocratique du Congo, et le Gabon au sud de la Guinée Equatoriale.



FIGURE 4.2 – *Cercopithecus solatus*, Gabon (photographie : Peggy Motsch)

les *C. nictitans* (Figure 4.4) et *C. pogonias* de la forêt de Makandé (Gabon) (Brugiere et al., 2002). Les insectes peuvent également représenter une grande partie du régime alimentaire des cercopithèques forestiers, atteignant 45 % de celui-ci pour les *C. mitis* de la forêt de Kibale (Uganda) (Butynski, 1990). Les cercopithèques les chassent activement et utilisent pour cela plusieurs techniques ; ils peuvent par exemple préférer attraper leur proie à la main, ou bien tenir le substrat (feuille, branche) où elle se trouve et l'attraper directement avec leur bouche, comme les *C. mitis* de la forêt de Kakamega (Kenya) (Cords, 1986). Les cercopithèques forestiers peuvent également consommer occasionnellement des lézards, serpents, oiseaux, écureuils, souris et chauve-souris ; voire même d'autres primates de petite taille, comme les galagos (Furuichi, 2006; Cords and Fuller, 2010; Lawes et al., 2013; Tapane et al., 2016).

Les cercopithèques de milieux ouverts ne consomment quant à eux que très peu de fruits ; ceux-ci ne représentent pas plus de 10 % de leur régime (Whitten, 1983; Nakagawa, 1989; Isbell, 1998). La sève solidifiée, ou gomme, des acacias est en revanche un aliment important pour eux ; elle constitue par exemple 37 % du régime des *E. patas* du Kenya (Isbell, 1998). Les acacias leur fournissent également des feuilles, des épines, des graines, des fruits, des fleurs et de l'écorce ; ces arbres représentent à eux seuls 75 % du régime des *Chlorocebus* du parc national d'Amboseli (Kenya) (Wrangham and Waterman, 1981) et 83 % du régime des *E. patas* du Kenya (Isbell, 1998; Pruetz and Isbell, 2000). Les cercopithèques de milieux ouverts se nourrissent également d'insectes, mais qu'ils ne chassent généralement pas de manière active ; les *E. patas* du Kenya consomment principalement des fourmis qu'ils obtiennent en mordant dans les grosses épines d'acacias où elles nichent, et en les extrayant avec leur langue (Isbell, 1998; Enstam and Isbell, 2007). Pour ne pas que les fourmis s'alarment et les attaquent, ils ne se nourrissent que sur une ou deux épines d'un même acacia, avant de migrer vers un nouvel arbre. Les *E. patas* consomment également et occasionnellement des criquets, auxquels ils donnent chasse dans les



FIGURE 4.3 – *Cercopithecus cephus*, Gabon (photographie : Peggy Motsch)

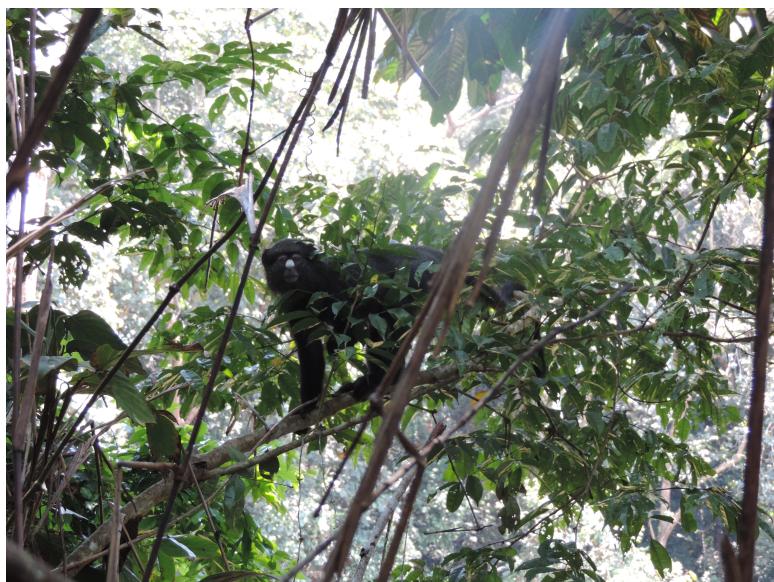


FIGURE 4.4 – *Cercopithecus nictitans*, Gabon (photographie : Peggy Motsch)

herbes (Enstam and Isbell, 2007).

Systèmes sociaux et associations polyspécifiques

La plupart des cercopithèques vivent en groupes se composant d'un mâle résident et de plusieurs femelles (Rowell, 1988; Cords, 2004), à l'exception des *Chlorocebus*, de *M. talapoin* et d'*A. nigroviridis* qui vivent en groupes multimâles et multifemelles (Struhsaker, 1967; Gautier-Hion, 1971, 1973; Gautier, 1985; Zieve, 1991). Les groupes multisexes de *M. talapoin* ne se forment cependant qu'en période de reproduction ; le reste du temps, les mâles et les femelles vivent séparés (Rowell, 1971; Rowell and Dixson, 1975). Dans des contextes de populations en déclin, il est également possible de retrouver des groupes se réduisant à un mâle et une femelle, comme observé pour *C. neglectus* (Gautier-Hion and Gautier, 1978). Chez toutes les espèces étudiées jusqu'ici, les mâles quittent leur groupe une fois atteint la maturité sexuelle, tandis que les femelles restent leur vie durant au sein du même groupe ; excepté dans des cas de déclin de populations où les femelles d'un groupe peuvent décider de se joindre à un groupe voisin (Hauser et al., 1986; Isbell et al., 1991). Les mâles partant peuvent intégrer un nouveau groupe en se battant éventuellement avec un mâle résidant pour prendre sa place (Enstam et al., 2002). Ce dernier n'est cependant pas toujours forcé de quitter le groupe lorsque déchu (Tsingalia and Rowell, 1984) ; plusieurs mâles peuvent, en particulier, cohabiter au sein d'un groupe normalement non-multimâle pendant les périodes de reproduction (Chism and Rowell, 1986; Harding and Olson, 1986; Cords, 1988, 2004). Un seul mâle ne serait en effet pas toujours en mesure de défendre un accès exclusif à plusieurs femelles lors de ces périodes (Cords, 2004). Suite à de telles instabilités, il peut arriver, bien que rarement, que le groupe se divise et se répartisse entre différents mâles (Struhsaker and Leland, 1988).

Etant donnée la structure sociale prédominante chez les cercopithèques, la majorité des interactions sociales au sein d'un groupe se fait entre femelles, et entre les femelles et le mâle résident ; ce principalement pendant la saison de reproduction, puisque le mâle adopte la plupart du temps une position périphérique par rapport au groupe (Hall, 1966; Chism et al., 1984; Rowell, 1988). Les interactions entre femelles d'un même groupe sont généralement amicales et consistent surtout en toilettage mutuel, ou grooming. Les femelles prennent également soin de leurs petits mais aussi des petits des autres femelles (Chism, 1978; Struhsaker and Leland, 1979; Zucker and Kaplan, 1981; Chism and Rogers, 2004). Chez la plupart des espèces de cercopithèques, il n'y a pas de hiérarchie clairement établie entre individus (Cords et al., 1987; Cords, 2000; Isbell and Pruetz, 1998). Pour ce qui est des relations entre groupes, celles-ci sont en revanche plutôt agressives (Cheney, 1987), que les individus défendent (*C. diana*, *C. ascanius*, *C. mitis*) ou non (*E. patas*, *C. neglectus*) un territoire particulier (Struhsaker and Leland, 1979; Cords et al., 1987; Hill, 1994). Lors des rencontres, les groupes échangent en premier lieu via des vocalisations dédiées aux situations de conflit, tant de la part des femelles que des mâles qui, stimulés par les cris des femelles, produisent des sons particulièrement graves (Hill, 1994). Les individus peuvent ensuite se chasser, parfois sur de longues distances, et s'agresser

physiquement si le conflit perdure (Chism, 1999; Windfelder and Lwanga, 2004).

Les cercopithèques vivant en forêt et adoptant un mode de vie arboricole, c'est-à-dire tous les *Miopithecus*, *Allenopithecus* et *Cercopithecus* - à l'exception de *C. neglectus*, *C. lhoesti*, *C. preussi* et *C. solatus* qui sont semi-terrestres - ont par ailleurs tendance à s'associer entre espèces au sein d'un même groupe, ou avec d'autres espèces de primates, de manière transitoire ou pérenne ; on parle alors d'associations polyspécifiques (Figure 4.5). Par exemple, *C. pogonias* et *C. nictitans* passent l'intégralité de leur temps associés l'un avec l'autre ou avec d'autres espèces de cercopithèques (Gautier-Hion and Gautier, 1974; Gautier-Hion et al., 1983). Dans de rares cas, ces associations résultent en croisements interspécifiques pouvant donner naissance à des hybrides fertiles (au moins partiellement)(Figure 4.6), comme reporté entre *C. mitis* et *C. ascanius* (Aldrich-Blake, 1968; Struhsaker and Leland, 1988). Les bénéfices des associations polyspécifiques entre cercopithèques reposeraient principalement sur l'accès à la nourriture et la détection des prédateurs. En effet, les groupes polyspécifiques ont tendance à explorer de plus vastes territoires, et accèdent ainsi à plus de ressources (Gautier-Hion, 1988) ; les espèces se guident les unes les autres en fonction de leur degré de connaissance de l'environnement (Struhsaker, 1981; Cords et al., 1987; Cords, 1990). De plus, de telles associations réduisent le risque pour une espèce d'accéder tardivement à un site déjà exploité par une autre espèce (Gautier-Hion et al., 1983). Les prédateurs contre lesquels les associations polyspécifiques entre cercopithèques permettent de lutter plus efficacement sont majoritairement l'aigle couronné, le chimpanzé, le léopard et l'homme (Gautier-Hion et al., 1983; Haltenorth and Diller, 1988; Struhsaker and Leakey, 1990). En effet, différentes espèces de cercopithèques peuvent être plus ou moins habiles à détecter un certain danger ; au sein d'un groupe trispécifique de *C. pogonias*, *C. cephus* et *C. nictitans* suivi au Gabon, ce sont ainsi principalement les *C. cephus* qui alarment en premiers les autres des dangers venant du sol, et les *C. pogonias* qui poussent les premiers cris d'alarme concernant les dangers venant du ciel (Gautier-Hion, 1988). Différentes espèces peuvent également se protéger plus activement : un mâle *C. nictitans* a été vu rentrer en conflit avec un aigle pour défendre un *C. cephus* de son groupe polyspécifique (Gautier-Hion, 1988). Les *C. cephus* sont en effet particulièrement sensibles à ce prédateur en raison de leur petite taille ; leur association avec des espèces comme *C. nictitans* et *C. pogonias* leur permet d'explorer avec plus de sécurité des milieux forestiers moins denses mais plus riches en ressources, où ils n'osent pas s'aventurer lorsqu'ils sont en groupes monospécifiques (Gautier-Hion et al., 1981, 1983).

Les cercopithèques vivant en forêt mais adoptant un mode de vie semi-terrestre, ne s'assortant pas en groupes polyspécifiques, ont développé des stratégies de survie différentes des cercopithèques arboricoles (Gautier, 1975; Quris, 1976; Gautier-Hion and Gautier, 1978). Ils vivent généralement en groupes plus restreints, et parcouruent moins de distance que les cercopithèques arboricoles pour se nourrir ; ils exploitent ainsi beaucoup plus activement les ressources qui s'offrent à eux. Par ailleurs, ils sont plutôt cryptiques et luttent contre les prédateurs en se cachant d'eux ; les *C. neglectus* n'ont même aucun cri d'alarme et préfèrent s'éloigner silencieusement en cas de danger ou bien se figer, pendant plusieurs heures si nécessaires. La stratégie



FIGURE 4.5 – *Cercopithecus nictitans* et *C. cephus* en association polyspécifique, Gabon (photographie : Peggy Motsch)



FIGURE 4.6 – Hybrides *C. nictitans* × *C. cephus*, Gabon (photographie : Peggy Motsch)



FIGURE 4.7 – Visage coloré du *C. cephus* (photographie : Peggy Motsch)

anti-prédation adoptée par les *Cercopithecus* semi-terrestres est en adéquation avec leurs robes plutôt neutres, qui tranchent avec les robes très colorées et contrastées des cercopithèques arboricoles (Rowell, 1988)(Figure 4.7). Les profils de couleurs des visages de ces derniers se seraient en effet diversifiés via la sélection de caractères leur permettant de mieux se distinguer entre espèces, et ainsi d'éviter la production coûteuse d'hybrides non ou partiellement fertiles (Kingdon, 1980, 2007; Allen et al., 2014).

1.4.2 Relations de parenté entre espèces

La collecte d'échantillons représentatifs sur un clade aussi large que les cercopithèques, composé d'espèces rares, peu évidentes à capturer et vivant dans des espaces naturels difficiles d'accès, n'est pas chose aisée ; ce qui se révèle être un frein à l'étude de leur évolution moléculaire. Les prélèvements faits sur des individus en captivité ont toutefois permis l'investigation des génomes de plusieurs espèces, via l'étude de leurs chromosomes et de leur ADN ; les progrès récents des méthodes d'exploitation de spécimens de collections muséologiques laissent par ailleurs espérer une compréhension plus exhaustive de l'histoire évolutive des cercopithèques.

Phylogénies chromosomiques

Les premières tentatives de résolution moléculaire des relations de parenté entre espèces de cercopithèques ont débuté dans les années 1970 et été poursuivies dans les années 1980, sur la base du partage de caractères chromosomiques (Dutrillaux et al., 1978, 1979, 1981, 1988a). Ces

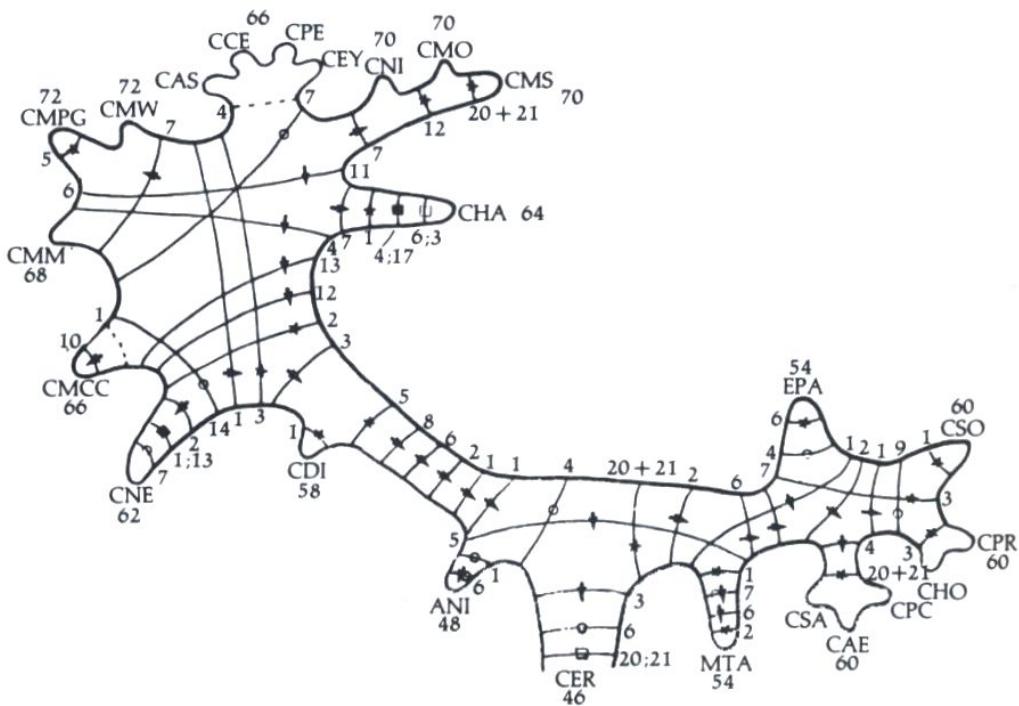


FIGURE 4.8 – Phylogénie chromosomique des cercopithèques, 1988 (d'après Dutrillaux et al. (1988b)). Flèche = fission, Etoile = relocalisation centromérique potentielle, Cercle = inversion, Carré = translocation. Le clade terrestre est visible à droite, avec les *Chlorocebus* (CSA *C. sabaeus*, CAE *C. aethiops*, CPC *C. cynosuros*), *E. patas* (EPA) et les *Cercopithecus* semi-terrestres (CHO *C. lhoesti*, CPR *C. preussi*, CSO *C. solatus*). Le clade arboricole est visible à gauche, avec les *Cercopithecus* arboricoles (CDI *C. diana*, CNE *C. neglectus*, CMCC *C. campbelli*, CMM *C. mona*, CMPG *C. pogonias*, CMW *C. wolfi*, CAS *C. ascanius*, CCE *C. cephus*, CPE *C. petaurista*, CEY *C. erythrotis*, CNI *C. nictitans*, CMO *C. mitis opisthostictus*, CMS *C. mitis stuhlmanni*, CHA *C. hamlyni*). ANI = *A. nigroviridis*, MTA = *M. talapoin*, CER = Ancêtre des cercopithèques. Le nombre de chromosomes dans le caryotype de chaque espèce est mentionné aux côtés de son nom. Les numéros associés aux remaniements correspondent à ceux des chromosomes ancestraux subissant ces remaniements.

analyses caryotypiques ont rapidement mis en évidence l'existence de deux clades distincts au sein des cercopithèques, l'un regroupant majoritairement les espèces arboricoles (la plupart des *Cercopithecus*) et l'autre les espèces semi-terrestres (*C. lhoesti*, *C. preussi* et *C. solatus* ainsi que les *Chlorocebus* et *Erythrocebus*). *Allenopithecus* serait plus proche du premier, appelé clade arboricole, et *Miopithecus* du second, appelé clade terrestre (Figure 4.8). Les relations de parenté au sein de chaque clade apparaissent obscures, en particulier pour le clade arboricole ; les espèces auraient en effet conservé des caractères chromosomiques identiques mais différentiellement combinés dans leurs génomes. Ce profil serait le résultat d'une évolution dite populationnelle ou réticulée des cercopithèques, au cours de laquelle différentes populations ancestrales auraient divergé les unes des autres en maintenant des flux géniques entre elles et en échangeant différentes portions de leurs génomes (Dutrillaux et al., 1988b). Ces résultats ont été corroborés par une phylogénie chromosomique plus récente, qui a également permis, grâce à de nouvelles méthodes d'analyse des chromosomes, de corriger quelques erreurs quant aux remaniements évolutifs inférés précédemment (Moulin et al., 2008)(Figure 4.9).

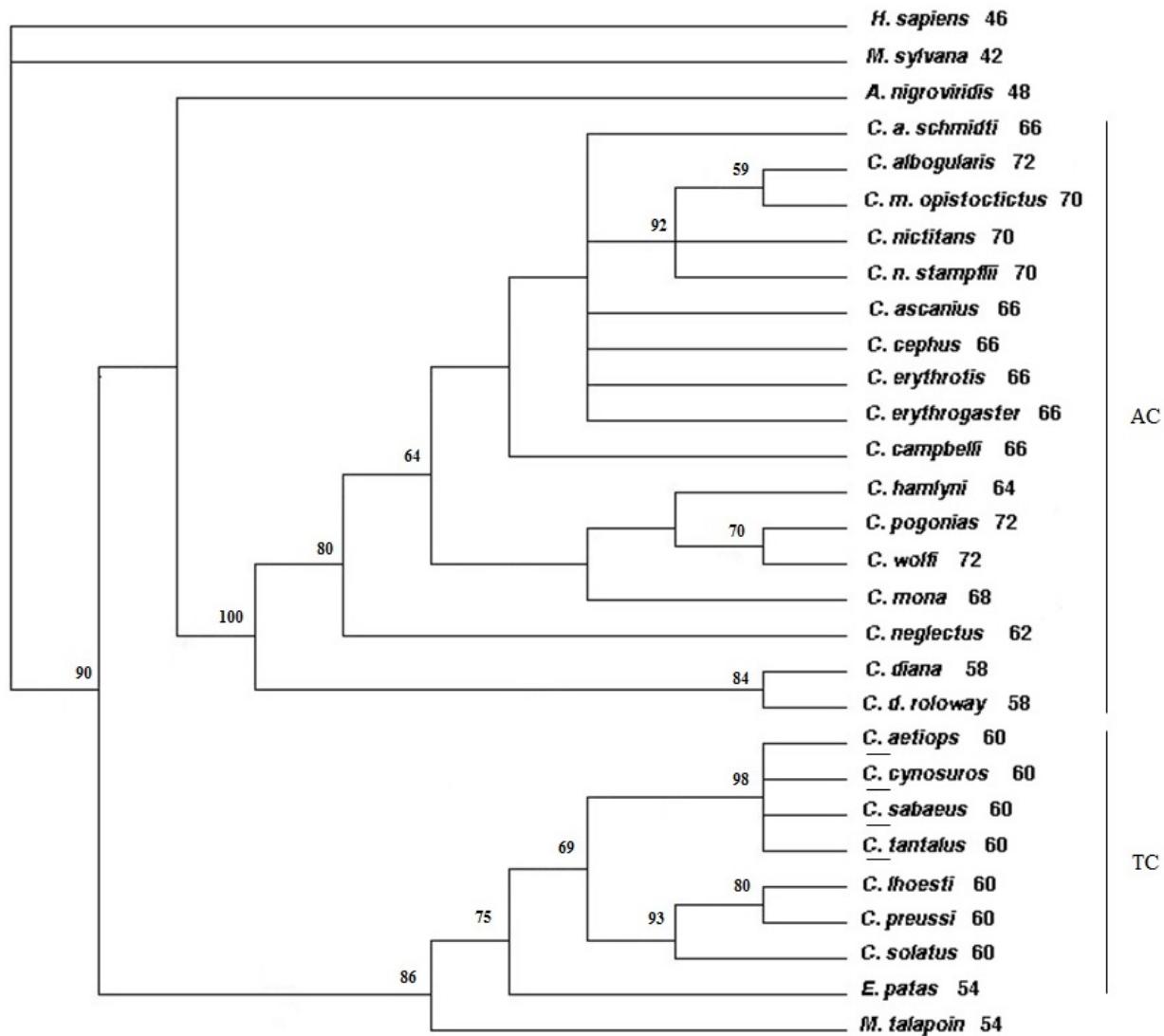


FIGURE 4.9 – Phylogénie chromosomique des cercopithèques, 2008 (d'après Moulin et al. (2008)). Méthode de parcimonie ; les valeurs de bootstrap supérieures à 50 sont apparentes. Le nombre de chromosomes dans le caryotype de chaque espèce est mentionné aux côtés de son nom. *C.* = *Chlorocebus*. AC = Clade arboricole, TC = Clade terrestre.

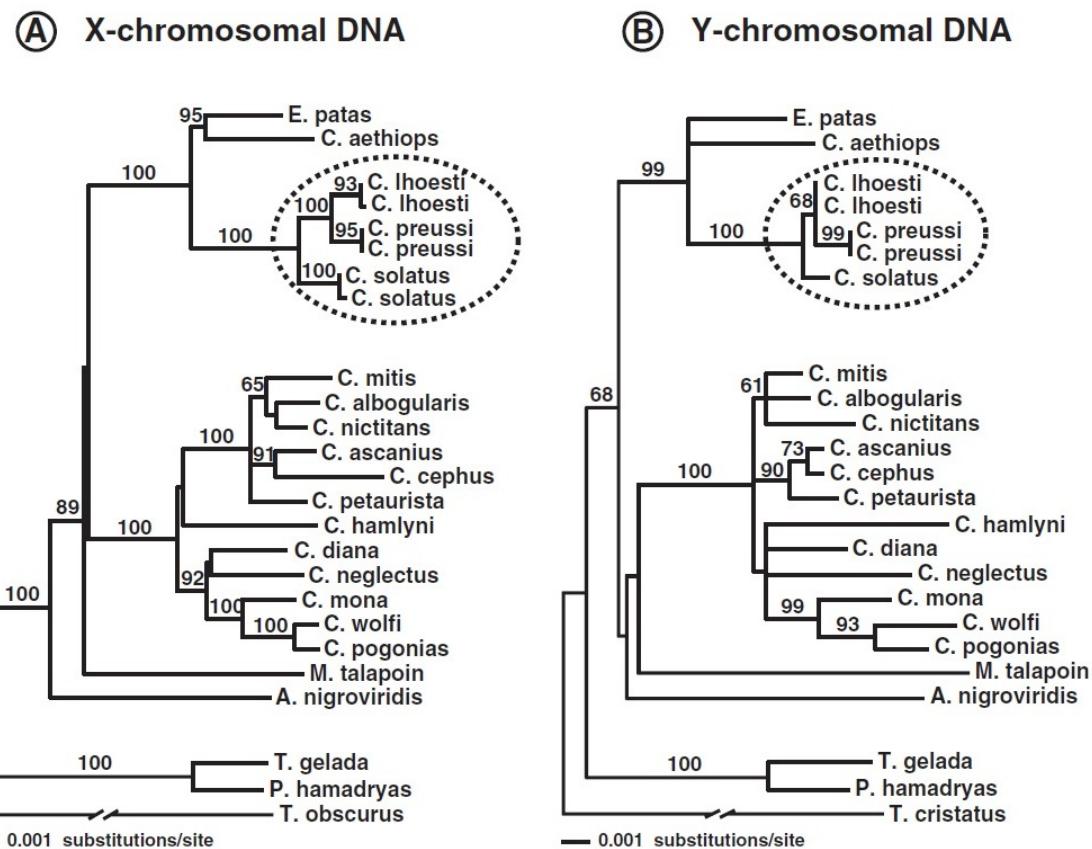


FIGURE 4.10 – Phylogénies nucléaires des cercopithèques (d'après Tosi (2008)). Méthode de vraisemblance ; les valeurs de bootstrap supérieures ou égales à 50 sont apparentes.

Phylogénies moléculaires

Les premières phylogénies moléculaires des cercopithèques ont été proposées dans les années 1990, et la comparaison interspécifique de marqueurs nucléaires - gènes du chromosome Y, gènes du chromosome X et loci d'insertion de SINE - confortent la présence d'un clade arboricole et d'un clade terrestre de cercopithèques (Tosi et al., 2002, 2003, 2005; Tosi, 2008; Xing et al., 2007)(Figure 4.10). La position des *Allenopithecus* et *Miopithecus* par rapport à ces clades reste cependant ambiguë et diffère en fonction des marqueurs utilisés. Leurs lignées exclusives se brancheraient néanmoins à la base de l'arbre des cercopithèques, voire à la base du clade arboricole ou du clade terrestre (Tosi, 2008; Perelman et al., 2011). Les relations de parenté au sein des clades arboricole et terrestre sont, elles aussi, généralement peu résolues. Par ailleurs, la place occupée par *C. lhoesti*, *C. preussi* et *C. solatus* au sein du clade terrestre, avec les *Chlorocebus* et *E. patas*, confirme le caractère paraphylétique du genre *Cercopithecus* : d'où la proposition récente du nouveau genre *Allochrocebus*, regroupant les trois espèces concernées ; ce genre n'est cependant pas encore globalement adopté (Kingdon et al., 2013; Motsch et al., 2015).

L'utilisation de marqueurs mitochondriaux pour reconstruire les relations de parentés des cercopithèques est longtemps restée rare (van der Kuyl et al., 1995; Disotell and Raaum, 2004), mais a récemment permis l'élaboration de la plus exhaustive des phylogénies proposées pour ce clade (Guschanski et al., 2013). Grâce à l'extraction d'ADN depuis de nombreux spécimens de collections suivi du séquençage intégral de mitogénomes, l'étude des relations entre la quasi-totalité des espèces de cercopithèques a pu être menée. L'existence d'une incongruence marquée des génomes nucléaires et mitochondriaux a ainsi pu être confortée ; les phylogénies mitochondrielles des cercopithèques ne soutiennent pas l'existence des clades terrestre et arboricole (Figures 4.11 et 4.12). Tandis que *C. solatus* et les *Chlorocebus* restent groupés dans un même clade mais avec *C. hamlyni*, *C. lhoesti*, *C. preussi* et *E. patas* apparaissent appartenir à un clade rassemblant également les espèces du clade arboricole chromosomique et nucléaire. Les lignées d'*A. nigroviridis* et des *Miopithecus* adoptent en revanche toujours une position basale dans l'arbre des cercopithèques.

Il n'est pas rare d'observer une discordance entre arbres nucléaires et mitochondriaux pour un clade donné ; les génomes nucléaires et mitochondriaux n'y auraient en fait pas la même histoire évolutive, et plusieurs causes peuvent être en jeu (Funk and Omland, 2003; Toews and Brelsford, 2012). D'une part, des hybridations répétées peuvent mener au transfert du génome mitochondrial d'une espèce dans une autre espèce ; c'est ce que l'on appelle l'introgression (voire hybridation introgressive ou encore flux génique interspécifique) mitochondriale. Il est ainsi possible que les ancêtres d'*E. patas*, ou bien ceux des espèces soeurs *C. lhoesti* et *C. preussi*, se soient vus accueillir le génome mitochondrial des ancêtres d'espèces du clade arboricole. Ce phénomène est d'autant plus probable que le génome mitochondrial se transmet de manière matrilinéaire, et que les hybrides femelles primates, portant le sexe homogamétique, sont moins affectées dans leur fertilité que les hybrides mâles (Haldane, 1922). Les introgressions mitochondrielles sont en effet restreintes dans les clades où le mâle porte le sexe homogamétique, comme chez les oiseaux (Sattler and Braun, 2000; Brumfield et al., 2001; Allen, 2002). D'autre part, des phénomènes de spéciation par hybridation, c'est à dire de formation d'une nouvelle espèce par l'hybridation de deux autres, peuvent également être à l'origine de l'incongruence des marqueurs nucléaires et mitochondriaux. Le génome nucléaire de la nouvelle espèce hybride serait en effet une mosaïque des génomes nucléaires de ses espèces mères, tandis que son génome mitochondrial lui serait transmis par l'une ou l'autre de celles-ci. Il est donc possible qu'*E. patas*, ou bien *C. lhoesti* et *C. preussi*, soient de nouvelles espèces nées respectivement de l'hybridation entre deux espèces ancestrales, l'une appartenant au clade arboricole et leur ayant légué son génome mitochondriale, et l'autre appartenant au clade terrestre des cercopithèques.

1.4.3 Cytogénétique évolutive des cercopithèques

Les chromosomes des Primates évoluent via divers remaniements, comme les fissions (centromériques ; non centromériques et associées à l'émergence de centromères), translocations (réci-proques ; robertsoniennes ; en tandem), inversions (péricentriques ; paracentriques) et relocali-

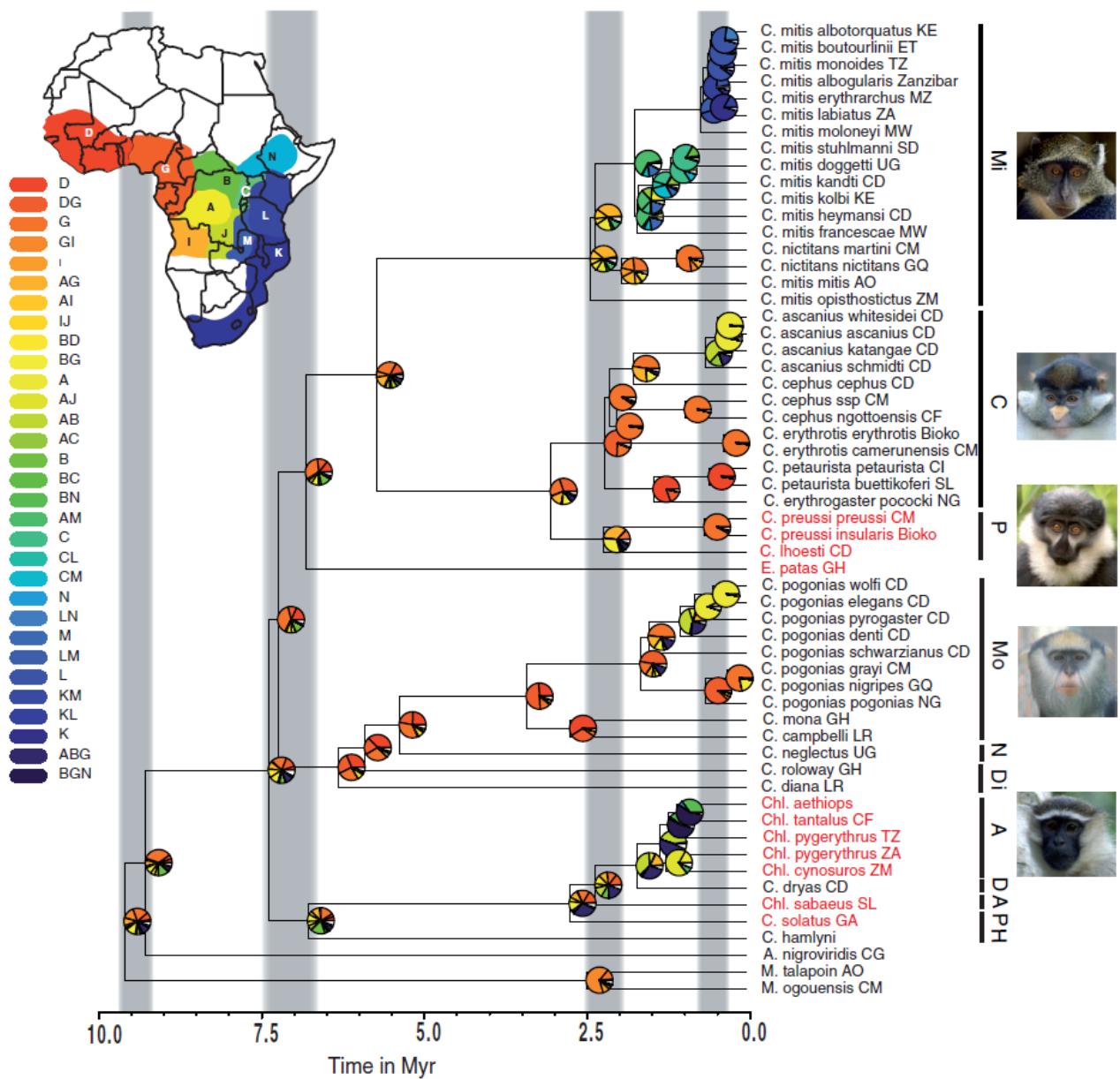


FIGURE 4.11 – Phylogénie mitochondriale des cercopithèques (d'après Guschanski et al. (2013)). Méthode de vraisemblance. Les diagrammes aux noeuds représentent les aires de distribution ancestrales. Les barres grises représentent 4 périodes principales de spéciation. Les noms colorés en rouge correspondent aux espèces classiquement assignées au clade terrestre des cercopithèques. Espèces : Mi = groupe de *C. mitis*, C = groupe de *C. cebus*, P = groupe de *C. preussi*, Mo : groupe de *C. mona*, N = groupe de *C. neglectus*, Di = groupe de *C. diana*, A = groupe de *C. aethiops*, D = groupe de *C. dryas*, H = groupe de *C. hamlyni*. Aires ancestrales : A = Bassin du Congo, B = nord de la République Démocratique du Congo, C : nord de la Vallée du Rift, Kenya, D = Haute-Guinée, G = Basse-Guinée, I = Angola, J = sud-est de la République Démocratique du Congo, K = sud-est de l'Afrique, L = nord-est de l'Afrique, M = Zambie, N = Ethiopie, Soudan. Myr = million years.

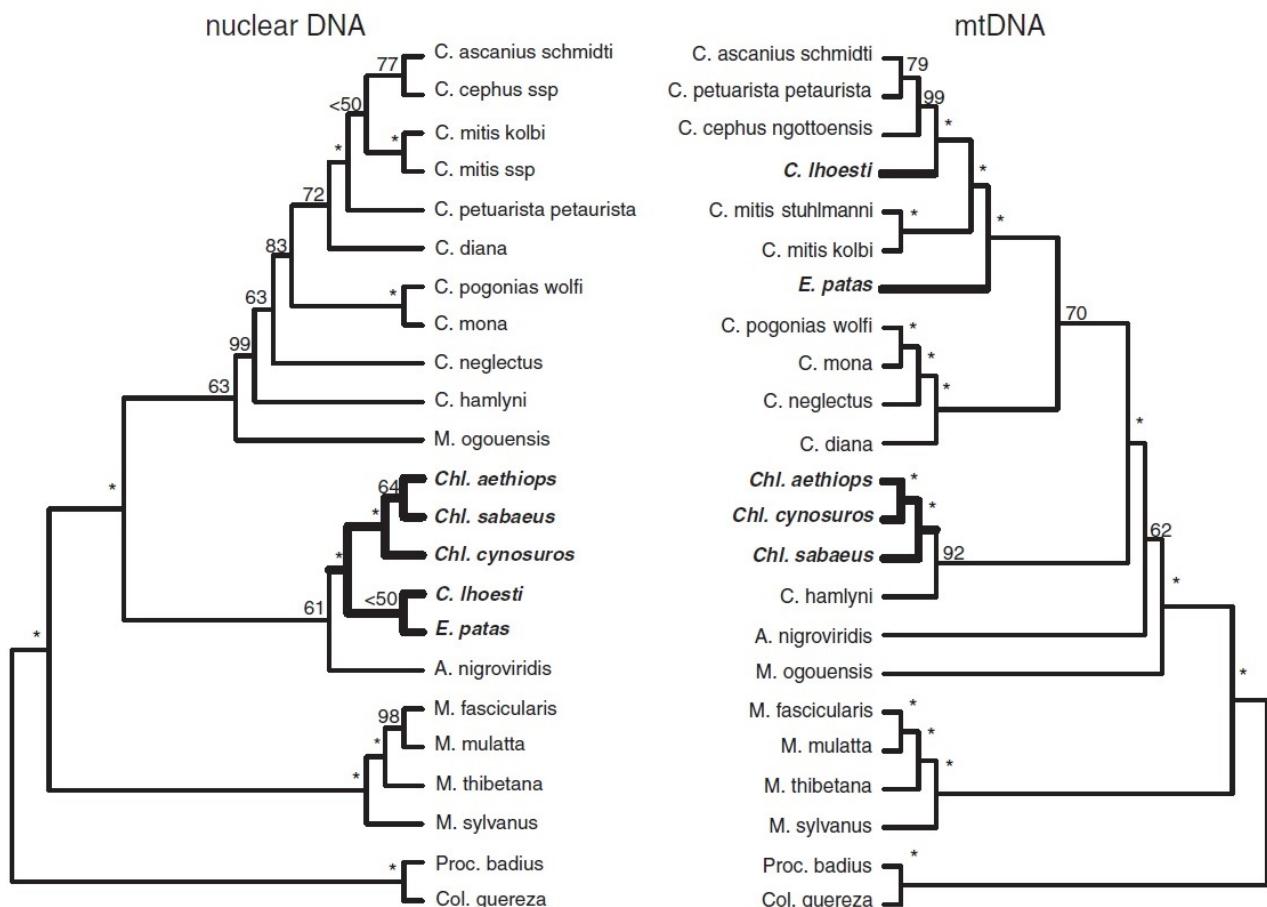


FIGURE 4.12 – Phylogénie nucléaire versus mitochondriale des cercopithèques (d'après Guschanski et al. (2013)). Méthode de vraisemblance. Les valeurs de boostrap de 100 sont représentées par un astérisque. Les noms en gras correspondent aux espèces classiquement assignées au clade terrestre des cercopithèques.

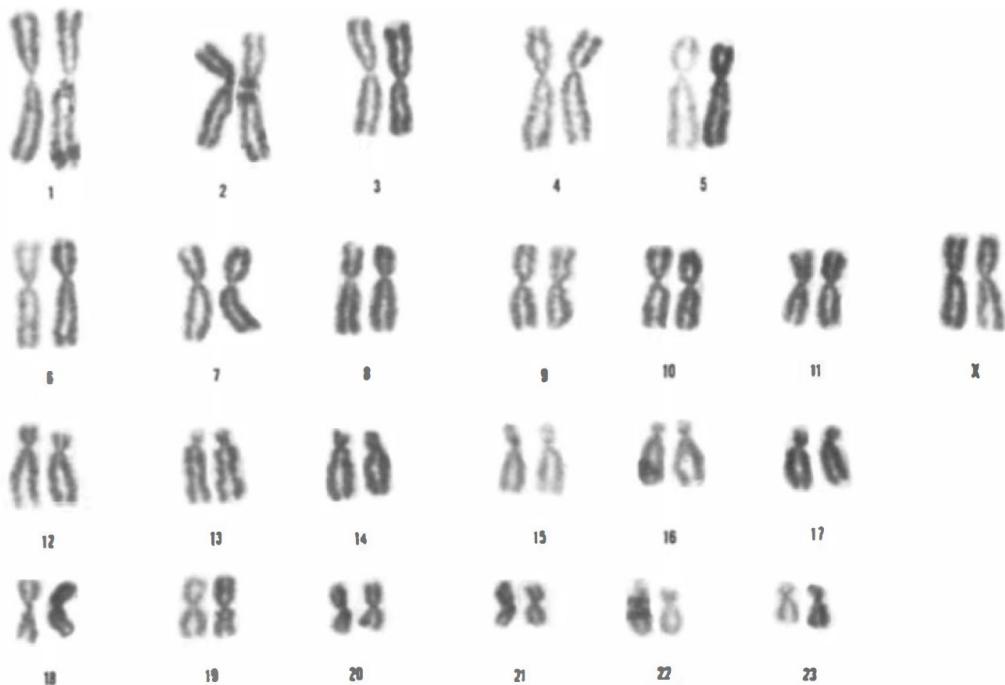


FIGURE 4.13 – Caryotype sans bandes du chimpanzé (d'après Grouchy et al. (1978)).

sations centromériques. Les chromosomes d'un génome peuvent ainsi être divisés en segments chromosomiques, conservés entre espèces mais parfois différemment associés entre eux, et porteurs ou non de centromères. La comparaison interspécifique des segments chromosomiques homologues et de leurs associations est rendue possible par diverses méthodes cytogénétiques, et peut aboutir à l'inférence des remaniements responsables de l'évolution chromosomique du clade considéré.

Aperçu historique des méthodes de cytogénétique comparative

Le premier caryotype complet de primate a été publié en 1939, pour *Macaca mulatta* ([Shiwago, 1939](#)). La cytogénétique comparative des Primates a alors principalement reposé, pendant quelques décennies, sur le nombre de chromosomes, leur taille relative et la position des centromères (Figure 4.13). Il a ainsi pu être mis en évidence des variations interspécifiques de ces paramètres, et donc que diverses espèces avaient des caryotypes différents ([Bender and Mettler, 1958](#); [Chu and Bender, 1961](#); [Borgaonkar, 1966](#)). L'hypothèse de remaniements chromosomiques évolutifs a été avancée ; par exemple, celle d'une translocation robertsonienne (fusion centromérique de deux chromosomes acrocentriques) pour expliquer les 46 chromosomes de l'homme, seul grand singe à ne pas en posséder 48. Ces hypothèses ne pouvaient cependant pas être testées avec les techniques d'alors.

Dans les années 1970, la mise au point de plusieurs techniques de marquage des chromosomes a permis de faire apparaître, le long de leurs bras, des bandes plus ou moins colorées dont la

TABLE 4.1 – Code des techniques de marquage chromosomique (d'après Popescu et al. (1998))

Code	Type de marquage chromosomique
QFQ	bandes Q, fluorescentes, colorées à la quinacrine
QFH	bandes Q, fluorescentes, colorées au Hoechst
GTG	bandes G, par action de la trypsine, colorées au Giemsa
GAG	bandes G, par action d'une solution saline acétique, colorées au Giemsa
GBG	bandes G, par incorporation de BrdU, colorées au Giemsa
CBG	bandes C, par action de l'hydroxyde de baryum, colorées au Giemsa
RFA	bandes R, fluorescentes, colorées à l'acridine orange
RHG	bandes R, par dénaturation thermique, colorées au Giemsa
RBG	bandes R, par incorporation de BrdU, colorées au Giemsa
RBA	bandes R, par incorporation de BrdU, colorées à l'acridine orange
RBP	bandes R, par incorporation de BrdU, colorées à l'iode de propidium

succession est précise, reproductible et conservée entre segments chromosomiques homologues. Trois types de marquage chromosomique sont alors principalement utilisés : les marquages en bandes Q, G et R (Table 4.1). Les bandes Q et G sont identiques et correspondent à des régions de réplication tardive, riches en AT et en SINE (Caspersson et al., 1968; Finaz and De Grouchy, 1971; Seabright, 1971; Weisblum and De Haseth, 1972; Lin et al., 1977). Les bandes R adoptent le profil inverse des bandes Q et G, et correspondent à des régions de réplication précoce, riches en GC et en LINE (Dutrillaux and Lejeune, 1971; Dutrillaux and Covic, 1974; Comings, 1973)(Figure 4.14). Le marquage dit en bandes RBP permet de faire apparaître des bandes R suite à un traitement au BrdU (5-bromo-2'-déoxyuridine) de cellules en cycle, durant la seconde moitié de leur phase S; cet analogue de la thymidine s'incorpore alors dans les régions de réplication tardive et y empêche a posteriori la fixation de l'iode de propidium, intercalant fluorescent de l'ADN (Popescu et al., 1998).

Bien que la compréhension des mécanismes à l'origine de ces bandes reste encore aujourd'hui incomplète (Popescu et al., 1998), leur profil a permis d'identifier précisément les chromosomes de chaque espèce primate étudiée, de retrouver les segments chromosomiques homologues entre de multiples espèces et d'étudier l'association différentielle de ces segments dans les génomes (Pearson et al., 1971; Finaz et al., 1973; Stock and Hsu, 1973; Dutrillaux et al., 1979, 1981)(Figure 4.15). Ainsi, la comparaison des chromosomes des grands singes a montré que le chromosome 2 humain étaient homologues à deux chromosomes chez les autres espèces, confirmant l'origine évolutive de ce chromosome par translocation robertsonienne (Turleau and De Grouchy, 1972; Grouchy et al., 1972; Lejeune et al., 1973). La conservation interspécifique de chromosomes ou d'associations de segments chromosomiques a permis l'élaboration, d'après le principe de parcimonie, des premières phylogénies chromosomiques (Dutrillaux et al., 1973, 1979, 1981; Dutrillaux and Lejeune, 1975). Les caryotypes ancestraux hypothétiques des différents clades

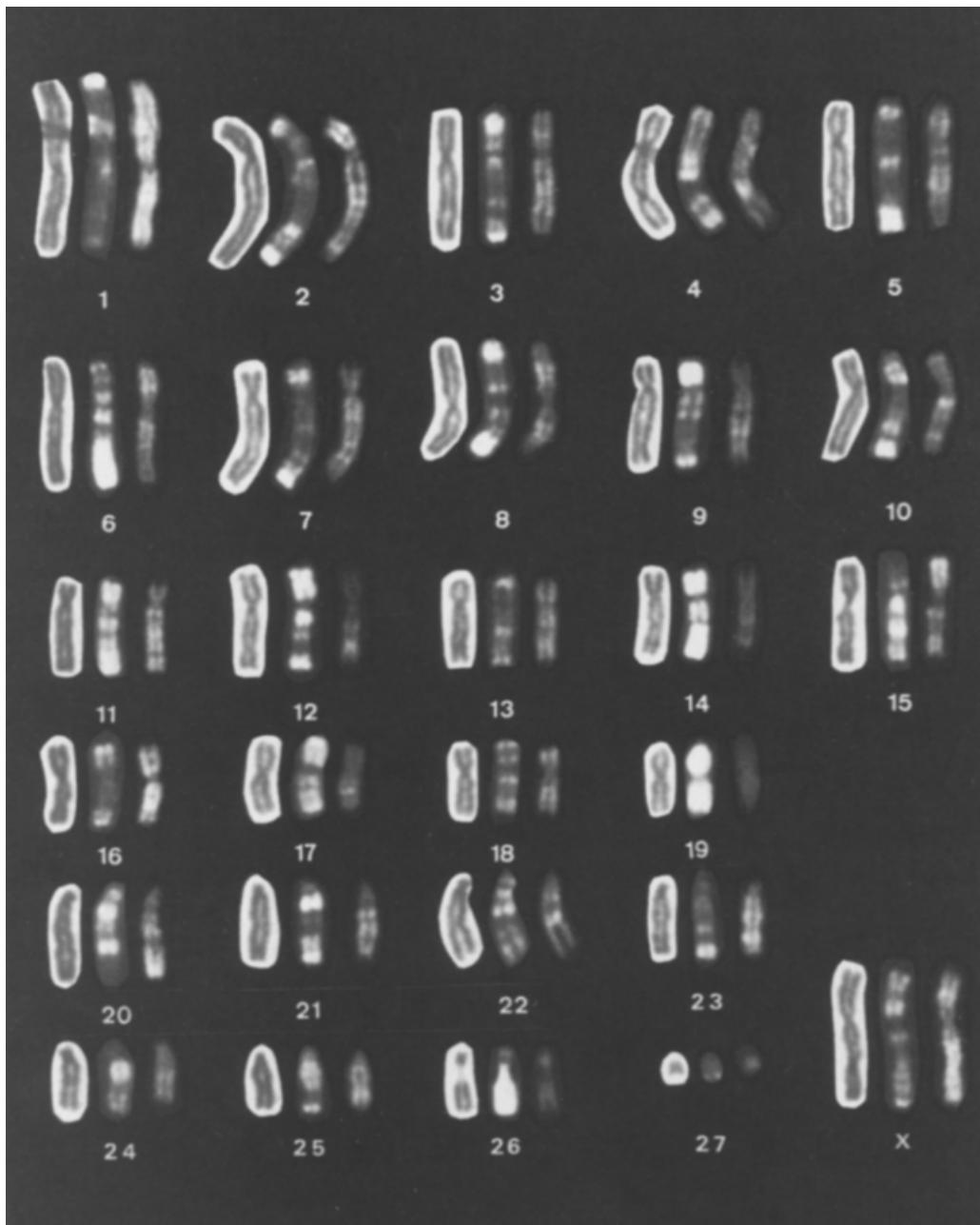


FIGURE 4.14 – Demi-caryotype de *M. talapoin* en Giemsa (gauche), bandes R (centre) et bandes Q (droite) (d'après Dutrillaux et al. (1978)). Le chromosome Y est placé en position 27.

étaient produits de manière conjointe aux phylogénies, en considérant qu'un chromosome ou une association de segments conservé chez différentes espèces leur avait été transmis pas leur ancêtre commun. *In fine*, l'évolution chromosomique d'un clade pouvait être inférée (Dutrillaux and Couturier, 1986) ; celle-ci s'est révélée très différente en fonction des grands groupes de primates, présentant pour chacun des remaniements évolutifs prédominants : comme les inversions chez les Hominidae et les Platyrrhini, les fissions et fusions chez les Cercopithecidae ou encore les translocations robertsoniennes chez les Strepsirrhini (Figure 4.16).

La cytogénétique comparative via le marquage en bandes des chromosomes s'est cependant heurtée aux espèces de la famille des gibbons : trop peu de chromosomes ou associations de segments chromosomiques ont été retrouvés conservés entre elles pour permettre de raconter l'histoire chromosomique de ce clade, et leur comparaison à l'homme s'est également révélée particulièrement délicate (Dutrillaux et al., 1975; Tantravahi et al., 1975). Le développement de la cytogénétique moléculaire comparative, dans les années 1990, a été initié pour justement palier à ce problème et rendre possible la comparaison interspécifique des chromosomes par FISH. Des peintures chromosomiques, ensembles de sondes contenant l'ADN d'un chromosome particulier, ont pour cela été utilisées (Cremer et al., 1988; Lichter et al., 1988; Pinkel et al., 1988). D'abord conçues à partir de librairies génomiques spécifiques de chromosomes, le perfectionnement du tri chromosomique par cytométrie en flux a depuis fait de cette méthode la source principale d'ADN spécifique de chromosomes pour l'élaboration de ces peintures (Gray et al., 1975; Telenius et al., 1992; Ferguson-Smith et al., 2004; Stanyon and Stone, 2008) ; la séparation des chromosomes y repose sur leur intensité de fluorescence suite à un marquage par des fluorochromes AT- ou GC-spécifiques. Un excès d'ADN enrichi en éléments répétés est systématiquement ajouté au cocktail d'hybridation d'un chromosome, afin de confiner l'hybridation des sondes à ce chromosome ou ses segments chromosomiques homologues ; on parle d'hybridation suppressive.

L'hybridation réciproque des peintures chromosomiques de diverses espèces a permis de confirmer ou corriger les homologies chromosomiques déterminées via les bandes, et d'en proposer de nouvelles ; tout en précisant la plupart des points de cassures impliqués dans les remaniements évolutifs survenus chez les Primates (Wienberg et al., 1990, 1992; Jauch et al., 1992; Stanyon et al., 1992, 1995; Canavez et al., 1996; Consigliere et al., 1996, 1998; Richard et al., 1996; Bigoni et al., 1997; Morescalchi et al., 1997; Best et al., 1998; Dumas et al., 2007)(Figures 4.17 et 4.18). L'élaboration des phylogénies chromosomiques se trouve facilitée et améliorée, tout comme celle des caryotypes ancestraux (Müller et al., 1999). L'évolution chromosomique des gibbons a enfin pu être retracée ; leurs génomes sont apparus largement remaniés par des translocations réciproques (Jauch et al., 1992; Koehler et al., 1995; Arnold et al., 1996; Schrock et al., 1996; Müller et al., 1997, 1998, 2003).

Les peintures chromosomiques restent toutefois de faible utilité pour l'étude des remaniements intrachromosomiques ; dans les années 2000, l'arrivée de la méthode de BAC-FISH dans le champ de la cytogénétique comparative permet de palier à ce défaut. Grâce à des sondes ci-

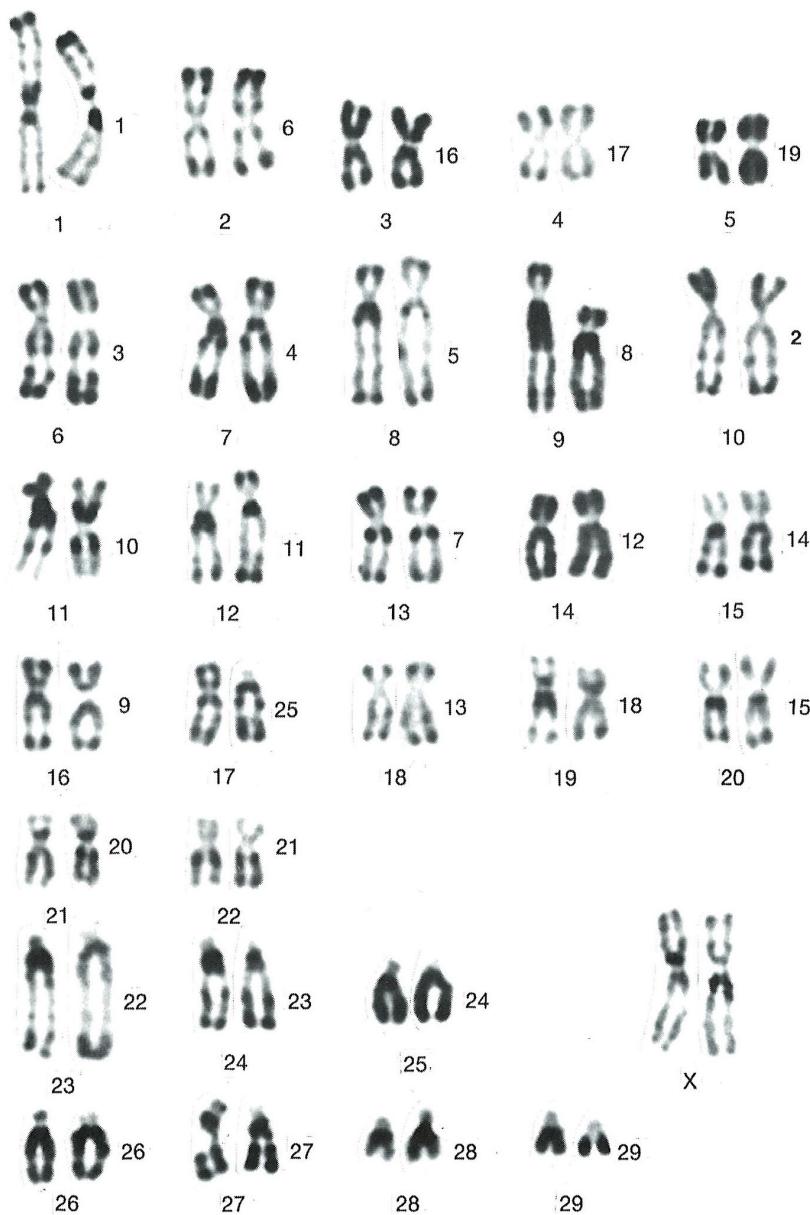


FIGURE 4.15 – Comparaison des caryotypes en bandes R de *C. solatus* (gauche) et *C. lhoesti* (droite) (d'après Dutrillaux et al. (1988a)). Les chromosomes sont classés selon le caryotype de *C. solatus*. Les chromosomes sont très conservés entre ces deux espèces ; on peut remarquer entre elles une inversion péricentrique ou relocalisation centromérique sur les paires homologues aux chromosomes 11 et 17 de *C. solatus*.

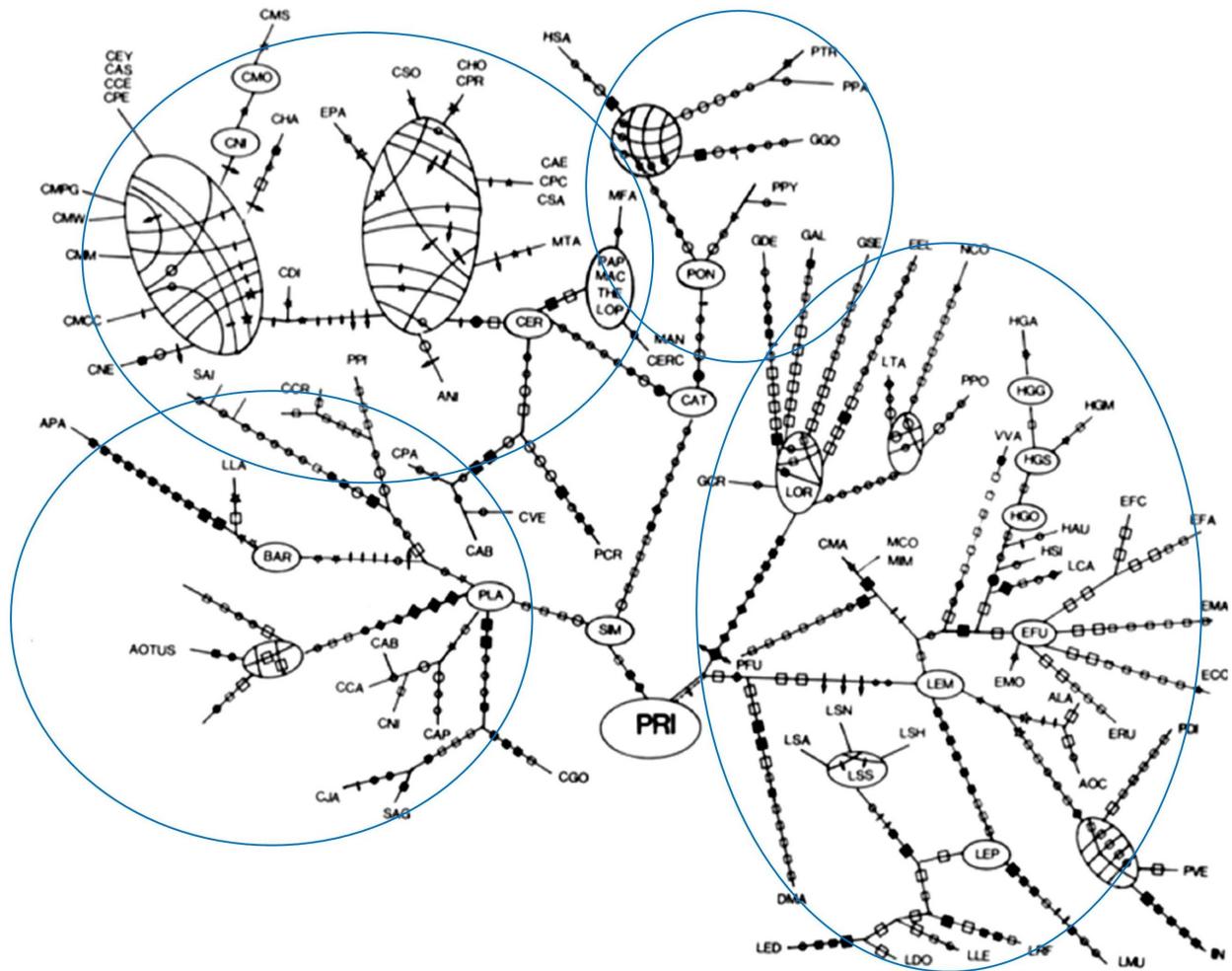


FIGURE 4.16 – Phylogénie chromosomique et prédominance des types de remaniements au cours de l'évolution des Primates (d'après Rumpler and Dutrillaux (1989)). Flèche = fission ou fusion, Cercle = inversion, Carré = translocation. Cercopithecidae (en haut à gauche) : fissions et fusions. Hominidae (en haut à droite) : inversions. Platyrrhini (en bas à gauche) : inversions et translocations. Strepsirrhini (en bas à droite) : translocations.

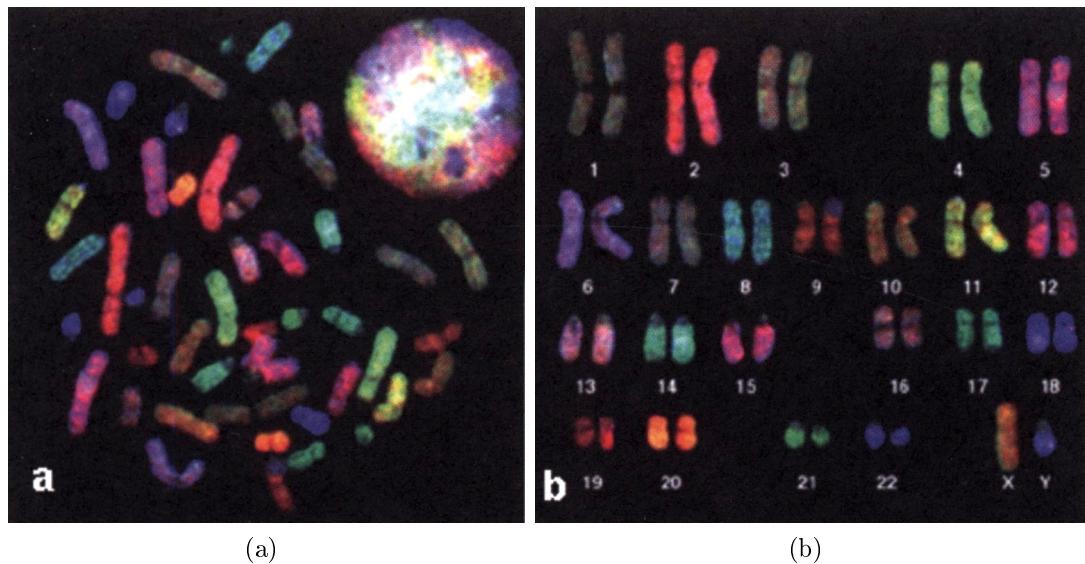


FIGURE 4.17 – Peintures chromosomiques humaines hybridées sur les chromosomes humains (d'après Schrock et al. (1996) et Wienberg and Stanyon (1998)). (A) Etalement métaphasique. (B) Caryotype.

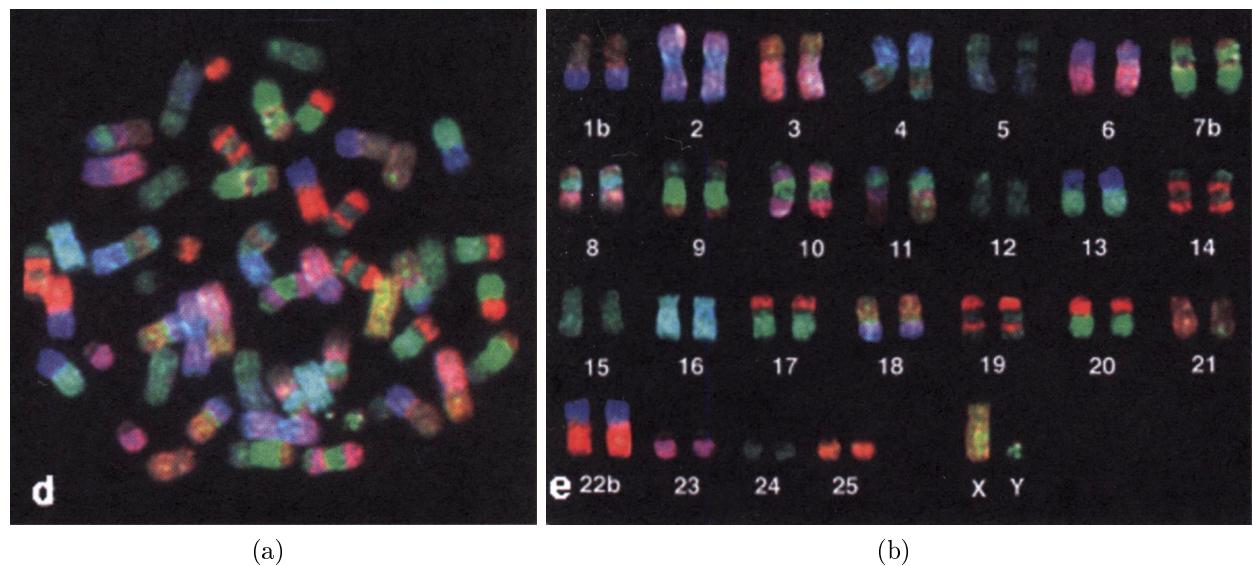


FIGURE 4.18 – Peintures chromosomiques humaines hybridées sur les chromosomes du gibbon *Hylobates syndactylus* (d'après Schrock et al. (1996) et Wienberg and Stanyon (1998)). (A) Etalement métaphasique. (B) Caryotype.

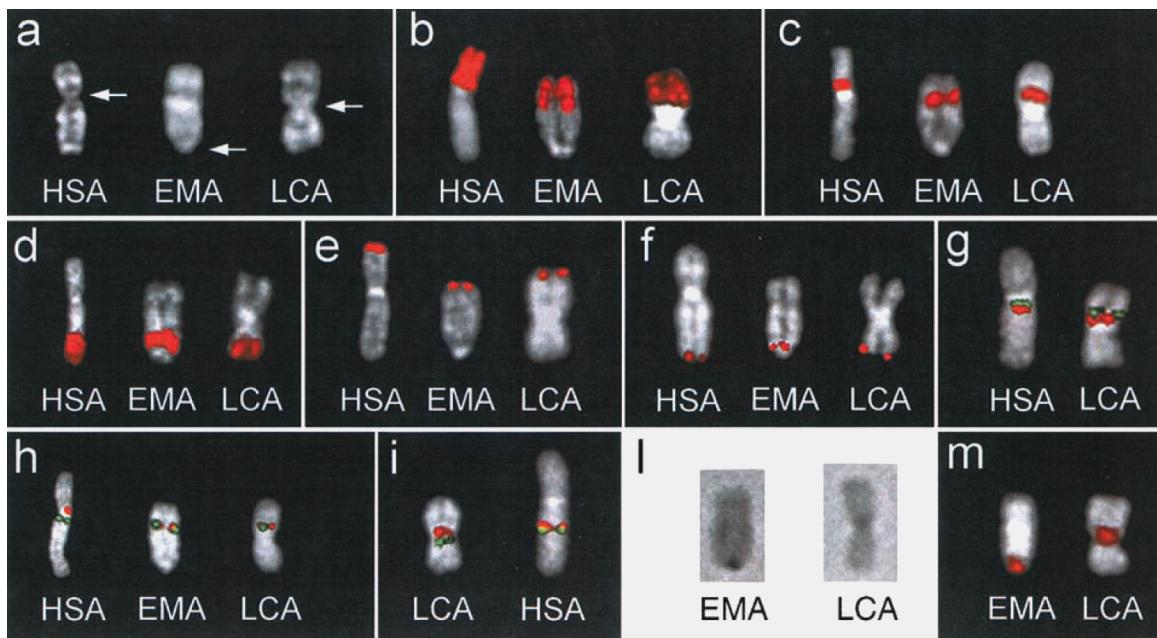


FIGURE 4.19 – Relocalisations centromériques sur le chromosome X au cours de l'évolution des Primates (d'après [Ventura et al. \(2001\)](#)). (A) Chromosome X de l'homme (*Homo sapiens*, HSA), du lémur noir (*Eulemur macaco*, EMA) et du lémur catta (*Lemur catta*, LCA), marquage en bandes Q et orientation homologue. Les flèches pointent les centromères, différemment positionnés sur l'X entre les espèces. (B-F) BAC-FISHs avec des sondes homologues. (G-I) Co-BAC-FISHs avec des sondes homologues. (L) Chromosome X en bandes C des lémuriens. (M) BAC-FISH sur l'X des lémuriens avec des sondes élaborées à partir de matériel disséqué depuis leurs centromères respectifs.

blant quelques centaines de kilobases seulement, l'ordre et l'association de courts segments chromosomiques peuvent être déterminés le long d'un chromosome, et comparés avec ceux retrouvés sur son homologue dans d'autres espèces. Cette méthode s'est avérée particulièrement utile pour distinguer les inversions péricentriques (avec des points de cassures de part et d'autre des centromères) et relocalisations centromériques ; et donc mettre en évidence les nouveaux centromères évolutifs présents dans les génomes des Primates (Montefalcone et al., 1999; [Ventura et al., 2001, 2007](#); Stanyon et al., 2008)(Figure 4.19). Il a ainsi pu être montré que des loci homologues étaient utilisés indépendamment au cours de l'évolution pour l'émergence des centromères (Stanyon et al., 2008). La BAC-FISH comparative a par ailleurs permis d'affiner la délimitation des points de cassures évolutifs et de montrer que là encore de même points étaient indépendamment utilisés au cours de l'évolution (Roberto et al., 2007; Francesca and Luca, 2010; Capozzi et al., 2012; Sangpakdee et al., 2016). Tout comme pour les peintures chromosomiques, la BAC-FISH comparative est utilisée pour continuer à proposer des phylogénies chromosomiques et des caryotypes ancestraux (Figure 4.20).

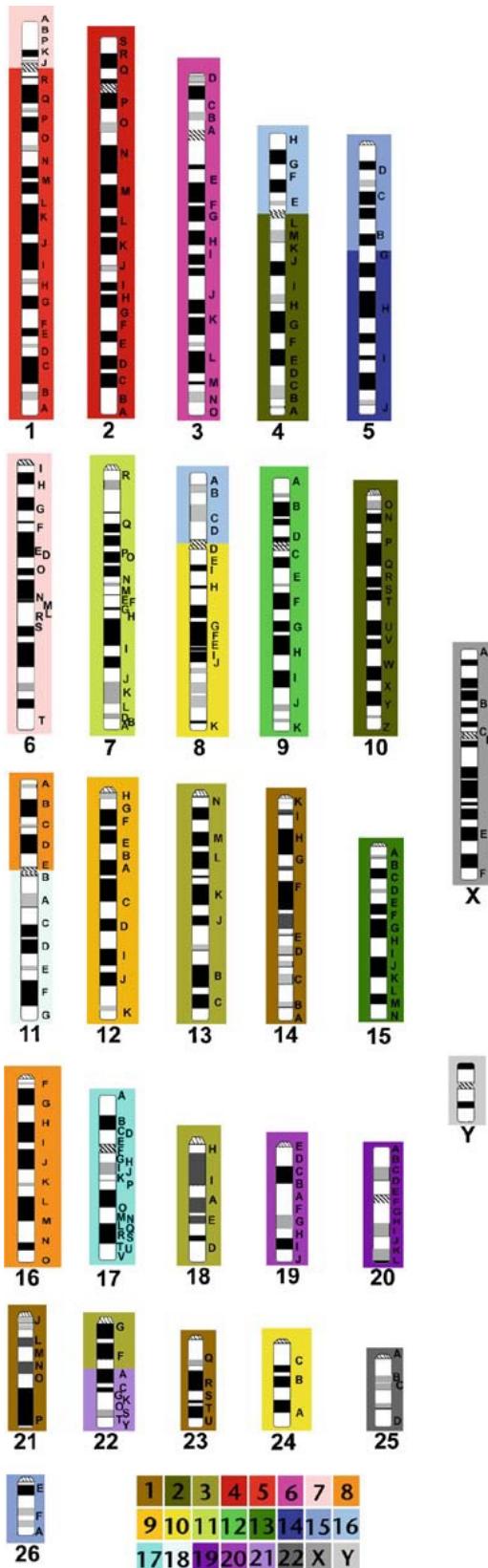


FIGURE 4.20 – **Caryotype ancestral des Platyrhini** (d'après Stanyon et al. (2008)). Les chromosomes sont classés en fonction de leur taille et colorés selon leurs homologies humaines. Les lettres assignées aux segments chromosomiques font référence à l'ordre de ces segments chez l'homme; ces segments ont fait l'objet d'études de BAC-FISH chez différentes espèces de Platyrhini.

Evolution des chromosomes chez les cercopithèques

Les cercopithèques sont, parmi les Cercopithecidae, la tribu la plus diversifiée en caryotypes. Les chromosomes de 26 espèces ont été étudiés jusqu'ici, via des méthodes de cytogénétique classique comme moléculaire ; leur nombre varie de 48 chez *A. nigroviridis* à 72 chez *C. pogonias*, *C. wolffii* et *C. albogularis* (Chiarelli, 1968; Dutrillaux et al., 1979, 1988a; Moulin et al., 2008). La comparaison interspécifique de leurs caryotypes et de ceux des autres Catarrhini a permis de reconstruire le caryotype ancestral hypothétique des cercopithèques, qui présente 46 chromosomes, tout comme celui des Cercopithecidae (Dutrillaux et al., 1979, 1981). Il a ainsi pu être déterminé que de nombreux phénomènes de fissions chromosomiques s'étaient produits durant l'évolution des cercopithèques, expliquant ainsi leur diversification caryotypique. De manière originale chez les Mammifères, ces fissions seraient pour la plupart non centromériques (ou non robertsoniennes) ; elles se seraient produites le long des bras chromosomiques, et non aux centromères (Dutrillaux et al., 1979; Finelli et al., 1999; Stanyon et al., 2004; Francesca and Luca, 2010). Ce type de fission, lorsque conservée, s'accompagne donc de l'émergence d'un nouveau centromère évolutif sur l'un des deux néo-chromosomes, lui permettant de ségrégner au cours des divisions et d'être maintenu dans les cellules. L'existence de centromères et télo-mères latents dans les génomes a ainsi été proposée pour expliquer la récurrence des fissions non centromériques chez les cercopithèques (Dutrillaux et al., 1979; Clemente et al., 1990). Par ailleurs, il a également été noté que la position des centromères différait régulièrement sur les chromosomes homologues des diverses espèces ; des inversions péricentriques pourraient expliquer ce phénomène (Dutrillaux et al., 1979), tout comme des relocalisations centromériques (Moulin et al., 2008). La mise au point et l'utilisation de BAC-FISH sur les chromosomes des cercopithèques seraient nécessaires pour déterminer dans quelle mesure ces deux remaniements ont pris part à leur évolution.

L'analyse descriptive des remaniements chromosomiques partagés par les différentes espèces de cercopithèques a pu montrer que ceux-ci s'étaient très différemment accumulés dans leurs génomes (Dutrillaux et al., 1979; Moulin et al., 2008). Deux remaniements seulement sont synapomorphiques à cette tribu et retrouvés chez toutes les espèces étudiées : une fission du chromosome homologue au chromosome 3 humain, et une fusion des segments homologues aux chromosomes 20 et 21 humains. Le caryotype d'*A. nigroviridis* ($2n = 48$) a accumulé très peu de remaniements depuis la divergence des cercopithèques ; il présente une fission seulement par rapport au caryotype hypothétique de leur ancêtre commun. Le clade terrestre des cercopithèques (*Chlorocebus* ($2n = 60$), *Erythrocebus* ($2n = 54$), *C. lhoesti*, *C. preussi* et *C. solatus* ($2n = 60$)) est soutenu par deux fissions synapomorphiques partagées par toutes ses espèces ; plusieurs autres fissions chromosomiques ont ponctué l'évolution de chacune, mais leur occurrence reste limitée. Le caryotype de *Miopithecus* ($2n = 54$) apparaît proche de ceux des espèces du clade terrestre. Le clade arboricole des cercopithèques (tous les *Cercopithecus* à l'exception de *C. lhoesti*, *C. preussi* et *C. solatus*) est supporté par trois fissions synapomorphiques partagées par toutes ses espèces. Certaines n'ont accumulé qu'une autre fission supplémentaire, comme celles du groupe dit de *C. diana* (*C. diana* et *C. roloway* ($2n = 58$))). La plupart ont cependant

vu leur génome profondément remanié, en particulier dans les groupes dits de *C. mona* (*C. mona* (2n = 68), *C. pogonias* et *C. wolfi* (2n = 72)), *C. cephush* (*C. cephush*, *C. erythrotis*, *C. ascanius*, *C. erythrotis* et *C. erythrogaster* (2n = 66)) et *C. mitis* (*C. mitis*, *C. nictitans* (2n = 70) et *C. albogularis* (2n = 72)). Les espèces du clade arboricole auraient ainsi vu émerger dans leurs génomes de nombreux nouveaux centromères évolutifs, de manière associée à ces fissions récurrentes.

L'origine des remaniements chromosomiques et leur impact sur la spéciation font l'objet de nombreuses interrogations et théories (Brown and O'Neill, 2010). Des architectures épigénétiques particulières pourraient faciliter les remaniements au cours de l'évolution, comme l'hypométhylation des points de cassures observée dans les génomes des gibbons (Carbone et al., 2009). Il a par ailleurs été proposé que, suite à l'apparition d'un remaniement chromosomique, les flux géniques entre le chromosome remanié et son homologue vierge soient réduits ; ils peuvent en effet rencontrer des difficultés à recombiner lors de la méiose (Rieseberg, 2001; Navarro and Barton, 2003a,b). Leur ADN respectif peut alors diverger considérablement si le remaniement se fixe dans sa population d'origine, et l'évolution indépendante des gènes concernés peut mener progressivement à une barrière reproductive entre individus porteurs et non porteurs du remaniement (Coluzzi, 1981; Rieseberg et al., 1999; Ayala and Coluzzi, 2005). D'autre part, cette divergence génétique peut mener à une adaptation différentielle des individus à certaines niches écologiques (Figure 4.21) ; ces deux phénomènes peuvent être à l'origine de la formation de nouvelles espèces. Il est ainsi possible que la diversification chromosomique des cercopithèques ait joué un rôle moteur dans leurs phénomènes de spéciation et dans leur adaptation à différents écosystèmes.

Conclusion

Les cercopithèques sont un clade à la richesse spécifique élevée et à l'évolution chromosomique originale. Les réarrangements ayant modelé leurs génomes laissent espérer la présence, chez certaines espèces, de nombreux nouveaux centromères évolutifs ; les cercopithèques apparaissent donc comme d'excellents modèles pour une étude comparative des séquences alpha satellites. Au-delà de l'apport de ces espèces à la compréhension de la dynamique évolutive des régions centromériques, il serait intéressant de découvrir si l'ADN alpha satellite pourrait à son tour nous éclairer sur l'histoire évolutive des cercopithèques, dont les relations de parenté restent encore aujourd'hui à préciser.

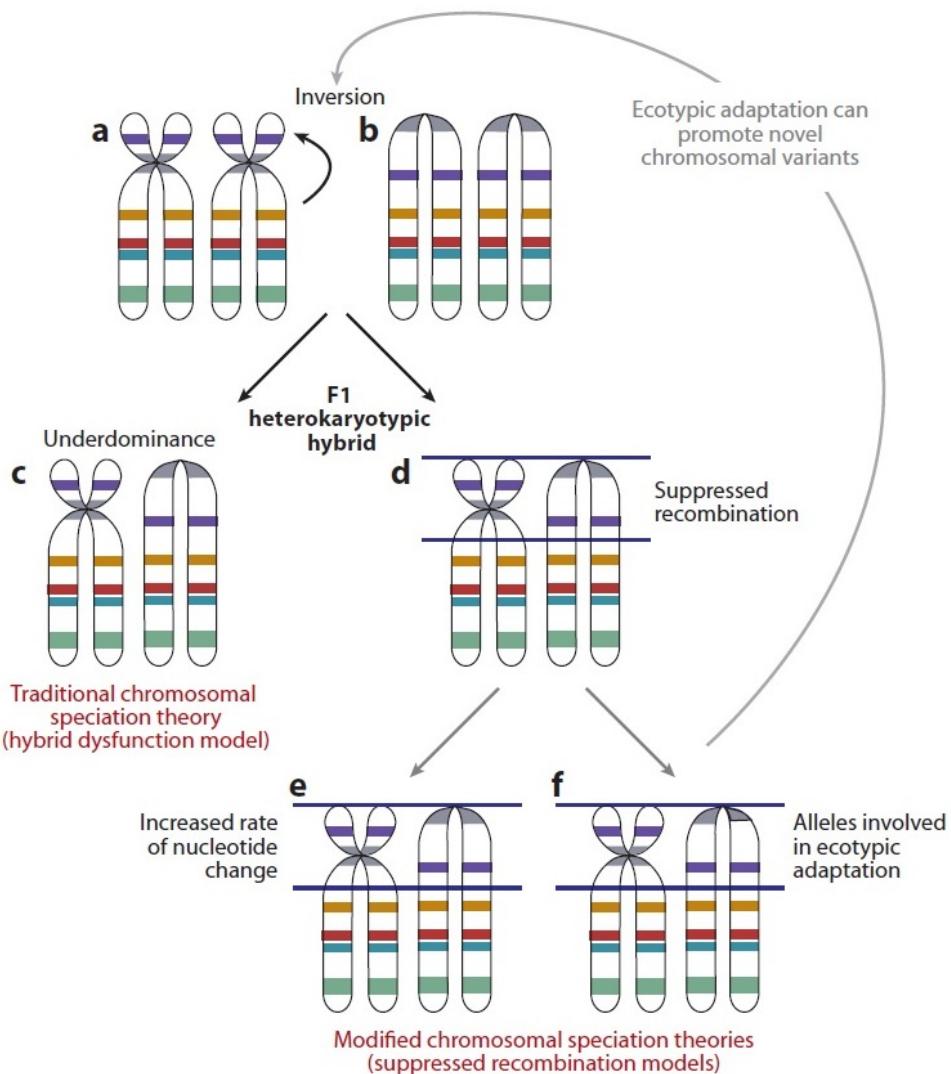


FIGURE 4.21 – Modèles de spéciation chromosomique (d'après Brown and O'Neill (2010)). L'exemple d'une inversion péricentrique est considéré ici. (A) Individu porteur de la forme métacentrique. (B) Individu porteur de la forme acrocentrique. (C et D) Hybride porteur des deux formes. Le modèle traditionnel de spéciation chromosomique (C) stipule que cet hybride sera stérile, ce qui constitue une barrière reproductive entre les individus porteurs de formes différentes. Le modèle modifié de spéciation chromosomique (D) stipule que les chromosomes concernés vont avoir des difficultés à recombiner, ce qui peut mener à (E) une accumulation rapide de mutations nucléotidiques différentes ou (F) la fixation d'allèles différents pouvant promouvoir une adaptation à des milieux différents.

1.5 Projet de thèse

Afin d'améliorer nos connaissances quant à la diversité de l'ADN alpha satellite chez les Primates et de mieux comprendre sa dynamique évolutive, nous avons tout d'abord caractérisé les séquences alpha satellites des espèces *C. solatus* ($2n = 60$) et *C. pogonias* ($2n = 72$), qui appartiennent respectivement au clade terrestre et au clade arboricole des cercopithèques. Notre étude de l'ADN alpha satellite s'est par la suite étendue à treize espèces de cercopithèques ainsi qu'à deux de leurs espèces proches. L'utilisation des nouvelles technologies de séquençage, de la bioinformatique et de la cytogénétique classique et moléculaire nous a permis d'apporter des éléments de réponse aux questions suivantes :

- Quelle est la diversité de l'ADN alpha satellite chez les cercopithèques ?
- Comment se distribuent les familles d'alpha satellites dans leurs génomes ?
- Quelle est l'histoire évolutive de l'ADN alpha satellite au sein de ce clade ?

Les deux premiers articles de cette thèse présentent ainsi les différentes familles d'alpha satellites identifiées dans les génomes de *C. solatus* et *C. pogonias*, les organisations qu'elles y adoptent et leur distribution intra- et interchromosomique. Le troisième article de cette thèse présente la distribution spécifique et chromosomique de ces familles chez les cercopithèques, et introduit la diversité de l'ADN alpha satellite sur les nouveaux centromères évolutifs. La dynamique d'émergence et d'extinction des familles d'alpha satellites y est discutée, ainsi que sa relation à la dynamique des remaniements chromosomiques. L'apport des connaissances sur l'ADN alpha satellite à celles de l'histoire évolutive des cercopithèques y est également abordé.

Chapitre 2

Résultats

2.1 Nouvelle approche d'étude de l'ADN alpha satellite appliquée à *Cercopithecus solatus*

Premier article - Diversity and distribution of alpha satellite DNA in the genome of an Old World monkey : *Cercopithecus solatus*

Lauriane Cacheux^{1,2}, Loïc Ponger¹, Michèle Gerbault-Seureau², Florence Anne Richard^{2,3} and Christophe Escudé¹

¹Structure et Instabilité des Génomes, U1154 INSERM, UMR7196 CNRS, Sorbonne Universités, Muséum national d'Histoire naturelle, Département Régulations, Développement et Diversité Moléculaire, Paris, France

²Institut de Systématique, Evolution, Biodiversité, UMR7205 CNRS, UPMC, EPHE, Sorbonne Universités, Muséum national d'Histoire naturelle, Département Systématique et Evolution, Paris, France

³Université Versailles St-Quentin, Montigny-le-Bretonneux, France

2.1.1 Abstract

Alpha satellite is the major repeated DNA element of primate centromeres. Evolution of these tandemly repeated sequences has led to the existence of numerous families of monomers exhibiting specific organizational patterns. The limited amount of information available in non-human primates is a restriction to the understanding of the evolutionary dynamics of alpha satellite DNA.

We carried out the targeted high-throughput sequencing of alpha satellite monomers and dimers from the *Cercopithecus solatus* genome, an Old World monkey from the Cercopithecini tribe. Computational approaches were used to infer the existence of sequence families and to study how these families are organized with respect to each other. While previous studies had suggested that alpha satellites in Old World monkeys were poorly diversified, our analysis provides evidence for the existence of at least four distinct families of sequences within the studied species, two of them belonging to a higher order organizational pattern. Fluorescence in situ hybridization using oligonucleotide probes that are able to target each family in a specific way showed that the different families had distinct distributions on chromosomes and were not homogeneously distributed between chromosomes.

Our new approach provides an unprecedented and comprehensive view of the diversity and organization of alpha satellites in a species outside the hominoid group. We consider these data with respect to previously known alpha satellite families and to potential mechanisms for satellite DNA evolution. Applying this approach to other species will open new perspectives regarding the integration of satellite DNA into comparative genomic and cytogenetic studies.

Keywords : Alpha satellite DNA, High-throughput sequencing, *Cercopithecus solatus*, Centromere genomics

2.1.2 Introduction

Centromeres are chromosomal regions that control chromosome segregation during cell division in eukaryotes, through kinetochore assembly and microtubule attachment. In almost all eukaryotes, the DNA underlying centromeres is made of large tracts of nearly identical tandem DNA repeats, known as satellite DNA ([Yunis and Yasmineh, 1971](#); [Warburton and Willard, 1990](#); [Schueler et al., 2001](#)). The remarkable variation of satellite DNAs between species has been an enigma ever since their discovery and different important roles have been ascribed to these sequences, from the imperative centromeric function in mitosis and meiosis to regulatory functions ([Plohl et al., 2008](#); [Feliciello et al., 2015](#)).

Alpha satellite DNA is the most abundant satellite DNA in Primates and is found both at the site of centromere attachment and in neighboring heterochromatic regions, referred to as

pericentromeres (She et al., 2004). Alpha satellite DNA was originally isolated as a highly repetitive component of the *Chlorocebus aethiops* (also called African green monkey) genome (Maio, 1971); homologous repeats were then described throughout the Primate order including apes, Old World and New World monkeys (Musich et al., 1980; Maio et al., 1981; Alves et al., 1994). Alpha satellite DNA is made of tandemly repeated AT-rich monomers that are about 170 bp in length and organized in head-to-tail orientation (Willard, 1991; Rudd et al., 2006). In the human genome, individual monomers share between 60 and 100 % sequence identity. The highly identical composition of successive repeats represents a technical challenge that has thwarted the complete assembly of centromeric DNA so far (Rudd and Willard, 2004; Miga, 2015). Nevertheless, over the last 30 years the systematic cloning and sequencing of many alpha satellite DNAs, combined with fluorescence in situ hybridization (FISH) experiments, has provided a thorough knowledge of alpha satellite DNA diversity and organizational patterns in the human genome (Willard, 1991; Schueler and Sullivan, 2006; Miga et al., 2014) and, to a much lesser extent, in other primates (Alexandrov et al., 2001; Alkan et al., 2007; Shepelev et al., 2009; Catacchio et al., 2015).

In human, alpha satellite DNA has been shown to adopt two different organizations. In the so-called higher order repeat (HOR) organizational pattern, highly conserved repeat units (97–100 % sequence identity), each made of multiple 171 bp monomers (up to more than 30), are found as an homogenized array that can extend over a multimegabase-sized region (Willard and Waye, 1987; Warburton and Willard, 1990; Alexandrov et al., 1993; Rudd and Willard, 2004; Hayden, 2012). This organization is typically found as very long arrays of alpha satellites at the centromere core of all human chromosomes. In pericentromeres, a second type of organization, called monomeric and involving arrays of single alpha satellite monomers which are less well conserved (70–90 % sequence identity), can coexist with HORs (Schueler et al., 2001; Rudd et al., 2006). Sequence comparisons between human alpha satellite monomers have led to the description of up to seventeen different alpha satellite families, or monomer types (Willard and Waye, 1987; Alexandrov et al., 1988; Lee et al., 1997; Shepelev et al., 2009). Although the alpha satellite component of other primate genomes has been less intensively studied, there is some evidence for similar organizations in great apes, but additional families have been described and the composition of HORs as well as their chromosomal distribution differ when compared with human (Jørgensen et al., 1987; Archidiacono et al., 1995; Warburton et al., 1996; Rudd et al., 2006; Catacchio et al., 2015). This implies that the structure and content of centromeric DNA can change in a few million years.

Although the mechanisms that gave rise to this diversity and organization are not precisely known, it is commonly accepted that the so-called concerted evolution of repetitive sequences is based on different mechanisms of non-reciprocal transfer occurring within or between chromosomes, such as unequal crossover, gene conversion, rolling circle replication and reinsertion, and transposon-mediated exchange (Malik and Henikoff, 2002; Plohl et al., 2012). Such mechanisms enable series of amplification events, thereby creating new arrays of alpha satellites (Warburton and Willard, 1995; Schindelhauer and Schwarz, 2002; Roizès, 2006; Rudd et al.,

2006; Miga et al., 2014). The analysis of the different alpha satellite families found in assembled pericentromeres from specific human chromosomes revealed an age gradient of the families along each chromosome arm, which lead to propose that during the course of evolution, new arrays of alpha satellites expand at the centromere core, thereby splitting and displacing older arrays distally onto each arm (Schueler et al., 2001; Rudd and Willard, 2004; She et al., 2004; Schueler et al., 2005; Shepelev et al., 2009).

Knowledge about alpha satellite DNA in species outside the hominoid group is very scarce, in particular in Old World monkeys, a clade that includes Colobinae, Papionini and Cercopithecini. The tribe Cercopithecini contains 35 species which have diversified within the last 10 million years (Wilson and Reeder, 2005; Guschanski et al., 2013) and therefore represents a particularly interesting group for studying the evolution of satellite DNA. Moreover, it has been reported that alpha satellite DNA is more abundant in some Cercopithecini species (up to 20 % of the genome of *C. aethiops*) (Madhani et al., 1986) than in great apes, where its contribution would reach only 3 % of the genome (Miga, 2015). Finally, enzymatic digestion of genomic DNA from various Old World monkey species can lead to a clear alpha satellite ladder pattern which is not observed when human or chimpanzee DNA is used, thereby pointing to different composition and organization of alpha satellite DNA in Old World monkeys (Fittler, 1977).

In the present work, we have undertaken the targeted sequencing of the alpha satellite component of *Cercopithecus solatus* (or Sun-tailed monkey) as a representative species for Cercopithecini (Harrison, 1988). Alpha satellite monomers and dimers were obtained by enzymatic digestion of genomic DNA and gel purification, then submitted to high-throughput sequencing. The obtained sequences were analyzed and classified into monomer families using computational approaches. Finally the genomic distribution of each family was studied by FISH using a collection of oligonucleotide probes that are able to distinguish different sequence variants. Our study provides evidence for the existence of two main families of monomers which differ in their chromosomal distribution, one being specifically distributed on centromeres while the other is found only at pericentromeric locations with a non-uniform distribution between chromosomes. Two other families are detected which are only found associated within a dimeric organization and are located for the greatest part on the Y chromosome and to a lesser extent on pericentromeres from other chromosomes. These data represent the most complete analysis of the diversity and distribution of alpha satellite sequences in an Old World monkey reported to date. Our experimental approach may be applied to other species, opening new perspectives regarding the integration of satellite DNA into comparative studies.

2.1.3 Results

Retrieval of alpha satellite sequences from the *C. solatus* genome

Work conducted in the early 1980s had shown that enzymatic digestion of genomic DNA from Old World monkeys with several restriction enzymes resulted in a migration profile that was characteristic for alpha satellite DNA, i.e. with bands corresponding to one and multiple repeat units of about $n \times 170$ bp in length (Musich et al., 1980; Lee and Singer, 1982). In silico analysis of several sequences isolated from *C. aethiops* led us to select the XmnI restriction endonuclease as a candidate that should cleave a majority of monomers. Experimental digestion of *C. solatus* genomic DNA with this enzyme revealed the expected banding pattern (Supplementary figure 1.1). We therefore decided to extract DNA from two bands corresponding to monomers and dimers of alpha satellites from an agarose gel and implemented high-throughput sequencing on an Ion Torrent sequencing platform providing reads up to 400 nucleotide in length (see Methods).

204,990 and 353,683 raw sequences were obtained for the monomer and dimer samples, respectively. Four in silico filters were applied successively to both datasets : a quality filter keeping sequences with a Phred quality score superior to 25 ; an extremity filter keeping sequences with the XmnI restriction site at both ends ; a length filter keeping sequences within the range 162-182 bp for monomers and 324-364 bp for dimers, and an alpha satellite filter keeping sequences similar to an alpha satellite reference sequence (see Methods). The number of sequences that remained after each filter is reported on Supplementary table 1.1. A total of 100,713 sequences fitting with all the criteria was obtained from the monomer sample and represents what we call from now on the monomer dataset. For the dimer sample, only 3,568 sequences were obtained, they represent the dimer dataset. The drastic reduction observed within the dimer dataset was mostly the consequence of the length filter and may reflect an intrinsic limitation of the sequencing technology, unable to obtain long reads when template sequences are made of two successive highly identical sequences. These sequences were nevertheless included for further analysis as they provided an additional source of information (see below).

Characterization of alpha satellite diversity in *C. solatus* monomer dataset

A principal component analysis (PCA) using the 5-mer nucleotide composition of DNA sequences was applied to the monomer dataset in order to compare these sequences and identify putative groups without direct alignment. Visualization of sequences into the plane formed by the two first components of the PCA revealed two main groups of alpha satellite monomers, as shown by the distribution of points on Figure 1.1A. Monomers were classified into each group by using a hierarchical clustering analysis (HCA) based on a subset of sequences followed by a linear discriminant analysis (LDA) to extend the classification to all the sequences (see colors

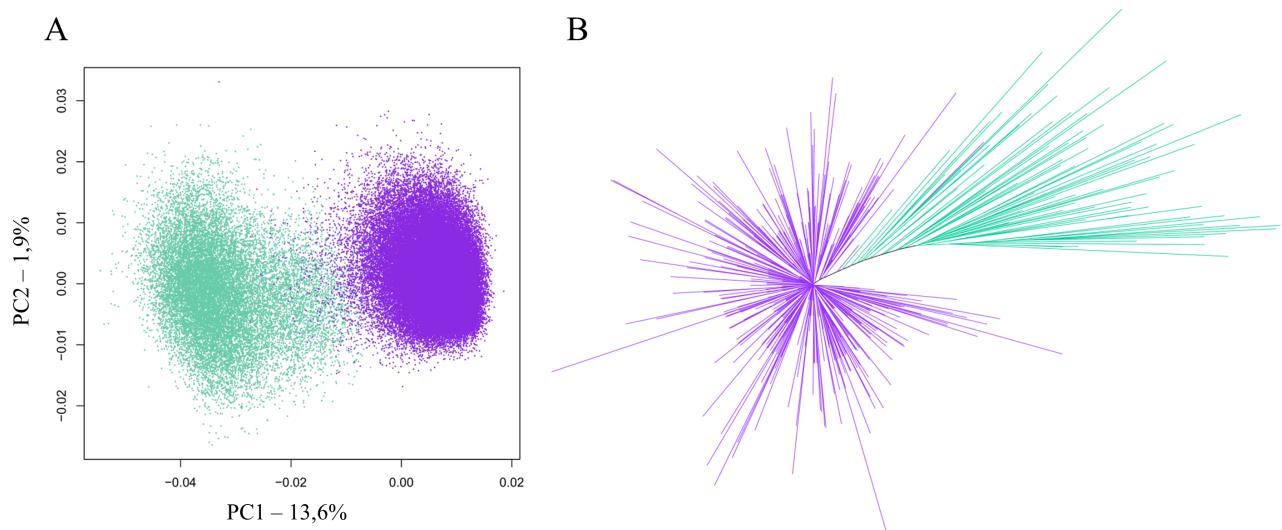


FIGURE 1.1 – Characterization of alpha satellite DNA diversity in *C.solatus* monomer dataset. (A) PCA projection on principal components 1 and 2 of the normalized 5-mer frequency vectors for all sequences. Each point represents a sequence and has been colored according to its assignment to the C1 (purple) or C2 (pastel green) alpha satellite family based on hierarchical classification method. (B) Phylogenetic tree (Neighbor-joining method, K2P model) for 500 randomly selected sequences. The color code matches the one described for (A) and (B). Bootstrap values for the branches leading to C1 and C2 are not indicated as they remain low, which may be due to the important number of sequences used and their relative divergence.

on Figure 1.1A). The most important group, called C1 and shown in purple, contained 82 % of the sequences and the other group, called C2 and shown in pastel green, contained the remaining 18 %. To address the quality of this classification, 500 sequences were then randomly selected within the complete monomer dataset and were used to generate a phylogenetic tree where branches were colored according to the monomer classification (Figure 1.1B). The disposition of sequences from the C1 and C2 groups on this tree provided a further support to our classification into two groups. Moreover, this tree showed a higher degree of divergence between C2 sequences compared to C1 sequences. Actually, the comparison of a subset of 500 randomly selected sequences within each group showed that the average sequence identity inside C1 was 95 %, whereas the average sequence identity inside C2 was only 85 %. The consensus sequences of C1 and C2 were 172 bp in length, and differed from each other by a total of 9 positions (Figure 1.2). Finally, monomers were searched for the presence of CENP-B and pJalpha boxes (Rosandić et al., 2006). A pJalpha box was present in the consensus of C1 and C2 and was found in 95 % of C1 sequences and 85 % of C2 sequences, whereas a CENP-B box was only found in 0.05 % and 0.04 % of these sequences, respectively.

In order to further characterize the sequence diversity within the monomer dataset, we searched for the presence of identical sequences and noticed that while the sequences within the C2 group were all unique, numerous identical sequences could be found within the C1 group. A total of 4,850 sequences were repeated at least twice, representing a total of 20,248 reads in our dataset. Among those sequences, 20 were repeated more than 40 times and one 2,678

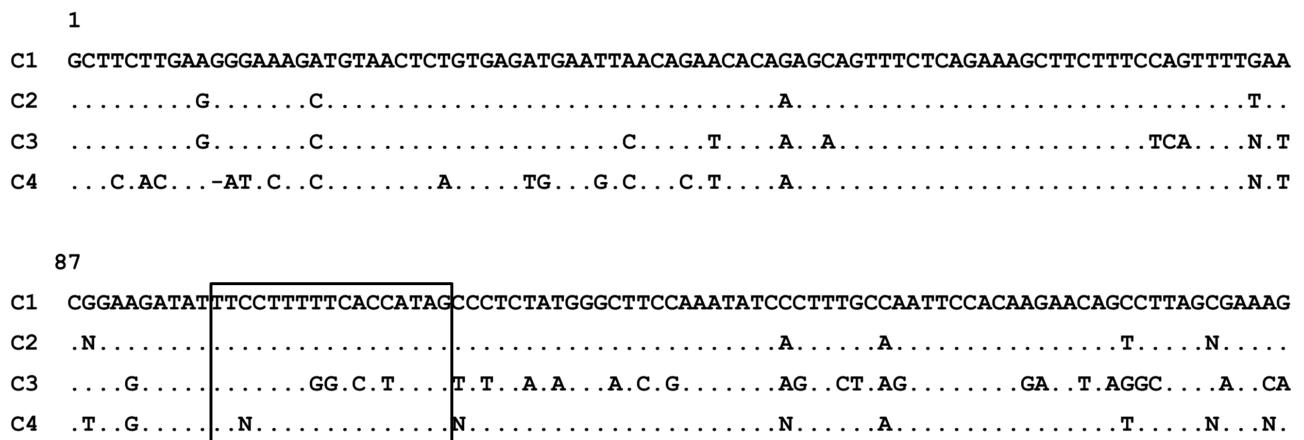


FIGURE 1.2 – Consensus sequences of the alpha satellite families identified in the *C. solatus* genome. The consensus sequences were determined following the alignment of 500 randomly selected sequences for the C1 and C2 families, and the alignment of the available 109 and 112 sequences for the C3 and C4 families respectively. Each position was considered unambiguous if more than 60 % of monomers had the same nucleotide at this position. A point at a position replaces a nucleotide identical to the nucleotide at the homologous position in the C1 consensus. The box shows the fixation site of the pJalpha protein, which is absent from the C3 consensus.

times. We decided to investigate further the 20 most abundant sequences. The most abundant sequence was exactly the consensus sequence of the C1 group, while all the others corresponded to this sequence with single nucleotide variations and/or deletions, as indicated on Table 1.1. The absence of repeated sequences in the C2 group let us hypothesize, by contrast, that the different repeats observed in the C1 group may directly reflect the presence of strictly identical sequences in the *C. solatus* genome. As Ion Torrent sequencing has been reported to give rise to sequencing errors, we decided to search if the identical sequences were obtained from reads collected in both orientations. We found that five out of the 20 sequences were associated with a strong bias for read orientation (Table 1.1). Within these five sequences, the three more abundant (2, 3, 8) represented deletions within a homopolymer tract, while the two others (15, 20) corresponded to the combination of the two most abundant deletions (found in 2 and 3) with the most abundant single nucleotide variation (found in 4). Deletions within homopolymer tracts have already been shown to be inherent to the Ion torrent Technology (Bragg et al., 2013) and the orientation bias we observed let us conclude that sequences displaying these deletions were non-relevant artifacts. On the contrary, all other sequences observed in high copy number, which were all obtained in both sequencing orientations, would correspond to sequence variants that are present with a high abundance in the *C. solatus* genome.

TABLE 1.1 – Analysis of alpha satellite sequences found in high copy number in *C. solatus* monomer dataset

Id	Sequence	Number	Forward (%)
1	Consensus C1	2678	46
2	C114Del	486	1*
3	T101Del	357	99*
4	T39G	242	46
5	A40C	101	56
6	T121A	92	41
7	T74G	78	47
8	T80Del	78	100*
9	G84C	78	41
10	C42G	76	53
11	G1A	74	43
12	A110G	65	48
13	A112T	61	38
14	T19C	59	37
15	T39G-C114Del	57	0*
16	A151C	56	52
17	G79C	55	53
18	C89T	53	49
19	G1T	46	63
20	T39G-T101Del	41	98*

NOTE - The sequences are ordered and numbered according to the number of identical copies of the sequence in the monomer dataset. The “Sequence” column indicates how each sequence differs from the consensus sequence of the C1 family, using standard notations. The “Number” column displays the number of identical copies of the sequence in the monomer dataset. The “Forward” column displays the percentage of reads obtained in the forward orientation (i.e. the orientation of our reference sequence). Strong biases for read orientation are shown with an asterix (*).

Characterization of alpha satellite diversity in *C. solatus* dimer dataset

Among the 3,568 sequences recovered from the dimer dataset, 1,095 contained an intact XmnI restriction site approximately located in the middle, suggesting that the enzymatic digestion was not complete. The remaining 2,473 sequences, which did not possess the XmnI restriction site, were split using an alignment-based process (see Methods) and the resulting monomers were submitted to a length filter, giving a total of 2,408 associated left and right monomers. We focused first on these sequences and submitted them to the same process as described above. PCA showed the existence of two groups for both the left and right monomers, which could be discriminated using HCA (Figure 1.3A and B). A comparison of the consensus sequences of each group revealed that the most abundant sequence set for both the left and right monomers belonged to the C2 group, while the two smaller sequence sets had consensus sequences that differed from each other and from the consensus of the C1 and C2 groups (Figure 1.2). These two new groups of sequences, which represented 5 % of the left or right monomers, will be from now on called C3 and C4 and shown in dark and light pink, respectively. We decided to build a phylogenetic tree with left and right monomers mixed together, using the described color code (Figure 1.3C). This tree confirmed the existence of C3 and C4 as separate groups. Their respective average sequence identities were measured to be 86 % and 83 %. We also checked that left and right monomers belonging to the C2 group could not be distinguished from each other on a phylogenetic tree (Supplementary figure 1.2), or from the C2 sequences present in the monomer dataset. The comparison of the consensus sequences showed that the C3 and C4 groups differed much more from each other and from the C1 and C2 groups than C1 and C2 differed from each other. Interestingly, C4 was the only group with a consensus length of 171 bp instead of 172 bp. A search for CENP-B and pJalpha boxes showed that most sequences within the C4 group contained a pJalpha box (75 %) while the CENP-B box was absent, like observed for the C1 and C2 groups. By contrast, neither the pJalpha box nor the CENP-B box was found in the sequences from the C3 group (Figure 1.2).

The dimer dataset was also used to infer information regarding how monomers belonging to different groups associated with each other. All left and right monomers were assigned to one of the C1 to C4 groups (see Methods). Supplementary table 1.2 reports the results of these assignments as well as associations between left and right monomers, distinguishing dimers that contained the XmnI site (X dataset) and those where the XmnI site was absent (noX dataset). We noticed that sequences from the C1 group were absent from the noX dataset and were poorly represented in the X dataset. This result may appear unexpected as 82 % of the sequences from the monomer dataset belonged to the C1 group. Two hypotheses may explain this observation : the high sequence identity within the C1 group may reduce both the likelihood of the inactivation of the XmnI digestion site through mutations and the sequencing efficiency of dimers (see above). A statistical analysis of the X dataset showed that left monomers from the C1 and C2 groups were preferentially associated to right monomers from the same group (Supplementary table 1.2), which suggests that sequences from the C1 and the C2 groups are tandemly repeated in the *C. solatus* genome. C2-C2 associations were also found to predominate

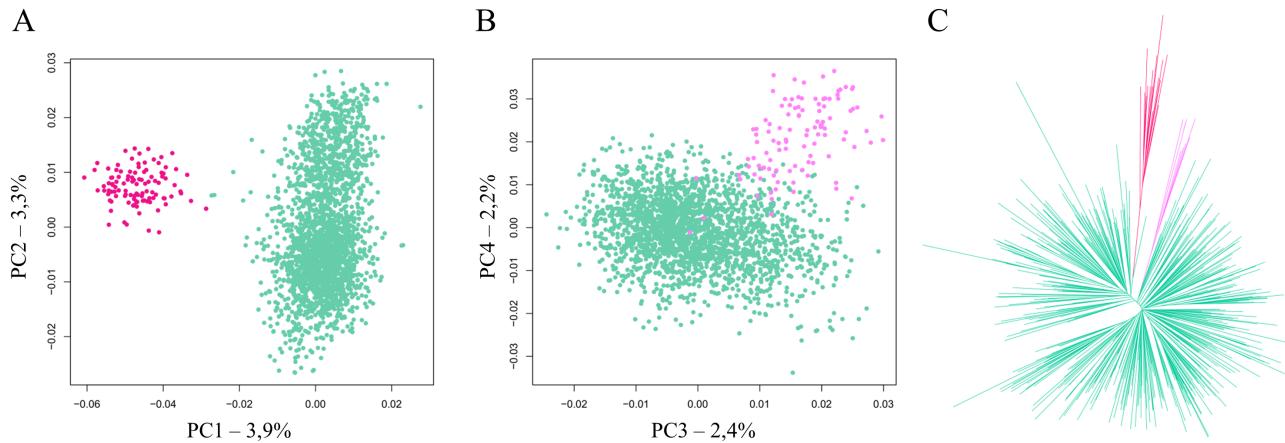


FIGURE 1.3 – Characterization of alpha satellite diversity in *C. solatus* dimer dataset. PCA projection of the normalized 5-mer frequency vectors are shown for (A) the left monomers and (B) the right monomers. Each point represents a sequence and has been colored according to its assignment to the C2 (pastel green), C3 (dark pink) or C4 (light pink) alpha satellite family based on hierarchical classification method. (C) Phylogenetic tree (Neighbor-joining method, K2P model, 100 bootstraps) for 250 randomly selected left monomers and 250 randomly selected right monomers. The color code matches the one described for (A) and (B). Bootstrap values for the branches leading to C3 and C4 are 87 and 55, respectively.

within the noX dataset. Interestingly, left monomers from the C3 group were preferentially associated to right monomers from the C4 group, suggesting the existence of a higher order organization with repeats containing at least two monomers belonging to different groups.

Distribution of alpha satellite families on *C. solatus* chromosomes

We were next interested in studying the genomic distribution of the four groups of sequences identified above. Short oligonucleotide probes have been shown to be more efficient at distinguishing alpha satellite sequences that differ by very few nucleotides compared with classical probes obtained by random priming or nick translation (O'Keefe and Matera, 2000; Silahtaroglu et al., 2004). We chose to use synthetic 18-mer oligonucleotides carrying locked nucleic acid (LNA) modifications at one out of two positions and capable of forming at least 7 GC base pairs, as previous work had demonstrated their interest for the detection of alpha satellite sequences (Ollion et al., 2015). An in silico probe selection process was implemented in order to identify among the most common 18-mer sequences within a group (found in more than 20 % of the monomers) those that were specific for this group (found in less than 3 % of the monomers of other groups). As we expected that oligonucleotide probes may still hybridize in the presence of one mismatch, we calculated the expected binding frequencies when one mismatch was present and applied the same selection criteria once again. Supplementary figure 1.3 reports the sequences that best fitted with our requirements, albeit not completely. Due to the high sequence similarity between sequences within the C1 and C2 groups, probes had to distinguish

sequences that differ mainly by only two nucleotides or even a single one (Supplementary figure 1.3). The two sets of probes selected to target the C1 and C2 groups were therefore designed so that they would compete with each other if used simultaneously. The detection systems (fluorophores or haptens) were chosen in order to allow various combinations of probes to be tested together.

A first series of FISH experiments on *C. solatus* metaphase spreads was performed using probes C1a and C2a or C1b and C2b simultaneously. Probes targeting the C1 group produced intense signals at the centromere (primary constriction) of all chromosomes except a single one (Figure 1.4A and C), while probes targeting the C2 group provided signals that are located at the pericentromeres (around the primary constriction) of several chromosomes pairs with different labeling patterns (Figure 1.4B and C and Supplementary figure 1.4A). Some chromosomes were extensively labeled by C2 probes on both sides of the centromere, others seem to be labeled at only one side, and others seemed to display no signal (see arrows on Figure 1.4D). Stronger signals were observed on the acrocentric chromosome short arms (Supplementary figure 1.4A). When these probes were used alone, each C1a or C2a probe produced a labeling pattern similar to what was observed in the presence of the other. On the contrary, each probe C1b or C2b used alone labeled regions that are larger than in the absence of the other. These experiments suggest that our probes may hybridize to sequences that differ by a single nucleotide (i.e. C1b binds to sequences from the C2 group and C2b probe binds to sequences from the C1 group) but that this binding is inhibited in the presence of an adequate competitor probe. In addition, when target sequences differ by at least two nucleotides, a specific detection is achieved in the absence of competitor.

Additional experiments showed that in presence of competitors, the signal produced by C1a overlapped with the signal produced by C1b and the signal produced by C2a almost perfectly overlapped with the one produced by C2b (Supplementary figure 1.5). This observation supports the idea that the labeling patterns observed with the chosen oligonucleotide probes reflect the distribution of the sequence groups identified by sequence analysis. Moreover, the absence of overlap between signals provided by probes targeting sequences from the C1 and C2 groups suggests that monomers within each group are clustered together and do not mix with each others. Combined with the arguments described above that are in favor of a tandem organization of monomers for both the C1 and C2 groups, these features support the fact that the C1 and C2 groups of sequences represent distinct families of alpha satellite DNA that display a monomeric organisation in the genome of *C. solatus*.

Further experiments were performed with probes targeting the C3 and C4 groups. All C3 and C4 probes provided identical labeling patterns, with a strong signal located on a single chromosome, as well as very weak pericentromeric signals on some other chromosomes (Figure 1.5 and Supplementary figure 1.5). The chromosome labeled by the C3 and C4 probes, which is in fact the chromosome that was not labeled by the probes targeting C1, was also identified to be the Y chromosome by cytogenetic experiments (Supplementary figure 1.4B). The colocalization

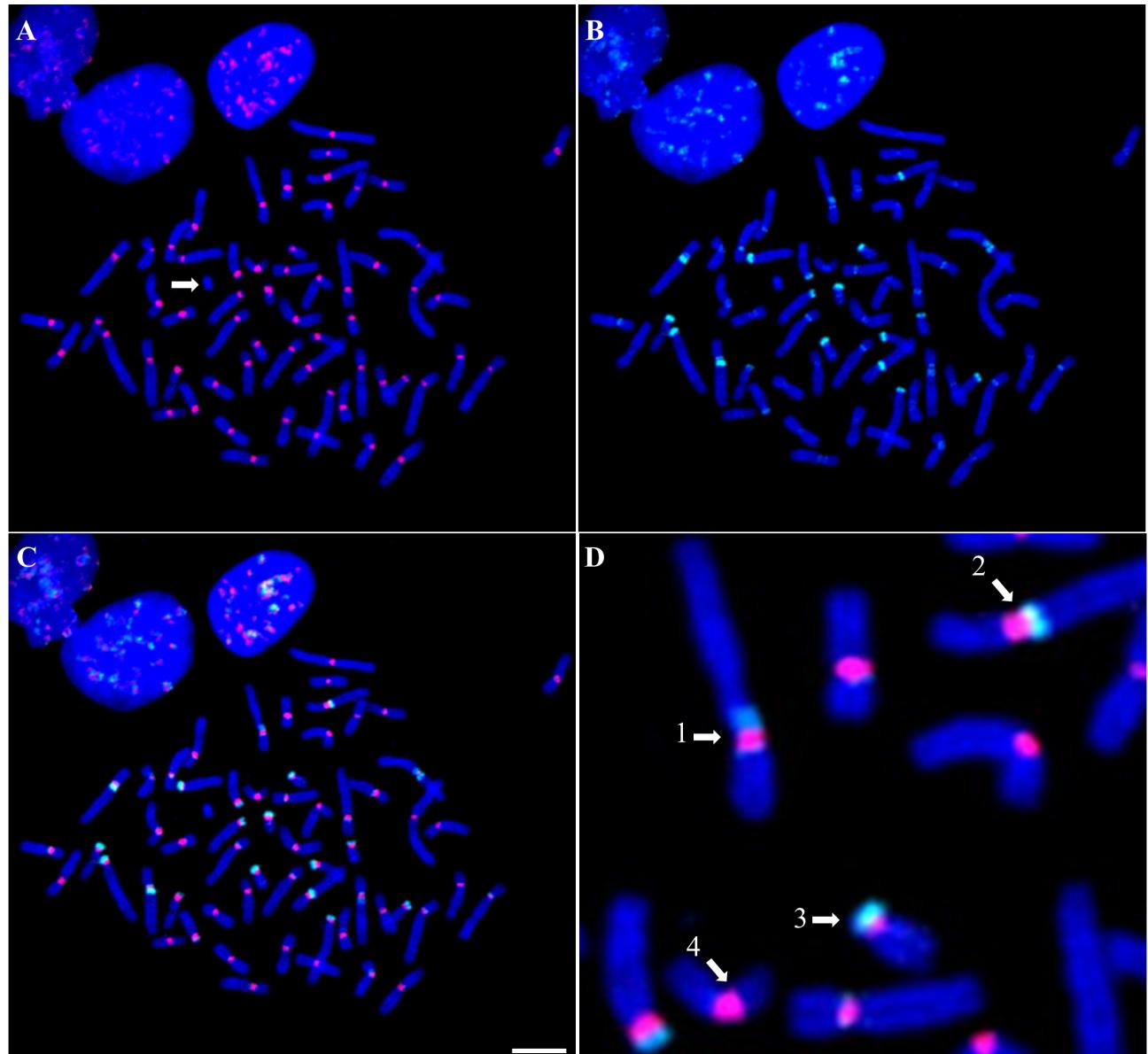


FIGURE 1.4 – FISH analysis of the C1 and C2 alpha satellite families on *C. solatus* chromosomes. Probes C1b and C2b are hybridized simultaneously to *C. solatus* metaphase chromosomes, which are colored in blue. (A) Hybridization of probe C1b is shown in red. The arrow points to a single unlabeled chromosome. (B) Hybridization of probe C2b is shown in green. (C) Combined signals from (A) and (B). (D) Focus on image (C) showing in details the different types of distribution of the C2b signals relatively to C1b. 1 : C2b labels both pericentromeres, 2 : C2b labels one pericentromere toward the long arm, 3 : C2b labels one pericentromere toward the short arm of an acrocentric chromosome, 4 : no C2b signal can be observed on this chromosome. Scale bar = 10 µm.

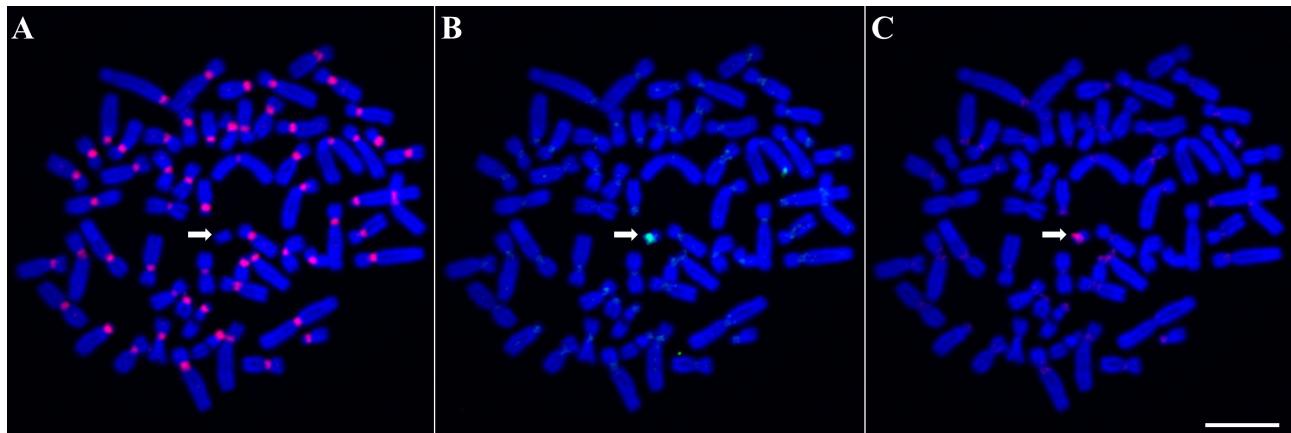


FIGURE 1.5 – FISH analysis of the C3 and C4 alpha satellite families on *C. solatus* chromosomes. Probes C1a, C3a and C4a are hybridized simultaneously to *C. solatus* metaphase chromosomes, which are colored in blue. (A) Hybridization of probe C1a is shown in red. A single chromosome (shown with an arrow) is not labeled. (B) Hybridization of probe C3a is shown in green. (C) Hybridization of probe C4a is shown in red. The pericentromeres of several chromosomes are sparingly labeled by C3a and C4a. Scale bar = 10 µm.

of probes targeting the C3 and C4 groups and the absence of overlap with probes targeting the C1 or C2 group is consistent with the sequence analysis described above. These results taken together suggest thus that sequences belonging to the C3 and C4 group represent additional families of alpha satellite DNA that display a higher order organization within the genome of *C. solatus*. As a further control of the consistency between the results from FISH experiments and sequence analysis, we showed that a 13-mer LNA probe that was designed to target the four C1 to C4 groups of sequences (called Cx) provided signals that overlapped with the combined signals of probes targeting each group, i.e. was able to label all chromosomes at centromeres and pericentromeres (Supplementary figure 1.6).

We were also interested in studying the chromosomal distribution of some of the repeated sequences found in high copy number in the monomer dataset. The results of our previously described FISH experiments suggest that the specific detection of single nucleotide variations may be difficult to achieve using individual probes but that using several probes in competition may provide the possibility to achieve the required level of specificity. Therefore, we designed new oligonucleotide probes targeting a common region, aiming at distinguishing three different highly repeated sequences with single nucleotide variations (Figure 1.6, see Methods). When all probes were used in combination, probes targeting sequence 4 (T39G variation) and sequence 10 (C42G variation) seemed to label all chromosomes, albeit with non-overlapping patterns (see for example insets in Figure 1.6F), while probe targeting sequence 5 (A40C variation) was clearly shown to produce a signal on only 8 chromosomes.

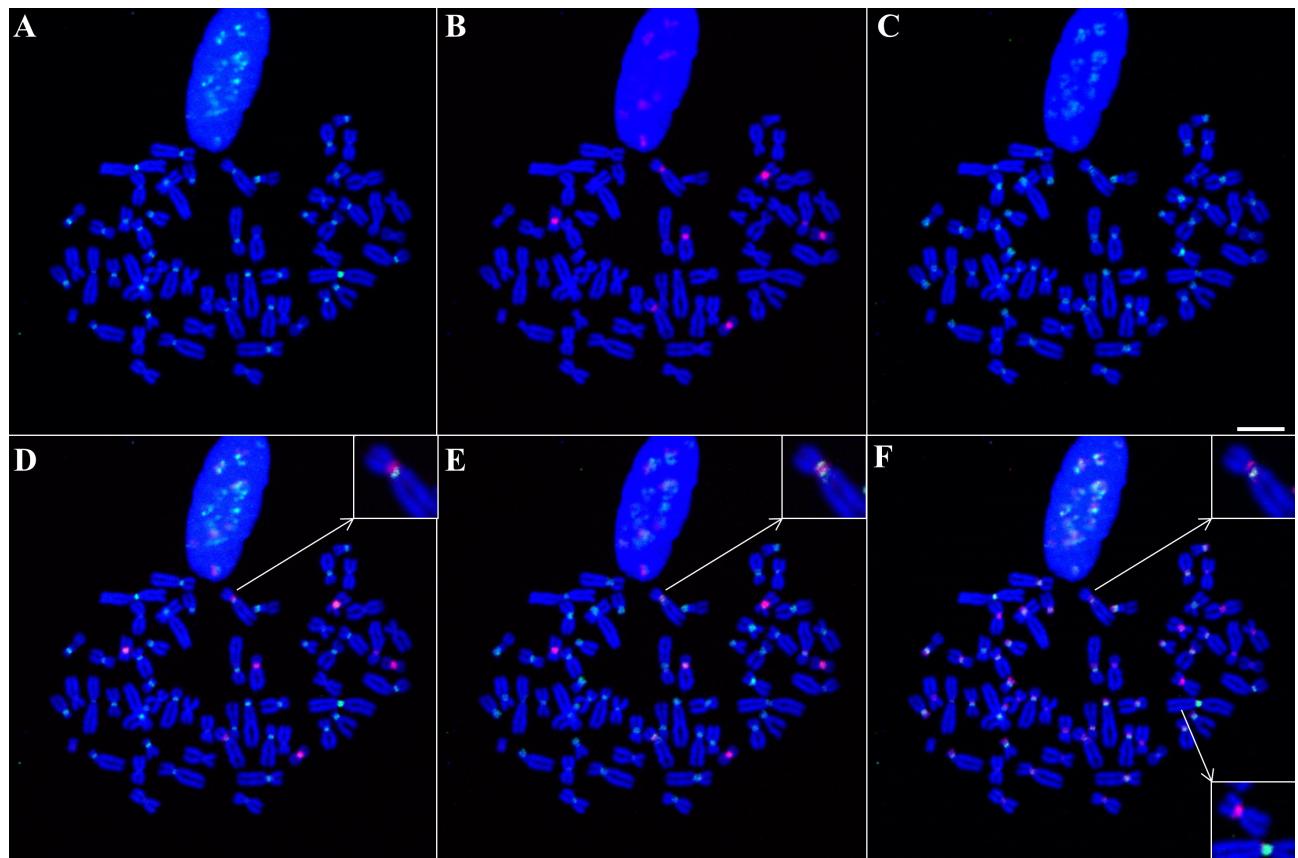


FIGURE 1.6 – FISH analysis of the high copy numbered alpha satellite sequences 4 (T39G), 5 (A40C) and 10 (C42G) on *C. solatus* chromosomes. Probes T39G, A40C and C42G and the competitor oligonucleotide TACco (complementary to the C1 consensus) are hybridized simultaneously to *C. solatus* metaphase chromosomes, which are colored in blue. (A) Hybridization of probe T39G is shown in green. (B) Hybridization of probe A40C is shown in red. 8 chromosomes are labeled. (C) Hybridization of probe C42G is shown in green. (D) Combined signals from (A) and (B). (E) Combined signals from (B) and (C). (F) Combined signals from (A) and (C) with probe T39G shown in green and probe C42G shown in red. Upper inset in (D), (E) and (F) shows one chromosome where signals from the 3 probes do not overlap. Lower inset in (F) shows two chromosomes, one being labeled by probe T39G and the other by probe A40C. Scale bar = 10 µm.

Comparison of *C. solatus* alpha satellite families with known primate families

The sequence families defined above may provide information regarding the evolutionary history of alpha satellite DNA in Primates. We were therefore interested in investigating phylogenetic relationships between these families and alpha satellite sequences that were previously described for other primate species. Interestingly, the first alpha satellite consensus sequence ever described, which was obtained for the cercopithecini *C. aethiops* (Rosenberg et al., 1978), was exactly the same as the consensus sequence of our C1 family, which is also the most abundant repeated sequence in our monomer dataset. This identity suggests the conservation of the C1 family between Cercopithecini species. Although very few sequences were available, a tentative classification was previously proposed for alpha satellite DNA present in Old and New World monkeys, involving five families termed S1 to S5 (Alexandrov et al., 2001). We built a phylogenetic tree containing 50 sequences randomly selected within each of our C1 to C4 families and several sequences representative for S1, S2, S4 and S5 (Figure 1.7A, see Methods). The S1 sequences obtained from *C. aethiops* were intermingled in this tree with our C1 and C2 sequences. Other sequences classified in S1 but obtained from other species were dispersed in other parts of the graph, suggesting that the proposed S1 family was not relevant. There was also no clear proximity of each one of the C1 to C4 family with sequences belonging to the so-called S2, S4 or S5 family. The phylogenetic tree showed on the contrary that sequences from macaque (identified as S1 or S2) may form a sister group of the C4 family whereas the only available baboon sequence (identified as S1) was close to the C3 family. None of the sequences from macaque or baboon resembled those from our C1 or C2 family. All these results suggest that, contrary to S4 and S5, S1 and S2 do not correspond to alpha satellite families. We also built a phylogenetic tree involving our C1-C4 families and seven families (termed M1, R1-2, V1, and H1 to H4) that were previously identified in human pericentromeres, some of them being reported as similar to sequences found in other primates (Shepelev et al., 2009) (Figure 1.7B). The tree suggests that alpha satellite families found in *C. solatus* have an evolutionary history that is largely independent from that of the alpha satellite families found in human pericentromeres.

2.1.4 Discussion

Despite the recent generalization of high-throughput sequencing, application of these new technologies to the study of repeated DNA remains scarce (Rojo et al., 2015; Ruiz-Ruano et al., 2016). Here, we present an original experimental and computational framework for studying repeated DNA. We have focused on a single Cercopithecini species where the diversity and organization of alpha satellite DNA are described in details. Our approach relies on sequencing of gel purified alpha satellite monomers and dimers obtained by restriction enzyme digestion of genomic DNA, followed by sequence analysis and FISH experiments with carefully designed probes.

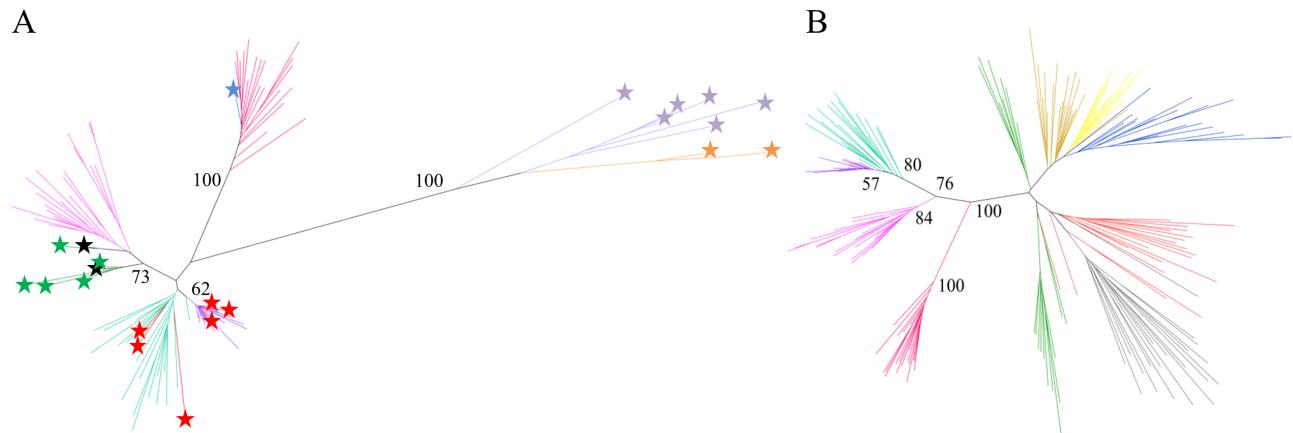


FIGURE 1.7 – Phylogenetic relationships between *C. solatus* alpha satellite families and other previously identified primate families. (A) Phylogenetic tree (Neighbor-joining method, K2P model, 100 bootstraps) for 20 randomly selected sequences within each C1 to C4 family and 22 monomers (labeled with stars) used to propose the S1-S5 families in Old and New World monkeys ((Alexandrov et al., 2001))(see Methods) : one S1 monomer from baboon (blue), two S1 monomers from macaque (black), five S2 monomers from macaque (green), six S1 monomers from *C. aethiops* (red), six S4 monomers (mauve) and two S5 monomers (orange) from New World monkeys. C1 is shown in purple, C2 in pastel green, C3 in dark pink and C4 in light pink. Bootstrap values are given for principal branches when superior to 50. (B) Phylogenetic tree (Neighbor-joining method, K2P model, 100 bootstraps) for 20 randomly selected sequences within each C1 to C4 family and within pericentromeric human families : R1-2 (blue), M1 (yellow), V1 (gold), H1-2 (green), H3 (red) and H4 (grey). C1 is shown in purple, C2 in pastel green, C3 in dark pink and C4 in light pink. Bootstrap values are given for *C. solatus* principal branches.

We detected four alpha satellite families, called C1 to C4, in the *C. solatus* genome. Additional families may have been missed by our approach, for example because they would not contain restriction sites for XmnI. Although some technical issue had drastically reduced the number of available amount of sequences containing two monomers, the dimer dataset provided information about the structural organization of each family, showing that the C1 and C2 families adopt a monomeric organization, while C3 and C4 would associate into HORs. Our data do not allow concluding if the C3-C4 dimers are tandemly repeated or represent only a part of a longer HOR involving other monomers, but suggest that such structures, which have also been observed in New World monkey genomes (Sujiwattanarat et al., 2015), may be widespread in Primates. It had previously been reported, using a limited number of sequences, that alpha satellite sequences in Old World monkeys contained a pJalpha binding site and no CENP-B binding site (Alexandrov et al., 1993; Goldberg et al., 1996; Yoda et al., 1996). Our data provide further support to this observation which holds true for three of the four newly identified families. The absence of any of these two binding sites in the C3 family represents an oddity but one should notice that as sequences from the C3 family are associated with sequences from the C4 family into a HOR organization, the pJalpha binding site remains present in the repeated motif. We detected several sequences in our dataset that were repeated identically a high number of times (up to several thousands). As our protocol does not contain any PCR amplification before capture of individual sequences on beads, the abundance of these sequences may reflect their natural abundance within the *C. solatus* genome, provided one is able to identify potential artifacts resulting from sequencing errors among those sequences.

The high similarity between *C. solatus* alpha satellite families, especially C1 and C2, the consensus of which differ at only a few nucleotide positions, required the implementation of a highly specific FISH detection to infer their chromosomal distribution. Our results emphasize the interest of short LNA-modified oligonucleotide probes that are here shown to be able to distinguish sequences that differ by only two nucleotides. It is even possible to distinguish a single nucleotide variation between two sequences by using two probes targeting each sequence variant simultaneously. In all our experiments, we cannot exclude the possibility that probes also hybridize to sequences that are not perfectly complementary, nor that some signals do not come from sequences that are present in the *C. solatus* genome but not found in our datasets. Nevertheless, the absence of cross labeling between the probes targeting different families and the consistency of hybridization results with predictions inferred from sequence analysis support our probe design strategy and stands for the accuracy and the exhaustiveness of our description of the alpha satellite component of *C. solatus*.

Our FISH experiments showed that the C1 family, which is the most conserved (95 % mean sequence identity), displays a centromeric localization while the more divergent C2 family (85 % mean sequence identity) displays a pericentromeric localization. According to the age-gradient based model for centromere evolution (Schueler et al., 2001, 2005), we may speculate that this pattern results from a peculiar evolutionary history where the C2 family, an old family of sequences, had occupied a centromeric position in an ancestor of *C. solatus*. This family

would then have been displaced towards pericentromeres following the amplification of more recent sequences from the C1 family at the centromere. Unequal crossing over at nearly identical repeats is thought indeed to lead to the homogenization of the core centromere, while mutations would only affect repeats outside of the core centromere (Smith, 1976; Schueler et al., 2001; Henikoff, 2002; Shepelev et al., 2009; Henikoff et al., 2015). An alternative but nonexclusive hypothesis would attribute distinct functional roles to both families, for example centromere function to C1 and cohesion of sister chromatids to C2, as it has been proposed for mouse minor and major satellite sequences, respectively (Guenatri et al., 2004). Interestingly, acrocentric chromosome short arms display a very large amount of C2 sequences as revealed by intense FISH signals. This observation provides support to a previous hypothesis according to which acrocentric chromosomes may physically interact and exchange genetic material (Choo et al., 1990; Warburton et al., 2008). The fact that the C3-C4 dimers are found on the Y chromosome and are almost absent from other chromosomes may be explained by the fact that the Y chromosome is excluded from recombination events with non-homologous chromosomes, as was observed in mice (Pertile et al., 2009). Finally, the observation of the distribution of one of the highly repeated sequence variants on only 8 chromosomes supports the existence of local alpha satellite homogenization events in the *C. solatus* genome.

Previous studies had considered alpha satellite DNA in Cercopithecini as poorly diversified (Alexandrov et al., 2001). Our results show that at least four alpha satellite families can be present in a single species, with complex chromosomal distribution and organizational patterns. Comparative studies including repetitive DNAs from different species have already been shown to provide new insights into genome and species evolution (Mravinac and Plohl, 2010). Our approach will permit not only to investigate the taxonomic distribution of alpha satellite families but also to study their organizational pattern, their chromosomal distribution as well as the existence of conserved highly repeated sequence variants. Phylogenetic analysis have demonstrated that the C1 to C4 families represent newly identified entities that do not correspond to previously proposed alpha satellite families. Although the available data are in favor of an apparent conservation of both the C1 and C2 families between *C. solatus* and *C. aethiops*, further studies will be required to better understand the dynamics of alpha satellite DNA in Old World monkeys and in other primates.

2.1.5 Conclusion

In summary, we have presented here a generally applicable strategy that provides, for a single species, a comprehensive description of alpha satellite sequence diversity and organization. Our approach, which is easy to implement and cost-effective, provides an opportunity to characterize satellite DNA in all species where a characteristic enzymatic ladder pattern can be obtained. Comparing different individuals and different species will provide new insights into the dynamics at which new satellite families or new highly repeated sequence variants appear during the course of evolution and transfer between chromosomes. The better description of the structure

of heterochromatic regions also provides potential for enhancing the epigenetic characterization of these regions as well as understanding the regulatory functions of heterochromatin.

2.1.6 Methods

DNA collection and metaphase preparations

Fibroblast samples of *Cercopithecus solatus* (ID : 2012-028, male sample, ethic permission n° FR1207510445-I) from the Collection of cryopreserved living tissues and cells of vertebrates (RBCell collection, Muséum national d'Histoire naturelle, Paris) were used for DNA extraction and metaphase preparations. DNA was extracted using the Omega Bioteck Tissue DNA Kit (Doraville, USA). Cell cultures and metaphase preparations were achieved according to Moulin et al. (2008).

Alpha satellite DNA isolation and sequencing

The Serial Cloner software (Serial Basics, serialbasics.free.fr) was used to perform in silico digestions of the Cercopithecini alpha satellite sequences registered as such in Genbank (Accession numbers : AM235889, AM235890, AM237210, AM237214, AM237213, AM237212, X04339, V00145, M26844 and AM237211), which contained both monomers and dimers. The restriction site of the XmnI restriction enzyme (GAANNNNNTTC) was observed once in a great proportion of monomers and twice in almost all dimers. XmnI was then used to digest *C. solatus* DNA in vitro. 10 µg of *C. solatus* genomic DNA were digested for 4 h 30 min at 37 °C with 60 units of XmnI activity (New England Biolabs) in a total volume of 34 µL. The enzyme was inactivated for 20 min at 65 °C. The sample was loaded on a 1 % agarose gel after addition of 6.8 µL loading buffer (50 % glycerol) and electrophoresis was performed in 0.5X Tris-borate-EDTA buffer, at room temperature for 2 h 45 min at 100 V. The gel was briefly stained with ethidium bromide and then imaged by UV transillumination. Bands corresponding to alpha satellite monomers (\approx 170 bp) and dimers (\approx 340 bp) were cut and DNA was extracted from the gel with the Omega Bioteck Gel extraction kit and resuspended in 100 µL of elution buffer. About 220 ng and 110 ng were obtained for the 170 bp and 340 bp samples, respectively.

Sequencing was performed on a PGM sequencing platform (Ion Torrent technology) using the 400 bp sequencing kit. Two libraries were generated using 50 ng of both blunt digest pools and the Ion Plus Fragment Library Kit (4471252, Life Technologies) and tagged with Ion Xpress barcode adapters (4471250, Life Technologies). After purification (1.8X) with Ampure XP Beads (A63880, Agencourt Bioscience, Beverly, USA), the libraries were quantitated using a Sybr Green qPCR assay (Biorad, Hercules, USA) based on a custom *Escherichia coli* reference library. After a dilution of each library down to 26 pM, 0.22 fmol for the 170 bp library and 0.44

fmol for the 340 bp library were pooled as templates for the clonal amplification on Ion Sphere particles during the emulsion PCR, performed on a One Touch2 emPCR robot according to the Ion PGM Template OT2 400 Kit user guide (4479878, Life Technologies). The amplification products were loaded onto an Ion 316v2 chip (4483324, Life Technologies), and subsequently sequenced according to the Ion PGM Sequencing 400 Kit user guide (4482002, Life Technologies). After standard filtration of the raw reads (polyclonal and low quality removal), the Ion Torrent sequencing yielded 204,990 sequences for the 170 bp pool and 353,683 sequences for the 340 bp pool. They were deposited in the NIH Short Read Archive (SRA accession numbers SRX1595681 and SRX1595679).

Alpha satellite sequence filtering

All sequences with an average Phred score lower than 25, a length outside the range 162-182 bp for monomers and 324-364 bp for dimers, and sequences without the XmnI digested sites at the extremities (5'-NNTTC ... GAANN-3') were not considered for further analysis. Alpha satellite sequences were identified with a BLAST search against a reference alpha satellite sequence of *C. aethiops* (AM23721) ([Altschul et al., 1990](#)). Using default BLAST parameters, all sequences exhibiting a hit longer than 80 bp for monomers and 160 bp for dimers were considered as alpha satellite sequences and conserved for the following analysis. All sequences were then reoriented if necessary in order to match the orientation of the reference alpha satellite sequence. The orientation information was preserved for investigations regarding reading biases.

Processing of dimeric sequences was performed as follows. When an XmnI site was present in the middle of these sequences, it was used for separating both monomers, providing the so-called left and right monomers located on the 5' side and on the 3' side of the sequence, respectively. Dimers that did not contain any XmnI site in the middle were aligned against a synthetic sequence formed by two consecutive copies of the reference sequence using the Needleman-Wunsch algorithm ([Needleman and Wunsch, 1970](#)) to identify the monomer limits and split them into left and right monomers according to the same rule as described above. All pairs with at least one monomer outside the 162-182 bp range were discarded. Pairing information was conserved to study association between left and right monomers.

Alpha satellite sequence characterization

Monomeric sequences were compared using their 5-mer composition in order to identify putative alpha satellite groups without direct alignment. For each set of monomers, the 5-mer frequency table was analyzed using a principal component analysis (PCA) to reduce the space complexity and enable data visualization on the first factorial planes. Sequences were classified into groups by using a hierarchical clustering method (HCA) based on the Ward criterion ([Ward Jr, 1963](#)) applied to the Euclidean distances calculated from the 100 first principal components of the

PCA. Because of the size of the monomer dataset, direct classification of the sequences using HCA was not possible. Instead, HCA was applied on 2,500 randomly selected sequences which were used to train a linear discriminant model. This model has been finally used to classify all the other monomers. The dimer dataset was analyzed in two different ways : 1) monomers extracted from dimers without Xmnl sites were classified by using an HCA based on a PCA, 2) monomers extracted from dimers with a Xmnl site have been classified by using a linear discriminant analysis trained to recognize the C1-C4 groups.

Because of the size of the datasets, the phylogenetic trees, the consensus sequences and the sequence distance analysis were conducted with different subsets of randomly selected sequences, using a homemade python script. The selected sequences were aligned using MUSCLE (Edgar, 2004) and analyzed with SeaView (Gouy et al., 2010). The phylogenetic trees were built by using the Neighbor joining algorithm and the Kimura 2-parameters distance. Reliability of nodes was assessed using 100 bootstrap iterations. The relatively low bootstrap values observed in the trees can be explained by a limited number of family specific sites, i.e., the informative sites, into the alignments. Nevertheless, the same clustering of the families and the same relationship between these families have been observed with all the trees generated with different randomly selected sequences.

CENP-B and pJalpha boxes were searched with the patterns TTCGTTGGAARCGGGA and TTCCTTTYCACCRTAG respectively (Rosandić et al., 2006) by using the program Fuzznuc (Rice et al., 2000) and allowing 2 mismatches. All statistical analyses were conducted with R (R Core Team, 2013). Our R scripts and other programs are available upon request.

The S1-S5 monomers used in Figure 1.7A have been isolated from the sequences described in Alexandrov et al. (2001). All these monomers have been extracted by using the homologous position of the Xmnl digestion site as a starting point (Xmnl phase) in order to be aligned with the monomers of *C. solatus*. Unfortunately, no full length S3 monomer was available in this phase. To obtain the Genbank accession numbers and the alignment of the used S1-S5 monomers, see Alexandrov et al. (2001) and Text S1. Human monomers from old and ancient families M1, R1-2, V1, H1-H4 used in Figure 1.7B have been isolated from the human Xp chromosome sequence (Genbank ID NT_011630) by using the homologous position of the Xmnl digestion site as a starting point. Monomers have been assigned to a family according to their location along the sequence and the annotations provided in Shepelev et al. (2009) (see Text S2 for alignment).

Oligonucleotide probes

Short oligonucleotide probes (18 nucleotides in length) were designed in order to target specifically the different alpha satellite families identified in *C. solatus*, by systematic prediction of binding frequencies based on the sequencing results. In some instances, when the 18-mer se-

quence did not allow forming at least 7 GC bp upon hybridization to the complementary strand, length was increased to 19. Sequences and binding frequencies are available in Supplementary figure 1.3, which also provides details about the positions of locked nucleic acid (LNA) modifications in the probes. These positions were selected based on previous experience in order to achieve a good binding affinity and specificity (Ollion et al., 2015). When possible, we selected probes that were perfectly complementary to more than 20 % of the sequences from the target group and to less than 3 % of the sequences from the other groups. Supplementary figure 1.3 also provides the expected binding frequencies if hybridization is possible despite the presence of one mismatch between the probe and its target. To target three sequences found in high copy number in the monomer dataset, we designed four LNA-modified probes (LNA are written in lower case and classic nucleotides are written in upper case) : probe T39G (5'TgTtCtGtT-CaTtCaTcTc3', 5'AlexaFluor488), probe A40C (5'TgTtCtGtGAaTtCaTcTc3', 3'Digoxygenin), probe C42G (5'TgTtCtCtTAaTtCaTcTc3', 3'Biotin) and probe TACco (5'TgTtCtGtTAaTt-CaTcTc3') which is complementary to the C1 consensus sequence. LNA-modified probes were purchased from Eurogentec (Seraing, Belgium).

FISH experiments

FISH were performed on metaphase chromosome preparations. Hybridization solutions were prepared by diluting the oligonucleotide probes to a final concentration of 0.1 μ M in a hybridization solution consisting of 2X SSC pH 6.3, 50 % deionized formamide, 1X Denhardt solution, 10 % dextran sulfate, and 0.1 % SDS. 20 μ L of the hybridization solution were deposited on each slide and covered with a coverslip. The slides were then heated for 3 min at 70 °C and hybridized for 1 h at 37 °C in a Thermobrite apparatus (Leica Biosystems). Then, each slide was washed twice in 2X SSC at 63 °C. Preparations were then incubated in blocking solution (4 % bovine serum albumin (BSA), 1X PBS, 0.05 % Tween 20) for 30 min at 37 °C to reduce nonspecific binding. Then, depending on the combination of probes, the following antibodies were used for subsequent revelations : Alexa 488-conjugated streptavidin (1 :200 ; Life Technologies, Foster City, USA), Cy5-conjugated streptavidin (1 :200 ; Caltag Laboratories, Burlingame, USA), FITC-conjugated sheep anti-digoxigenin (1 :200 ; Roche, Lewes, UK), and Rhodamine-conjugated sheep anti-digoxigenin (1 :200 ; Roche). All antibodies were diluted in blocking solution containing 1X PBS, 0.05 % Tween 20, and 4 % BSA. Antibody incubation lasted for 30 min at 37 °C. All washings were performed in 2X SSC, 0.05 % Tween 20. Chromosomes were counterstained with DAPI (4',6-diamidino-2-phenylindole) by pipetting 40 μ L of a 5 μ g/mL solution onto the slides, incubating for 5 min and then briefly washing in 1X PBS. Slides were mounted by adding a drop of Vectashield Antifade Mounting Medium (Vector Laboratories, Burlingame, USA) and covering with a coverslip.

Image acquisition and analysis

Metaphases were imaged using an Axio Observer Z1 epifluorescent inverted microscope (Zeiss) coupled to an ORCA R2 cooled CDD camera (Hamamatsu). The Axio Observer Z1 was equipped with a Plan-Apochromat 63× 1.4 NA oil-immersion objective and the following filters set : 49 shift free for DAPI (G365 / FT395 / BP445/50), 38 HE shift free for FITC/Alexa488 (BP470/40 / FT495 / BP525/50), homemade sets for Rhodamine (BP546/10 / FF555 / BP 583/22) and for Cy5 (BP643/20 / FF660 / BP684/24). The light source was LED illumination (wavelengths : 365 nm, 470 nm or 625 nm) except for Rhodamine, for which a metal halide lamp HXP120 was preferred. Immersion oil of refractive index 1.518 at 23 °C was used. Color-combined images were reconstructed using ImageJ ([Abràmoff et al., 2004](#)). At least ten metaphases were visualized for each experiment, which all confirmed the described patterns.

2.1.7 Additional files

Supplementary figure 1.1 : Migration profiles of *C. solatus* genomic DNA digested with XmnI

Supplementary figure 1.2 : Phylogenetic tree for left and right alpha satellite monomers from *C. solatus* dimer dataset

Supplementary figure 1.3 : LNA-modified probes used to target the C1 to C4 alpha satellite families on *C. solatus* chromosomes

Supplementary figure 1.4 : Distribution pattern of the C2 and C3 alpha satellite families on *C. solatus* chromosomes

Supplementary figure 1.5 : Comparison of FISH signals from different probes targeting identical alpha satellite families on *C. solatus* chromosomes

Supplementary figure 1.6 : Comparison of the hybridization pattern of probe Cx with those of probes targeting the C1 to C4 alpha satellite families on *C. solatus* chromosomes

Supplementary table 1.1 : Filtering steps from *C. solatus* raw data to alpha satellite monomer and dimer datasets

Supplementary table 1.2 : Alpha satellite family associations in *C. solatus* dimer dataset

Supplementary text 1.1. Alignment used to generate the phylogenetic tree from Figure 1.7A

Supplementary text 1.2. Alignment used to generate the phylogenetic tree from Figure 1.7B

RESEARCH ARTICLE

Open Access



CrossMark

Diversity and distribution of alpha satellite DNA in the genome of an Old World monkey: *Cercopithecus solatus*

Lauriane Cacheux^{1,2}, Loïc Ponger¹, Michèle Gerbault-Seureau², Florence Anne Richard^{2,3} and Christophe Escudé^{1*}

Abstract

Background: Alpha satellite is the major repeated DNA element of primate centromeres. Evolution of these tandemly repeated sequences has led to the existence of numerous families of monomers exhibiting specific organizational patterns. The limited amount of information available in non-human primates is a restriction to the understanding of the evolutionary dynamics of alpha satellite DNA.

Results: We carried out the targeted high-throughput sequencing of alpha satellite monomers and dimers from the *Cercopithecus solatus* genome, an Old World monkey from the Cercopithecini tribe. Computational approaches were used to infer the existence of sequence families and to study how these families are organized with respect to each other. While previous studies had suggested that alpha satellites in Old World monkeys were poorly diversified, our analysis provides evidence for the existence of at least four distinct families of sequences within the studied species and of higher order organizational patterns. Fluorescence in situ hybridization using oligonucleotide probes that are able to target each family in a specific way showed that the different families had distinct distributions on chromosomes and were not homogeneously distributed between chromosomes.

Conclusions: Our new approach provides an unprecedented and comprehensive view of the diversity and organization of alpha satellites in a species outside the hominoid group. We consider these data with respect to previously known alpha satellite families and to potential mechanisms for satellite DNA evolution. Applying this approach to other species will open new perspectives regarding the integration of satellite DNA into comparative genomic and cytogenetic studies.

Keywords: Alpha satellite DNA, High-throughput sequencing, *Cercopithecus solatus*, Centromere genomics

Background

Centromeres are chromosomal regions that control chromosome segregation during cell division in eukaryotes, through kinetochore assembly and microtubule attachment. In almost all eukaryotes, the DNA underlying centromeres is made of large tracts of nearly identical tandem DNA repeats, known as satellite DNA [1–3]. The remarkable variation of satellite DNAs between species has been an enigma ever since their discovery and different important roles have been ascribed to these

sequences, from the imperative centromeric function in mitosis and meiosis to regulatory functions [4, 5].

Alpha satellite DNA is the most abundant satellite DNA in Primates and is found both at the site of centromere attachment and in neighboring heterochromatic regions, referred to as pericentromeric regions [6]. Alpha satellite DNA was originally isolated as a highly repetitive component of the *Chlorocebus aethiops* (also called African green monkey) genome [7]; homologous repeats were then described throughout the Primate order including apes, Old World and New World monkeys [8–10]. Alpha satellite DNA is made of tandemly repeated AT-rich monomers that are about 170 bp in length and organized in head-to-tail orientation [11, 12]. In the human genome, individual monomers share between 60 and 100% sequence identity. The highly identical composition of successive repeats represents a

* Correspondence: escude@mnhn.fr

¹Département Régulations, Développement et Diversité Moléculaire, Structure et Instabilité des Génomes, INSERM U1154, CNRS UMR7196, Sorbonne Universités, Muséum national d'Histoire naturelle, Paris, France
Full list of author information is available at the end of the article

technical challenge that has thwarted the complete assembly of centromeric DNA so far [13, 14]. Nevertheless, over the last 30 years, the systematic cloning and sequencing of many alpha satellite DNAs, combined with fluorescence in situ hybridization (FISH) experiments, has provided a thorough knowledge of alpha satellite DNA diversity and organization patterns in the human genome [11, 15, 16] and, to a much lesser extent, in other primates [17–20].

In human, alpha satellite DNA has been shown to adopt two different organizations. In the so-called higher order repeat (HOR) organizational pattern, highly conserved repeat units (97–100% sequence identity), each made of multiple 171 bp monomers (up to more than 30), are found as an homogenized array that can extend over a multimegabase-sized region [2, 13, 21–23]. This organization is typically found as very long arrays of alpha satellites at the centromere core of all human chromosomes. In pericentromeres, a second type of organization, called monomeric and involving arrays of single alpha satellite monomers which are less well conserved (70–90% sequence identity), can coexist with HORs [3, 12]. Sequence comparisons between human alpha satellite monomers have led to the description of up to seventeen different alpha satellite families, or monomer types [19, 21, 24, 25]. Although the alpha satellite component of other primate genomes has been less intensively studied, there is some evidence for similar organizations in great apes, but additional families have been described and the composition of HORs as well as their chromosomal distribution differ when compared with human [12, 20, 26–28]. This implies that the structure and content of centromeric DNA can change in a few million years.

Although the mechanisms that gave rise to this diversity and organization are not precisely known, it is commonly accepted that the so-called concerted evolution of repetitive sequences is based on different mechanisms of non-reciprocal transfer occurring within or between chromosomes, such as unequal crossover, gene conversion, rolling circle replication and reinsertion, and transposon-mediated exchange [4, 29]. Such mechanisms enable series of amplification events, thereby creating new arrays of alpha satellites [12, 16, 30–32]. The analysis of the different alpha satellite families found in assembled pericentromeric regions from specific human chromosomes revealed an age gradient of the families along each chromosome arm, which led to propose that during the course of evolution, new arrays of alpha satellites expand at the centromere core, thereby splitting and displacing older arrays distally onto each arm [3, 6, 13, 19, 33].

Knowledge about alpha satellite DNA in species outside the hominoid group is very scarce, in particular in Old World monkeys, a clade that includes Colobinae, Papionini and Cercopithecini. The tribe Cercopithecini contains 35 species which have diversified within the last 10 million years [34, 35] and therefore represents a particularly interesting group for studying the evolution of satellite DNA.

Moreover, it has been reported that alpha satellite DNA is more abundant in some Cercopithecini species (up to 20% of the genome of *Chlorocebus aethiops*) [36] than in great apes, where its contribution would reach only 3% of the genome [14]. Finally, enzymatic digestion of genomic DNA from various Old World monkey species can lead to a clear alpha satellite ladder pattern which is not observed when human or chimpanzee DNA is used, thereby pointing to different composition and organization of alpha satellite DNA in Old World monkeys [37].

In the present work, we have undertaken the targeted sequencing of the alpha satellite component of *Cercopithecus solatus* (or Sun-tailed monkey) as a representative species for the Cercopithecini [38]. Alpha satellite monomers and dimers were obtained by enzymatic digestion of genomic DNA and gel purification, then submitted to high-throughput sequencing. The obtained sequences were analyzed and classified into monomer families using computational approaches. Finally the genomic distribution of each family was studied by FISH using a collection of oligonucleotide probes that are able to distinguish different sequence variants. Our study provides evidence for the existence of two main families of monomers which differ in their chromosomal distribution, one being specifically distributed on centromeres while the other is found only at pericentromeric locations with a non-uniform distribution between chromosomes. Two other families are detected which are only found associated within a dimeric organization and are located for the greatest part on the Y chromosome and to a lesser extent on pericentromeres from other chromosomes. These data represent the most complete analysis of the diversity and distribution of alpha satellite sequences in an Old World monkey reported to date. Our experimental approach may be applied to other species, opening new perspectives regarding the integration of satellite DNA into comparative studies.

Results

Retrieval of alpha satellite sequences from the *Cercopithecus solatus* genome

Work conducted in the early 1980s had shown that enzymatic digestion of genomic DNA from Old World monkeys with several restriction enzymes resulted in a migration profile that was characteristic for alpha satellite DNA, i.e. with bands corresponding to one and multiple repeat units of about $n \times 170$ bp in length [8, 39]. In silico analysis of several sequences isolated from *Chlorocebus aethiops* led us to select the XmnI restriction endonuclease as a candidate that should cleave a majority of monomers. Experimental digestion of *Cercopithecus solatus* genomic DNA with this enzyme revealed the expected banding pattern (Additional file 1: Figure S1). We therefore decided to extract DNA from two bands corresponding to monomers and dimers of alpha satellites from an agarose gel and

implemented high throughput sequencing on an Ion Torrent sequencing platform providing reads up to 400 nucleotide in length (see Methods).

204,990 and 353,683 raw sequences were obtained for the monomer and dimer samples, respectively. Four in silico filters were applied successively to both datasets: a quality filter keeping sequences with a Phred quality score superior to 25; an extremity filter keeping sequences with the XmnI restriction site at both ends; a length filter keeping sequences within the range 162–182 bp for monomers and 324–364 bp for dimers, and an alpha satellite filter keeping sequences similar to an alpha satellite reference sequence (see Methods). The number of sequences that remained after each filter is reported on Additional file 2: Table S1. A total of 100,713 sequences fitting with all the criteria was obtained from the monomer sample and represents what we call from now on the monomer dataset. For the dimer sample, only 3,568 were obtained, they represent the dimer dataset. The drastic reduction observed within the dimer dataset was mostly the consequence of the length filter and may reflect an intrinsic limitation of the sequencing technology, unable to obtain long reads when template sequences are made of two successive highly identical sequences. These sequences were nevertheless included for further analysis as they provided an additional source of information (see below).

Characterization of alpha satellite diversity in the monomer dataset

A principal component analysis (PCA) using the 5-mer nucleotide composition of DNA sequences was applied to

the monomer dataset in order to compare these sequences and identify putative groups without direct alignment. Visualization of sequences into the plane formed by the two first components of the PCA revealed two main groups of alpha satellite monomers, as shown by the distribution of points on Fig. 1a. Monomers were classified into each group by using a hierarchical clustering analysis (HCA) based on a subset of sequences followed by a linear discriminant analysis (LDA) to extend the classification to all the sequences (see colors on Fig. 1a). The most important group, called C1 and shown in purple, contained 82% of the sequences and the other group, called C2 and shown in pastel green, contained the remaining 18%. To address the quality of this classification, 500 sequences were then randomly selected within the complete monomer dataset and were used to generate a phylogenetic tree where branches were colored according to the monomer classification (Fig. 1b). The disposition of sequences from the C1 and C2 groups on this tree provided a further support to our classification into two groups. Moreover, this tree showed a higher degree of divergence between C2 sequences compared to C1 sequences. Actually, the comparison of a subset of 500 randomly selected sequences within each group showed that the average sequence identity inside C1 was 95%, whereas the average sequence identity inside C2 was only 85%. The consensus sequences of C1 and C2 were 172 bp in length, and differed from each other by a total of 9 positions (Fig. 2). Finally, monomers were searched for the presence of CENP-B and pJalpha boxes [40]. A pJalpha box was present in the

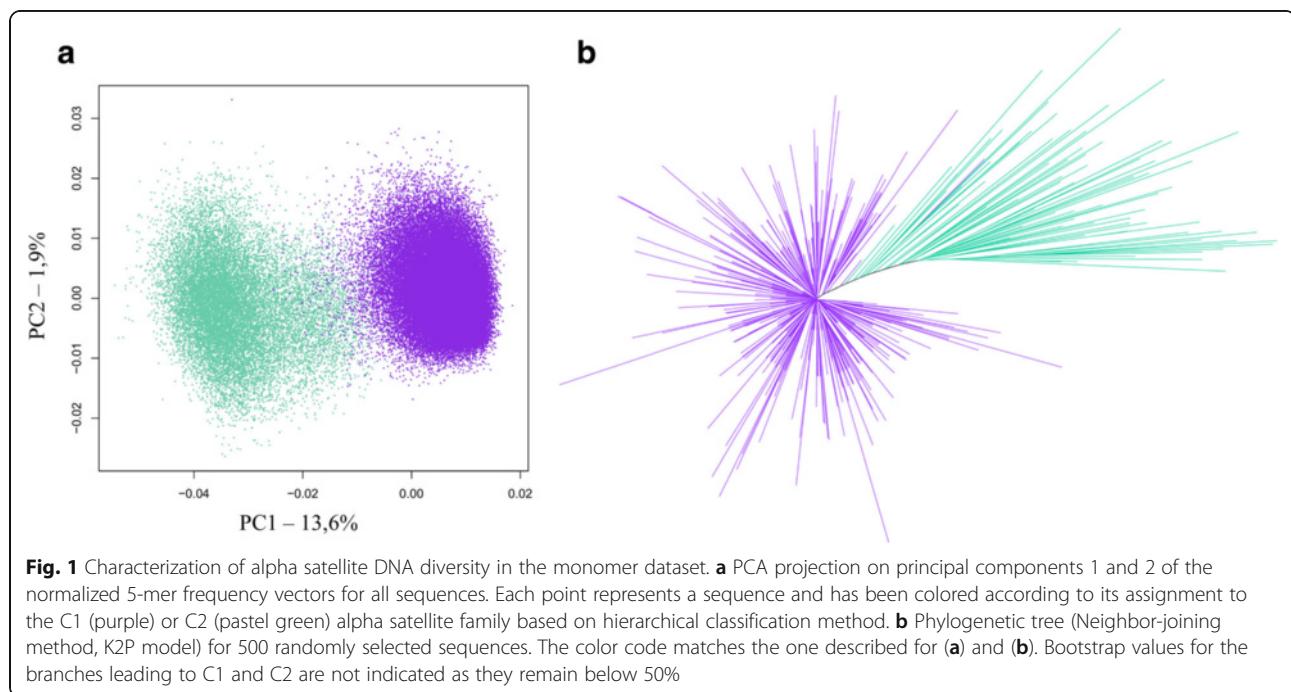




Fig. 2 Consensus sequences of the alpha satellite families identified in the *Cercopithecus solatus* genome. The consensus sequences were determined following the alignment of 500 randomly selected sequences for the C1 and C2 families, and the alignment of the available 109 and 112 sequences for the C3 and C4 families respectively. Each position was considered unambiguous if more than 60% of monomers had the same nucleotide at this position. A point at a position replaces a nucleotide identical to the nucleotide at the homologous position in the C1 consensus. The box shows the fixation site of the pJalpha protein, which is absent from the C3 consensus

consensus of C1 and C2 and was found in 95% of C1 sequences and 85% of C2 sequences, whereas a CENP-B box was only found in 0.05% and 0.04% of these sequences, respectively.

In order to further characterize the sequence diversity within the monomer dataset, we searched for the presence of identical sequences and noticed that while the sequences within the C2 group were all unique, numerous identical sequences could be found within the C1 group. A total of 4,850 sequences were repeated at least twice, representing a total of 20,248 reads in our dataset. Among those sequences, 20 were repeated more than 40 times and one 2,678 times. We decided to investigate further the 20 most abundant sequences. The most abundant sequence was exactly the consensus sequence of the C1 group, while all the others corresponded to this sequence with single nucleotide variations and/or deletions, as indicated on Table 1. The absence of repeated sequences in the C2 group let us hypothesize, by contrast, that the different repeats observed in the C1 group may directly reflect the presence of strictly identical sequences in the *Cercopithecus solatus* genome. As Ion Torrent sequencing has been reported to give rise to sequencing errors, we decided to search if the identical sequences were obtained from reads collected in both orientations. We found that five out of the 20 sequences were associated with a strong bias for read orientation (Table 1). Within these five sequences, the three more abundant (2, 3, 8) represented deletions within a homopolymer tract, while the two others (15, 20) corresponded to the combination of the two most abundant deletions (found in 2 and 3) with the most abundant single nucleotide variation (found in 4). Deletions within homopolymer tracts have already been shown to be inherent to the Ion torrent Technology [41] and the orientation bias we observed let us conclude that sequences displaying these deletions were non-relevant

artifacts. On the contrary, all other sequences observed in high copy number, which were all obtained in both sequencing orientations, would correspond to sequence variants that are present with a high abundance in the *Cercopithecus solatus* genome.

Table 1 Analysis of alpha satellite sequences found in high copy number in the monomer dataset

Id	Sequence	Number	Forward (%)
1	Consensus	2678	46
2	C114Del	486	1*
3	T101Del	357	99*
4	T39G	242	46
5	A40C	101	56
6	T121A	92	41
7	T74G	78	47
8	T80Del	78	100*
9	G84C	78	41
10	C42G	76	53
11	G1A	74	43
12	A110G	65	48
13	A112T	61	38
14	T19C	59	37
15	T39G-C114Del	57	0*
16	A151C	56	52
17	G79C	55	53
18	C89T	53	49
19	G1T	46	63
20	T39G-T101Del	41	98*

The sequences are ordered and numbered according to the number of identical copies of the sequence in the monomer dataset. The "Sequence" column indicates how each sequence differs from the consensus sequence of the C1 family, using standard notations. The "Number" column displays the number of identical copies of the sequence in the monomer dataset. The "Forward" column displays the percentage of reads obtained in the forward orientation (i.e. the orientation of our reference sequence)

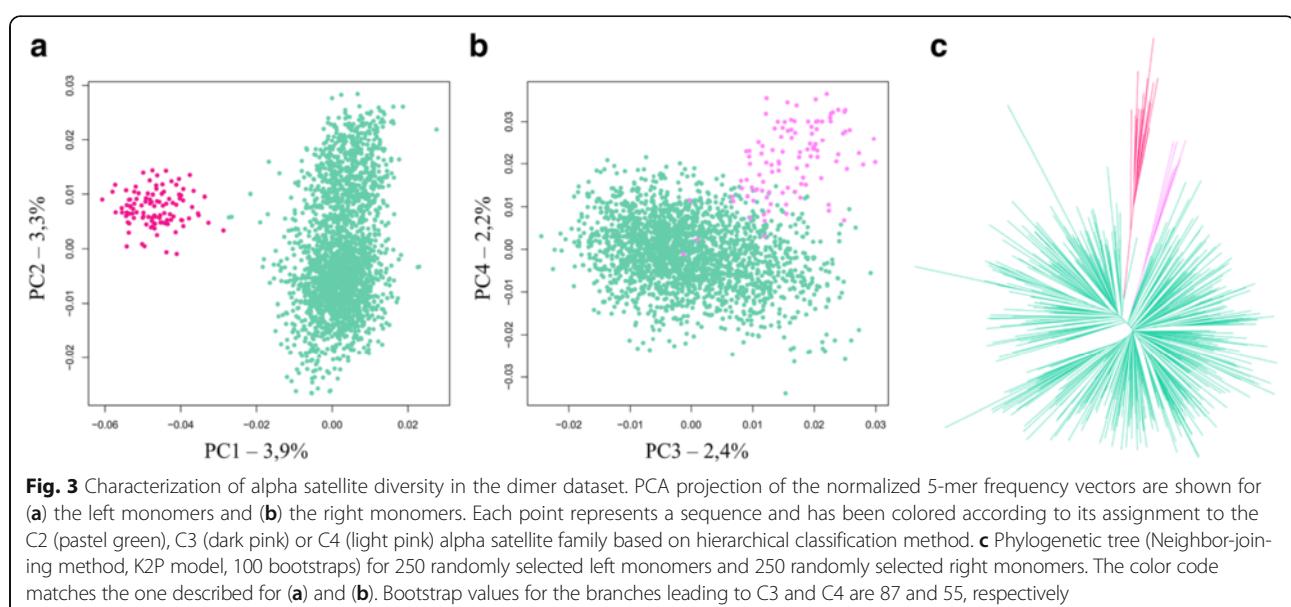
Strong biases for read orientation are shown with an asterix (*)

Characterization of alpha satellite diversity in the dimer dataset

Among the 3,568 sequences recovered from the dimer dataset, 1,095 contained an intact XmnI restriction site approximately located in the middle, suggesting that the enzymatic digestion was not complete. The remaining 2,473 sequences, which did not possess the XmnI restriction site, were split using an alignment-based process (see Methods) and the resulting monomers were submitted to a length filter, giving a total of 2,408 associated left and right monomers. We focused first on these sequences and submitted them to the same process as described above. PCA showed the existence of two groups for both the left and right monomers, which could be discriminated using HCA (Fig. 3a and b). A comparison of the consensus sequences of each group revealed that the most abundant sequence set for both the left and right monomers belonged to the C2 group, while the two smaller sequence sets had consensus sequences that differed from each other and from the consensus of the C1 and C2 groups (Fig. 2). These two new groups of sequences, which represented 5% of the left or right monomers, will be from now on called C3 and C4 and shown in dark and light pink, respectively. We decided to build a phylogenetic tree with left and right monomers mixed together, using the described color code (Fig. 3c). This tree confirmed the existence of C3 and C4 as separate groups. Their respective average sequence identities were measured to be 86% and 83%. We also checked that left and right monomers belonging to the C2 group could not be distinguished from each other on a phylogenetic tree (Additional file 1: Figure S2), or from the C2 sequences present in the monomer dataset. The comparison of the consensus sequences showed that the C3

and C4 groups differed much more from each other and from the C1 and C2 groups than C1 and C2 differed from each other. Interestingly, C4 was the only group with a consensus length of 171 bp instead of 172 bp. A search for CENP-B and pJalpha boxes showed that most sequences within the C4 group contained a pJalpha box (75%) while the CENP-B box was absent, like observed for the C1 and C2 groups. By contrast, neither the pJalpha box nor the CENP-B box was found in the sequences from the C3 group (Fig. 2).

The dimer dataset was also used to infer information regarding how monomers belonging to different groups associated with each other. All left and right monomers were assigned to one of the C1 to C4 groups (see Methods). Additional file 2: Table S2 reports the results of these assignments as well as associations between left and right monomers, distinguishing dimers that contained the XmnI site (X dataset) and those where the XmnI site was absent (noX dataset). We noticed that sequences from the C1 group were absent from the noX dataset and were poorly represented in the X dataset. This result may appear unexpected as 82% of the sequences from the monomer dataset belonged to the C1 group. Two hypotheses may explain this observation: the high sequence identity within the C1 group may reduce both the likelihood of the inactivation of the XmnI digestion site through mutations and the sequencing efficiency of dimers (see above). A statistical analysis of the X dataset showed that left monomers from the C1 and C2 groups were preferentially associated to right monomers from the same group (Additional file 2: Table S2), which suggests that sequences from the C1 and the C2 groups are tandemly repeated in the *Cercopithecus solatus* genome. C2-



C2 associations were also found to predominate within the noX dataset. Interestingly, left monomers from the C3 group were preferentially associated to right monomers from the C4 group, suggesting the existence of a higher order organization with repeats containing at least two monomers belonging to different groups.

Genomic distribution of alpha satellite families on *Cercopithecus solatus* chromosomes

We were next interested in studying the genomic distribution of the four groups of sequences identified above. Short oligonucleotide probes have been shown to be more efficient at distinguishing alpha satellite sequences that differ by very few nucleotides compared with classical probes obtained by random priming or nick translation [42, 43]. We chose to use synthetic 18-mer oligonucleotides carrying locked nucleic acid (LNA) modifications at one out of two positions and capable of forming at least 7 GC base pairs, as previous work had demonstrated their interest for the detection of alpha satellite sequences [44]. An in silico probe selection process was implemented in order to identify among the most common 18-mer sequences within a group (found in more than 20% of the monomers) those that were

specific for this group (found in less than 3% of the monomers of other groups). As we expected that oligonucleotide probes may still hybridize in the presence of one mismatch, we calculated the expected binding frequencies when one mismatch was present and applied the same selection criteria once again. Additional file 1: Figure S3 reports the sequences that best fitted with our requirements, albeit not completely. Due to the high sequence similarity between sequences within the C1 and C2 groups, probes had to distinguish sequences that differ mainly by only two nucleotides or even a single one (Additional file 1: Figure S3). The two sets of probes selected to target the C1 and C2 groups were therefore designed so that they would compete with each other if used simultaneously. The detection systems (fluorophores or haptens) were chosen in order to allow various combinations of probes to be tested together.

A first series of FISH experiments on *Cercopithecus solatus* metaphase spreads was performed using probes C1a and C2a or C1b and C2b simultaneously. Probes targeting the C1 group produced intense signals at the centromere (primary constriction) of all chromosomes except a single one (Fig. 4a and c), while probes targeting the C2 group provided signals that are located in the pericentromeric

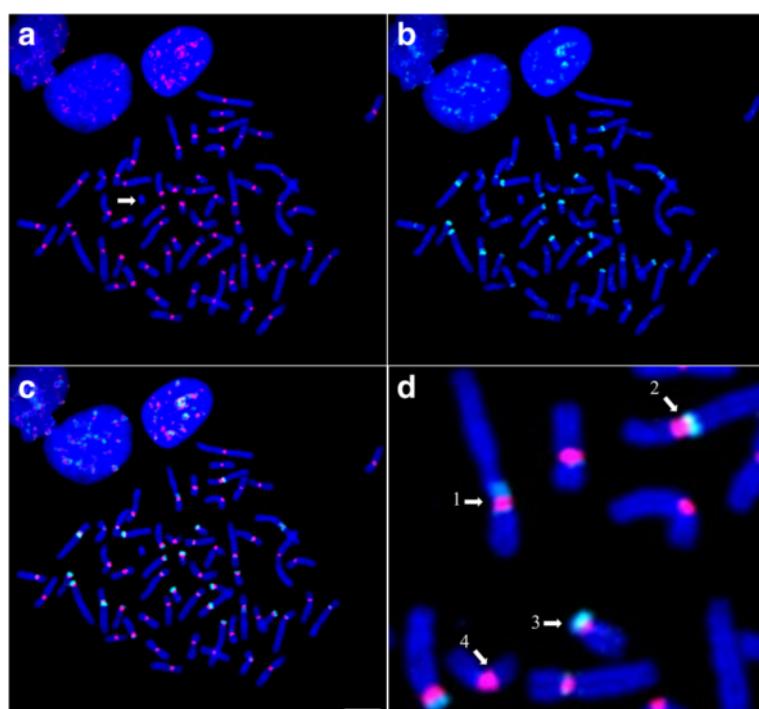


Fig. 4 FISH analysis of the C1 and C2 alpha satellite families characterized in the monomer dataset. Probes C1b and C2b are hybridized simultaneously to *Cercopithecus solatus* chromosomes, which are colored in blue. **a** Hybridization of probe C1b is shown in red. The arrow points to a single unlabeled chromosome. **b** Hybridization of probe C2b is shown in green. **c** Combined signals from (a) and (b). **(d)** Focus on image (c) showing in details the different types of distribution of the C2b signals relatively to C1b. 1: C2b labels both pericentromeric regions, 2: C2b labels one pericentromeric region toward the long arm, 3: C2b labels one pericentromeric region toward the short arm of an acrocentric chromosome, 4: no C2b signal can be observed on this chromosome. Scale bar = 10 μm

regions (around the primary constriction) of several chromosomes pairs with different labeling patterns (Fig. 4b and c and Additional file 1: Figure S4a). Some chromosomes were extensively labeled by C2 probes on both sides of the centromere, others seem to be labeled at only one side, and others seemed to display no signal (see arrows on Fig. 4d). Stronger signals were observed on the acrocentric chromosome short arms (Additional file 1: Figure S4a). When these probes were used alone, each C1a or C2a probe produced a labeling pattern similar to what was observed in the presence of the other. On the contrary, each C1b or C2b probe used alone labeled regions that are larger than in the absence of the other. These experiments suggest that our probes may hybridize to sequences that differ by a single nucleotide (i.e. C1b binds to sequences from the C2 group and C2b binds to sequences from the C1 group) but that this binding is inhibited in the presence of an adequate competitor probe. In addition, when target sequences differ by at least two nucleotides, a specific detection is achieved in the absence of competitor.

Additional experiments showed that in presence of competitors, the signal produced by C1a overlapped with the signal produced by C1b and the signal produced by C2a almost perfectly overlapped with the one produced by C2b (Additional file 1: Figure S5). This observation supports the idea that the labeling patterns observed with the chosen oligonucleotide probes reflect the distribution of the sequence groups identified by sequence analysis. Moreover, the absence of overlap between signals provided by probes targeting sequences from the C1 and C2 groups suggests that monomers within each group are clustered together and do not mix with each other. Combined with the arguments described above that are in favor of a tandem organization of monomers for both the C1 and C2 groups, these features support the fact that the C1 and C2 groups of sequences represent distinct families of alpha satellite DNA that display a monomeric organization in the genome of *Cercopithecus solatus*.

Further experiments were performed with probes targeting the C3 and C4 groups. All C3 and C4 probes provided identical labeling patterns, with a strong signal located on a single chromosome, as well as very weak pericentromeric signals on some other chromosomes (Fig. 5 and Additional file 1: Figure S5). The chromosome labeled by the C3 and C4 probes, which is in fact the chromosome that was not labeled by the probes targeting C1, was also identified to be the Y chromosome by cytogenetic experiments (Additional file 1: Figure S4b). The colocalization of probes targeting the C3 and C4 groups and the absence of overlap with probes targeting the C1 or C2 group is consistent with the sequence analysis described above. These results taken together suggest thus that sequences belonging to the C3 and C4 groups represent additional families of alpha satellite DNA that display a higher order organization within the genome of *Cercopithecus solatus*. As a further control of the consistency between the results from FISH experiments and sequence analysis, we showed that a 13-mer LNA probe that was designed to target the four C1 to C4 groups of sequences (called Cx) provided signals that overlapped with the combined signals of probes targeting each group, i.e. was able to label all chromosomes within the centromeric and pericentromeric regions (Additional file 1: Figure S6).

We were also interested in studying the chromosomal distribution of some of the repeated sequences found in high copy number in the monomer dataset. The results of our previously described FISH experiments suggest that the specific detection of single nucleotide variations may be difficult to achieve using individual probes but that using several probes in competition may provide the possibility to achieve the required level of specificity. Therefore, we designed new oligonucleotide probes targeting a common region, aiming at distinguishing three different highly repeated sequences with single nucleotide variations (Fig. 6, see Methods). When all probes were used in combination, probes targeting sequence 4 (T39G

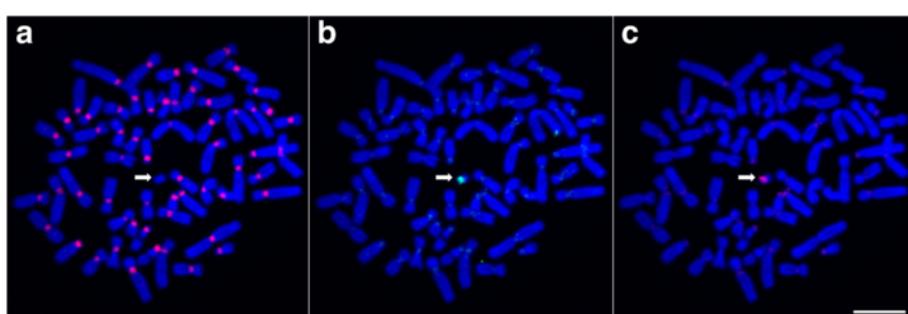


Fig. 5 FISH analysis of the C3 and C4 alpha satellite families characterized in the dimer dataset. Probes C1a, C3a and C4a are hybridized simultaneously to *Cercopithecus solatus* chromosomes, which are colored in blue. **a** Hybridization of probe C1a is shown in red. A single chromosome (shown with an arrow) is not labeled. **b** Hybridization of probe C3a is shown in green. **c** Hybridization of probe C4a is shown in red. The pericentromeric regions of several chromosomes are sparingly labeled by C3a and C4a. Scale bar = 10 μ m

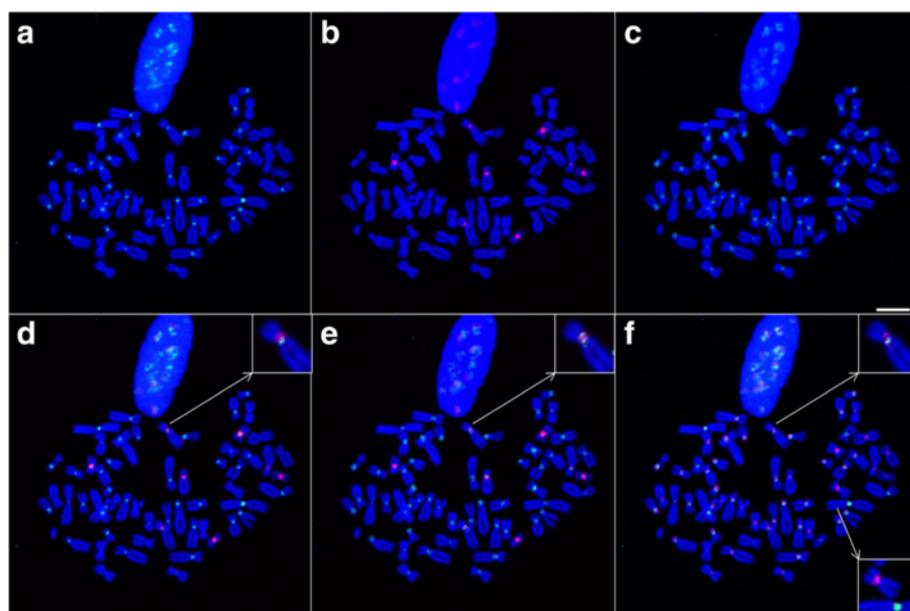


Fig. 6 FISH analysis of the chromosomal distribution of three sequences found in high copy number. Probes T39G, A40C and C42G and the competitor oligonucleotide TAC₂₀ (complementary to the C1 consensus) are hybridized simultaneously to *Cercopithecus solatus* chromosomes, which are colored in blue. **a** Hybridization of probe T39G is shown in green. **b** Hybridization of probe A40C is shown in red. Eight chromosomes are labeled. **c** Hybridization of probe C42G is shown in green. **d** Combined signals from (a) and (b). **e** Combined signals from (b) and (c). **f** Combined signals from (a) and (c) with probe T39G shown in green and probe C42G shown in red. Upper inset in (d), (e) and (f) shows one chromosome where signals from the 3 probes do not overlap. Lower inset in (f) shows two chromosomes, one being labeled by probe T39G and the other by probe A40C. Scale bar = 10 μ m

variation) and sequence 10 (C42G variation) seemed to label all chromosomes, albeit with non-overlapping patterns (see for example insets in Fig. 6f), while probe targeting sequence 5 (A40C variation) was clearly shown to produce a signal on only 8 chromosomes.

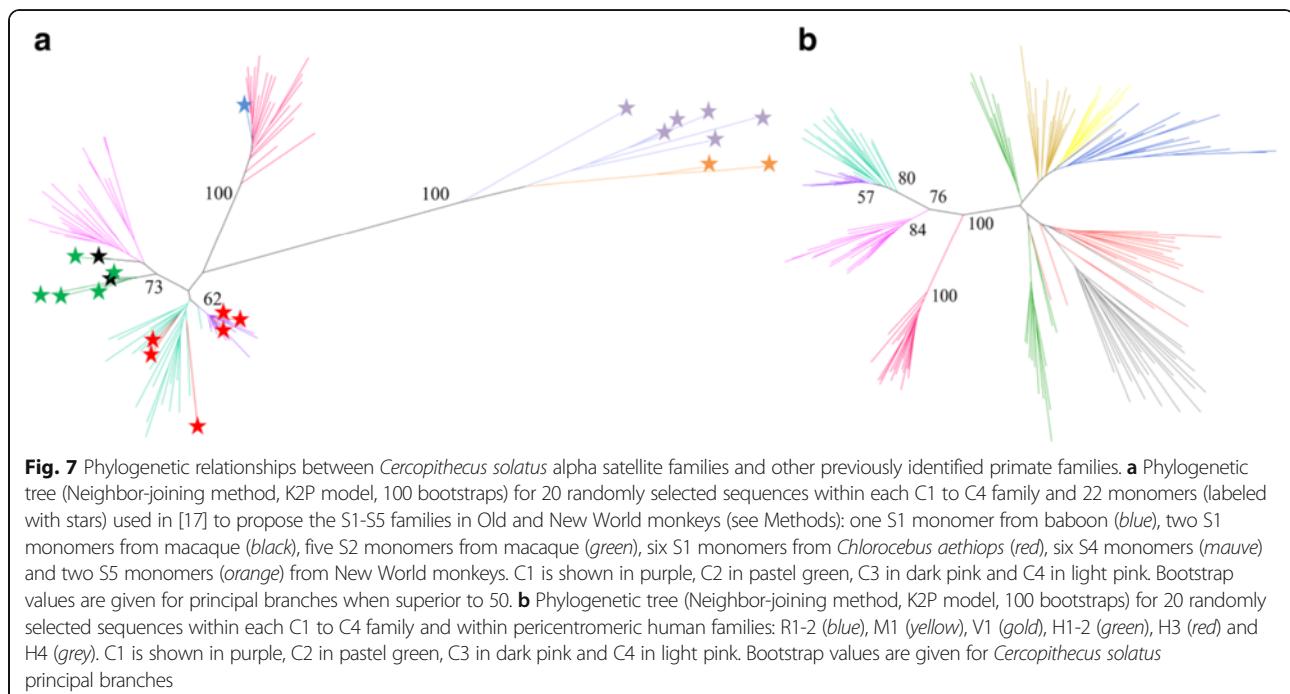
Comparison of *Cercopithecus solatus* alpha satellite families with known primate families

The sequence families defined above may provide information regarding the evolutionary history of alpha satellite DNA in Primates. We were therefore interested in investigating phylogenetic relationships between these families and alpha satellite sequences that were previously described for other primate species. Interestingly, the first alpha satellite consensus sequence ever described, which was obtained for the cercopithecini *Chlorocebus aethiops* [45], was exactly the same as the consensus sequence of our C1 family, which is also the most abundant repeated sequence in our dataset. This identity suggests the conservation of the C1 family between Cercopithecini species. Although very few sequences were available, a tentative classification was previously proposed for alpha satellite DNA present in Old and New World monkeys, involving five families termed S1 to S5 [17]. We built a phylogenetic tree containing 50 sequences randomly selected within each of our C1 to C4 families and several sequences representative for S1, S2, S4 and S5 (Fig. 7a, see Methods). The S1 sequences obtained

from *Chlorocebus aethiops* were intermingled in this tree with our C1 and C2 sequences. Other sequences classified in S1 but obtained from other species were dispersed in other parts of the graph, suggesting that the proposed S1 family was not relevant. There was also no clear proximity of each one of the C1 to C4 family with sequences belonging to the so-called S2, S4 or S5 family. The phylogenetic tree showed on the contrary that sequences from macaque (identified as S1 or S2) may form a sister group of the C4 family whereas the only available baboon sequence (identified as S1) was close to the C3 family. None of the sequences from macaque or baboon resembled those from our C1 or C2 family. All these results suggest that, contrary to S4 and S5, S1 and S2 do not correspond to alpha satellite families. We also built a phylogenetic tree involving our C1-C4 families and seven families (termed M1, R1-2, V1, and H1 to H4) that were previously identified in human pericentromeric regions, some of them being reported as similar to sequences found in other primates [19] (Fig. 7b). The tree suggests that alpha satellite families found in *Cercopithecus solatus* have an evolutionary history that is largely independent from that of the alpha satellite families found in human pericentromeric regions.

Discussion

Despite the recent generalization of high-throughput sequencing, application of these new technologies to the



study of repeated DNA remains scarce [46, 47]. Here, we present an original experimental and computational framework for studying repeated DNA. We have focused on a single Cercopithecini species where the diversity and organization of alpha satellite DNA are described in details. Our approach relies on sequencing of gel purified alpha satellite monomers and dimers obtained by restriction enzyme digestion of genomic DNA, followed by sequence analysis and FISH experiments with carefully designed probes.

We detected four alpha satellite families, called C1 to C4, in the *Cercopithecus solatus* genome. Additional families may have been missed by our approach, for example because they would not contain restriction sites for XmnI. Although some technical issue had drastically reduced the number of available amount of sequences containing two monomers, the dimer dataset provided information about the structural organization of each family, showing that the C1 and C2 families adopt a monomeric organization, while C3 and C4 would associate into HORs. Our data do not allow concluding if the C3-C4 dimers are tandemly repeated or represent only a part of a longer HOR involving other monomers, but suggest that such structures, which have also been observed in New World monkey genomes [48], may be widespread in Primates. It had previously been reported, using a limited number of sequences, that alpha satellite sequences in Old World monkeys contained a pJalpha binding site and no CENP-B binding site [22, 49, 50]. Our data provide further support to this observation which holds true for three of the four newly identified

families. The absence of any of these two binding sites in the C3 family represents an oddity but one should notice that as sequences from the C3 family are associated with sequences from the C4 family into a HOR organization, the pJalpha binding site remains present in the repeated motif. We detected several sequences in our dataset that were repeated identically a high number of times (up to several thousands). As our protocol does not contain any PCR amplification before capture of individual sequences on beads, the abundance of these sequences may reflect their natural abundance within the *Cercopithecus solatus* genome, provided one is able to identify potential artifacts resulting from sequencing errors among those sequences.

The high similarity between *Cercopithecus solatus* alpha satellite families, especially C1 and C2, the consensus of which differ at only a few nucleotide positions, required the implementation of a highly specific FISH detection to infer their chromosomal distribution. Our results emphasize the interest of short LNA-modified oligonucleotide probes that are here shown to be able to distinguish sequences that differ by only two nucleotides. It is even possible to distinguish a single nucleotide variation between two sequences by using two probes targeting each sequence variant simultaneously. In all our experiments, we cannot exclude the possibility that probes also hybridize to sequences that are not perfectly complementary, nor that some signals do not come from sequences that are present in the *Cercopithecus solatus* genome but not found in our datasets. Nevertheless, the absence of cross labeling between the probes targeting different families and the consistency of hybridization results with predictions inferred from sequence analysis

support our probe design strategy and stands for the accuracy and the exhaustiveness of our description of the alpha satellite component of *Cercopithecus solatus*.

Our FISH experiments showed that the C1 family, which is the most conserved (95% mean sequence identity), displays a centromeric localization while the more divergent C2 family (85% mean sequence identity) displays a pericentromeric localization. According to the age-gradient based model for centromere evolution [3, 33], we may speculate that this pattern results from a peculiar evolutionary history where the C2 family, an old family of sequences, had occupied a centromeric position in an ancestor of *Cercopithecus solatus*. This family would then have been displaced towards pericentromeric regions following the amplification of more recent sequences from the C1 family at the centromere. Unequal crossing over at nearly identical repeats is thought indeed to lead to the homogenization of the core centromere, while mutations would only affect repeats outside of the core centromere [3, 19, 51–53]. An alternative but non-exclusive hypothesis would attribute distinct functional roles to both families, for example centromere function to C1 and cohesion of sister chromatids to C2, as it has been proposed for mouse minor and major satellite sequences, respectively [54]. Interestingly, acrocentric chromosome short arms display a very large amount of C2 sequences as revealed by intense FISH signals. This observation provides support to a previous hypothesis according to which acrocentric chromosomes may physically interact and exchange genetic material [55, 56]. The fact that the C3-C4 dimers are found on the Y chromosome and are almost absent from other chromosomes may be explained by the fact that the Y chromosome is excluded from recombination events with non-homologous chromosomes, as was observed in mice [57]. Finally, the observation of the distribution of one of the highly repeated sequence variants on only 8 chromosomes supports the existence of local alpha satellite homogenization events in the *Cercopithecus solatus* genome.

Previous studies had considered alpha satellite DNA in Cercopithecini as poorly diversified [17]. Our results show that at least four alpha satellite families can be present in a single species, with complex chromosomal distribution and organizational patterns. Comparative studies including repetitive DNAs from different species have already been shown to provide new insights into genome and species evolution [58]. Our approach will permit not only to investigate the taxonomic distribution of alpha satellite families but also to study their organizational pattern, their chromosomal distribution as well as the existence of conserved highly repeated sequence variants. Phylogenetic analysis have demonstrated that the C1 to C4 families represent newly identified entities that do not correspond to previously proposed alpha satellite families. Although the available data are in favor of an apparent conservation of

both the C1 and C2 families between *Cercopithecus solatus* and *Chlorocebus aethiops*, further studies will be required to better understand the dynamics of alpha satellite DNA in Old World monkeys and in other primates.

Conclusions

In summary, we have presented here a generally applicable strategy that provides, for a single species, a comprehensive description of alpha satellite sequence diversity and organization. Our approach, which is easy to implement and cost-effective, provides an opportunity to characterize satellite DNA in all species where a characteristic enzymatic ladder pattern can be obtained. Comparing different individuals and different species will provide new insights into the dynamics at which new satellite families or new highly repeated sequence variants appear during the course of evolution and transfer between chromosomes. The better description of the structure of heterochromatic regions also provides potential for enhancing the epigenetic characterization of these regions as well as understanding the regulatory functions of heterochromatin.

Methods

DNA collection and metaphase preparations

Fibroblast samples of *Cercopithecus solatus* (ID: 2012–028, male sample, ethic permission n° FR1207510445-I) from the Collection of cryopreserved living tissues and cells of vertebrates (RBCell collection, Muséum national d'Histoire naturelle, Paris) were used for DNA extraction and metaphase preparations. DNA was extracted using the Omega Bioteck Tissue DNA Kit (Doraville, USA). Cell cultures and metaphase preparations were achieved according to [59].

Alpha satellite DNA isolation and sequencing

The Serial Cloner software (Serial Basics, serialbasics-free.fr) was used to perform in silico digestions of the Cercopithecini alpha satellite sequences registered as such in Genbank (Accession numbers: AM235889, AM235890, AM237210, AM237214, AM237213, AM237212, X04339, V00145, M26844 and AM237211), which contained both monomers and dimers. The restriction site of the XmnI restriction enzyme (GAANNNNTTC) was observed once in a great proportion of monomers and twice in almost all dimers. XmnI was then used to digest *Cercopithecus solatus* DNA in vitro. 10 µg of *Cercopithecus solatus* genomic DNA were digested for 4 h 30 min at 37°C with 60 units of XmnI activity (New England Biolabs) in a total volume of 34 µL. The enzyme was inactivated for 20 min at 65°C. The sample was loaded on a 1% agarose gel after addition of 6.8 µL loading buffer (50% glycerol) and electrophoresis was performed in 0.5X Tris-borate-EDTA buffer, at room temperature for 2 h 45 min at 100 V. The gel was briefly stained with ethidium bromide and then imaged by UV

transillumination. Bands corresponding to alpha satellite monomers (~170 bp) and dimers (~340 bp) were cut and DNA was extracted from the gel with the Omega Biotek Gel extraction kit and resuspended in 100 μ l of elution buffer. About 220 ng and 110 ng were obtained for the 170 bp and 340 bp samples, respectively.

Sequencing was performed on a PGM sequencing platform (Ion Torrent technology) using the 400 bp sequencing kit. Two libraries were generated using 50 ng of both blunt digest pools and the Ion Plus Fragment Library Kit (4471252, Life Technologies) and tagged with Ion Xpress barcode adapters (4471250, Life Technologies). After purification (1.8X) with Ampure XP Beads (A63880, Agencourt Bioscience, Beverly, USA), the libraries were quantitated using a SsoAdvanced Sybr Green qPCR assay (Biorad, Hercules, USA) based on a custom *E. coli* reference library. After a dilution of each library down to 26 pM, 0.22 fmol for the 170 bp library and 0.44 fmol for the 340 bp library were pooled as templates for the clonal amplification on Ion Sphere particles during the emulsion PCR, performed on a One Touch2 emPCR robot according to the Ion PGM Template OT2 400 Kit user guide (4479878, Life Technologies). The amplification products were loaded onto an Ion 316v2 chip (4483324, Life Technologies), and subsequently sequenced according to the Ion PGM Sequencing 400 Kit user guide (4482002, Life Technologies). After standard filtration of the raw reads (polyclonal and low quality removal), the Ion Torrent sequencing yielded 204,990 sequences for the 170 bp pool and 353,683 sequences for the 340 bp pool. They were deposited in the NIH Short Read Archive (SRA accession numbers SRX1595681 and SRX1595679).

Alpha satellite sequence filtering

All sequences with an average Phred score lower than 25, a length outside the range 162–182 bp for monomers and 324–364 bp for dimers, and sequences without the XmnI digested sites at the extremities (5'-NNNTTC ... GAANN-3') were not considered for further analysis. Alpha satellite sequences were identified with a BLAST search against a reference alpha satellite sequence of *Chlorocebus aethiops* (AM23721) [60]. Using default BLAST parameters, all sequences exhibiting a hit longer than 80 bp for monomers and 160 bp for dimers were considered as alpha satellite sequences and conserved for the following analysis. All sequences were then reoriented if necessary in order to match the orientation of the reference alpha satellite sequence. The orientation information was preserved for investigations regarding reading biases.

Processing of dimeric sequences was performed as follows. When an XmnI site was present in the middle of these sequences, it was used for separating both monomers, providing the so-called left and right monomers located on the 5' side and on the 3' side of the sequence, respectively.

Dimers that did not contain any XmnI site in the middle were aligned against a synthetic sequence formed by two consecutive copies of the reference sequence using the Needleman-Wunsch algorithm [61] to identify the monomer limits and split them into left and right monomers according to the same rule as described above. All pairs with at least one monomer outside the 162–182 bp range were discarded. Pairing information was conserved to study association between left and right monomers.

Alpha satellite sequence characterization

Monomeric sequences were compared using their 5-mer composition in order to identify putative alpha satellite groups without direct alignment. For each set of monomers, the 5-mer frequency table was analyzed using a principal component analysis (PCA) to reduce the space complexity and enable data visualization on the first factorial planes. Sequences were classified into groups by using a hierarchical clustering method (HCA) based on the Ward criterion [62] applied to the Euclidean distances calculated from the 100 first principal components of the PCA. Because of the size of the monomer dataset, direct classification of the sequences using HCA was not possible. Instead, HCA was applied on 2,500 randomly selected sequences which were used to train a linear discriminant model. This model has been finally used to classify all the other monomers. The dimer dataset was analyzed in two different ways: 1) monomers extracted from dimers without XmnI sites were classified by using an HCA based on a PCA, 2) monomers extracted from dimers with a XmnI site have been classified by using a LDA trained to recognize the C1-C4 groups.

Because of the size of the datasets, the phylogenetic trees, the consensus sequences and the sequence distance analysis were conducted with different subsets of randomly selected sequences, using a homemade python script. The selected sequences were aligned using MUSCLE [63] and analyzed with SeaView [64]. The phylogenetic trees were built by using the Neighbor joining algorithm and the Kimura 2-parameters distance. Reliability of nodes was assessed using 100 bootstrap iterations. The relatively low bootstrap values observed in the trees can be explained by a limited number of family specific sites, i.e., the informative sites, into the alignments. Nevertheless, the same clustering of the families and the same relationship between these families have been observed with all the trees generated with different randomly selected sequences.

CENP-B and pJalpha boxes were searched with the patterns TTCGTTGGAARCGGGA and TTCCCTTTY-CACCRTAG respectively [40] by using the program Fuzznuc [65] and allowing 2 mismatches. All statistical analyses were conducted with R [66]. Our R scripts and other programs are available upon request.

The S1-S5 monomers used in Fig. 7a have been isolated from the sequences described in [17]. All these monomers have been extracted by using the homologous position of the XmnI digestion site as a starting point (XmnI phase) in order to be aligned with the monomers of *Cercopithecus solatus*. Unfortunately, no full length S3 monomer was available in this phase. To obtain the Genbank accession numbers and the alignment of the used S1-S5 monomers, see [17] and Additional file 3: Text S1. Human monomers from old and ancient families (M1, R1-2, V1, H1-H4) used in Fig. 7b have been isolated from the human Xp chromosome sequence (Genbank ID NT_011630) by using the homologous position of the XmnI digestion site as a starting point. Monomers have been assigned to a family according to their location along the sequence and the annotations provided in [19] (see Additional file 4: Text S2 for alignment).

Oligonucleotide probes

Short oligonucleotide probes (18 nucleotides in length) were designed in order to target specifically the different alpha satellite families identified in *Cercopithecus solatus*, by systematic prediction of binding frequencies based on the sequencing results. In some instances, when the 18-mer sequence did not allow forming at least 7 GC bp upon hybridization to the complementary strand, length was increased to 19. Sequences and binding frequencies are available in Additional file 1: Figure S3, which also provides details about the positions of locked nucleic acid (LNA) modifications in the probes. These positions were selected based on previous experience in order to achieve a good binding affinity and specificity [44]. When possible, we selected probes that were perfectly complementary to more than 20% of the sequences from the target group and to less than 3% of the sequences from the other groups. Additional file 1: Figure S3 also provides the expected binding frequencies if hybridization is possible despite the presence of one mismatch between the probe and its target. To target three sequences found in high copy number in the monomer dataset, we designed four LNA-modified probes (LNA are written in lower case and classic nucleotides are written in upper case): probe T39G (5' TgTtCtGtTCaTtCaTcTc3', 5' AlexaFluor488), probe A40C (5' TgTtCtGtGAaTtCaTcTc3', 3' Digoxigenin), probe C42G (5' TgTtCtCtTAaTtCaTcTc3', 3' Biotin) and probe TACco (5' TgTtCtGtTAaTtCaTcTc3') which is complementary to the C1 consensus sequence. LNA-modified probes were purchased from Eurogentec (Seraing, Belgium).

FISH experiments

FISH were performed on metaphase chromosome preparations. Hybridization solutions were prepared by diluting the oligonucleotide probes to a final concentration

of 0.1 μM in a hybridization solution consisting of 2X SSC pH 6.3, 50% deionized formamide, 1X Denhardt solution, 10% dextran sulfate, and 0.1% SDS. 20 μL of the hybridization solution were deposited on each slide and covered with a coverslip. The slides were then heated for 3 min at 70°C and hybridized for 1 h at 37°C in a Thermobrite apparatus (Leica Biosystems). Then, each slide was washed twice in 2X SSC at 63°C. Preparations were then incubated in blocking solution (4% bovine serum albumin (BSA), 1X PBS, 0.05% Tween 20) for 30 min at 37°C to reduce nonspecific binding. Then, depending on the combination of probes, the following antibodies were used for subsequent revelations: Alexa 488-conjugated streptavidin (1:200; Life Technologies, Foster City, USA), Cy5-conjugated streptavidin (1:200; Caltag Laboratories, Burlingame, USA), FITC-conjugated sheep anti-digoxigenin (1:200; Roche, Lewes, UK), and Rhodamine-conjugated sheep anti-digoxigenin (1:200; Roche). All antibodies were diluted in blocking solution containing 1X PBS, 0.05% Tween 20, and 4% BSA. Antibody incubation lasted for 30 min at 37°C. All washings were performed in 2X SSC, 0.05% Tween 20. Chromosomes were counterstained with DAPI (4',6-diamidino-2-phenylindole) by pipetting 40 μL of a 5 μg/mL solution onto the slides, incubating for 5 min and then briefly washing in 1X PBS. Slides were mounted by adding a drop of Vectashield Antifade Mounting Medium (Vector Laboratories, Burlingame, USA) and covering with a coverslip.

Image acquisition and analysis

Metaphases were imaged using an Axio Observer Z1 epifluorescent inverted microscope (Zeiss) coupled to an ORCA R2 cooled CDD camera (Hamamatsu). The Axio Observer Z1 was equipped with a Plan-Apochromat 63x 1.4 NA oil-immersion objective and the following filters set: 49 shift free for DAPI (G365 / FT395 / BP445/50), 38 HE shift free for FITC/Alexa488 (BP470/40 / FT495 / BP525/50), homemade sets for Rhodamine (BP546/10 / FF555 / BP 583/22) and for Cy5 (BP643/20 / FF660 / BP684/24). The light source was LED illumination (wavelengths: 365 nm, 470 nm or 625 nm) except for Rhodamine, for which a metal halide lamp HXP120 was preferred. Immersion oil of refractive index 1.518 at 23° C was used. Color-combined images were reconstructed using ImageJ [67]. At least ten metaphases were visualized for each experiment, which all confirmed the described patterns.

Additional files

Additional file 1: Figure S1. Migration profiles of *Cercopithecus solatus* genomic DNA digested with XmnI. **Figure S2.** Phylogenetic tree for left and right alpha satellite monomers from the dimer dataset. **Figure S3.** LNA-modified probes used to target the C1 to C4 alpha satellite families on

Cercopithecus solatus chromosomes. **Figure S4.** Distribution pattern of C2 and C3 alpha satellite families on *Cercopithecus solatus* chromosomes. **Figure S5.** Comparison of FISH signals from different probes targeting identical alpha satellite families. **Figure S6.** Comparison of the hybridization pattern of the Cx probe with those of probes targeting the C1 to C4 alpha satellite families. (DOCX 1508 kb)

Additional file 2: Table S1. Filtering steps from *Cercopithecus solatus* raw data to alpha satellite monomer and dimer datasets. **Table S2.** Alpha satellite family associations in *Cercopithecus solatus* dimer dataset. (DOCX 40 kb)

Additional file 3: Text S1. Alignment used to generate the phylogenetic tree from Fig. 7a. (TXT 28 kb)

Additional file 4: Text S2. Alignment used to generate the phylogenetic tree from Fig. 7b. (TXT 48 kb)

Acknowledgments

This manuscript is dedicated to the memory of our friend and colleague Florence Richard. We are grateful to Peggy Motsch, the ECOSOL project (ECOlogie de *Cercopithecus SOLatus*) and the Centre International de Recherches Médicales de Franceville for providing the *Cercopithecus solatus* samples to the RBCell collection (MNHN). We thank the Service de Systématique Moléculaire (UMS2700, MNHN), and especially Delphine Gey, Régis Debruyne and José Utge for the sequencing experiment. We are much obliged to Anne-Marie and Bernard Dutrillaux for their advice in metaphase preparation and analysis as well as to François Loll for the technical support he provided regarding FISH experiments and image acquisition.

Funding

This work was supported by the Actions Thématisques Muséum "Génomique et Collections" and "Emergence".

Availability of data and material

The datasets supporting the conclusions of this article are available in the NIH Short Read Archive under SRX1595681 (<http://www.ncbi.nlm.nih.gov/sra/SRX1595681>) and SRX1595679 (<http://www.ncbi.nlm.nih.gov/sra/SRX1595679>) IDs.

Authors' contribution

CE, FR, LC and LP conceived and designed the experiments. LC, LP and MG performed the experiments. LC, LP and CE analyzed the data. LP and CE contributed reagents/materials/analysis tools. LC, CE and LP contributed to the writing of the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declared that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

¹Département Régulations, Développement et Diversité Moléculaire, Structure et Instabilité des Génomes, INSERM U1154, CNRS UMR7196, Sorbonne Universités, Muséum national d'Histoire naturelle, Paris, France.
²Département Systématique et Evolution, Institut de Systématique, Evolution, Biodiversité, UMR 7205 MNHN, CNRS, UPMC, EPHE, Sorbonne Universités, Muséum national d'Histoire naturelle, Paris, France. ³Université Versailles St-Quentin, Montigny-le-Bretonneux, France.

Received: 14 April 2016 Accepted: 2 November 2016

Published online: 14 November 2016

References

- Yunis JJ, Yasmineh WG. Heterochromatin, satellite DNA, and cell function. Structural DNA of eucaryotes may support and protect genes and aid in speciation. *Science*. 1971;174:1200–9.
- Warburton PE, Willard HF. Genomic analysis of sequence variation in tandemly repeated DNA. Evidence for localized homogeneous sequence domains within arrays of alpha-satellite DNA. *J Mol Biol*. 1990;216(1):3–16.
- Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF. Genomic and genetic definition of a functional human centromere. *Science*. 2001;294(October):109–15.
- Plohl M, Luchetti A, Meštrović N, Mantovani B. Satellite DNAs between selfishness and functionality: Structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. *Gene*. 2008;409:72–82.
- Feliciello I, Akrap I, Ugarković D. Satellite DNA Modulates Gene Expression in the Beetle *Tribolium castaneum* after Heat Stress. *PLOS Genet*. 2015;11(8):e1005466. Available from: <http://dx.plos.org/10.1371/journal.pgen.1005466>.
- She X, Horvath JE, Jiang Z, Liu G, Furey TS, Christ L, et al. The structure and evolution of centromeric transition regions within the human genome. *Nature*. 2004;430(7002):857–64.
- Maio JJ. DNA strand reassociation and polyribonucleotide binding in the African green monkey, *Cercopithecus aethiops*. *J Mol Biol*. 1971;56(3):579–95.
- Musich PR, Brown FL, Maio JJ. Highly repetitive component alpha and related alploid DNAs in man and monkeys. *Chromosoma*. 1980;80:331–48.
- Maio JJ, Brown FL, McKenna WG, Musich PR. Toward a molecular paleontology of primate genomes. II. The KpnI families of alploid DNAs. *Chromosoma*. 1981;83:127–44.
- Alves G, Seuanez HN, Fanning T. Alpha satellite DNA in neotropical primates (Platyrrhini). *Chromosoma*. 1994;103(4):262–7.
- Willard HF. Evolution of alpha satellite. *Curr Opin Genet Dev*. 1991;1(4):509–14.
- Rudd MK, Wray GA, Willard HF. The evolutionary dynamics of alpha-satellite. *Genome Res*. 2006;16:88–96.
- Rudd MK, Willard HF. Analysis of the centromeric regions of the human genome assembly. *Trends Genet*. 2004;20:529–33.
- Miga KH. Completing the human genome: the progress and challenge of satellite DNA assembly. *Chromosom Res*. 2015;421–26. Available from: <http://link.springer.com/10.1007/s10577-015-9488-2>.
- Schueler MG, Sullivan BA. Structural and functional dynamics of human centromeric chromatin. *Annu Rev Genomics Hum Genet*. 2006;7:301–13.
- Miga KH, Newton Y, Jain M, Altemose N, Willard HF, Kent WJ. Centromere reference models for human chromosomes X and Y satellite arrays. *2014;24(4):697–707*.
- Alexandrov I, Kazakov A, Tumeneva I, Shepelev V, Yurov Y. Alpha-satellite DNA of primates: old and new families. *Chromosoma*. 2001;110:253–66.
- Alkan C, Ventura M, Archidiacono N, Rocchi M, Sahinalp SC, Eichler EE. Organization and evolution of primate centromeric DNA from whole-genome shotgun sequence data. *PLoS Comput Biol*. 2007;3(9):1807–18.
- Shepelev VA, Alexandrov AA, Yurov YB, Alexandrov IA. The evolutionary origin of man can be traced in the layers of defunct ancestral alpha satellites flanking the active centromeres of human chromosomes. *PLoS Genet*. 2009;5(9). Available from: <http://dx.doi.org/10.1371/journal.pgen.1000641>.
- Catacchio CR, Ragone R, Chiatante G, Ventura M. Organization and evolution of Gorilla centromeric DNA from old strategies to new approaches. *Sci Rep*. 2015;5:14189. Available from: <http://www.nature.com/doifinder/10.1038/srep14189>.
- Willard HF, Waye JS. Hierarchical order in chromosome-specific human alpha satellite DNA. *Trends Genet*. 1987;3(7):192–19.
- Alexandrov IA, Medvedev LI, Mashkova TD, Kisselev LL, Romanova LY, Yurov YB. Definition of a new alpha satellite suprachromosomal family characterized by monomeric organization. *Nucleic Acids Res*. 1993;21(9):2209–15.
- Hayden KE. Human centromere genomics: Now it's personal. *Chromosom Res*. 2012;20(July):621–33.
- Alexandrov IA, Mitkevich SP, Yurov YB. The phylogeny of human chromosome specific alpha satellites. *Chromosoma*. 1988;96:443–53.
- Lee C, Wevrick R, Fisher RB, Ferguson-Smith MA, Lin CC. Human centromeric DNAs. *Hum Genet*. 1997;100:291–304.
- Jorgensen AL, Jones C, Bostock CJ, Bak AL. Different subfamilies of alploid repetitive DNA are present on the human and chimpanzee homologous chromosomes 21 and 22. *EMBO J*. 1987;6(6):1691–6.
- Archidiacono N, Antonacci R, Finelli P, Lonoce A, Rocchi M. Comparative Mapping of Human Alploid Sequences in Great Apes Using Fluorescence. *Genomics*. 1995;48:477–84.
- Warburton PE, Haaf T, Gosden J, Lawson D, Willard HF. Characterization of a chromosome-specific chimpanzee alpha satellite subset: evolutionary relationship to subsets on human chromosomes. *Genomics*. 1996;33(2):220–8.
- Malik HS, Henikoff S. Conflict begets complexity: The evolution of centromeres. *Curr Opin Genet Dev*. 2002;12:711–8.

30. Warburton PE, Willard HF. Interhomologue sequence variation of alpha satellite DNA from human chromosome 17: evidence for concerted evolution along haplotypic lineages. *J Mol Evol.* 1995;41(6):1006–15.
31. Schindelhauer D, Schwarz T. Evidence for a fast, intrachromosomal conversion mechanism from mapping of nucleotide variants within a homogeneous alpha-satellite DNA array. *Genome Res.* 2002;12:1815–26.
32. Roizès G. Human centromeric alploid domains are periodically homogenized so that they vary substantially between homologues. Mechanism and implications for centromere functioning. *Nucleic Acids Res.* 2006;34(6):1912–24.
33. Schueler MG, Dunn JM, Bird CP, Ross MT, Viggiani L, Rocchi M, et al. Progressive proximal expansion of the primate X chromosome centromere. *Proc Natl Acad Sci U S A.* 2005;102(30):10563–8.
34. Guschanski K, Krause J, Sawyer S, Valente LM, Bailey S, Finstermeier K, et al. Next-Generation Museomics Disentangles One of the Largest Primate Radiations. *Syst Biol.* 2013;62(4):539–54. Available from: <http://sysbio.oxfordjournals.org/cgi/doi/10.1093/sysbio/syt018>.
35. Mammal Species of the World: A Taxonomic and Geographic Reference (3rd ed). 2005. Wilson D, Reeder D, editors. Johns Hopkins University Press.
36. Madhani HD, Leadon SA, Smith CA, Hanawalt PC. a DNA in African green monkey cells is organized into extremely long tandem arrays. *J Biol Chem.* 1986;261:2314–8.
37. Fittler F. Analysis of the a-Satellite DNA from African Green Monkey Cells by Restriction Nucleases. *Eur J Biochem.* 1977;352:343–52.
38. Harrison JS, International C, Medicales DR. A new species of guenon (genus *Cercopithecus*) from Gabon. *J Zool.* 1984;198:561–75.
39. Lee TN, Singer MF. Structural organization of alpha-satellite DNA in a single monkey chromosome. *J Mol Biol.* 1982;161:323–42.
40. Rosandić M, Paar V, Basar I, Glunčić M, Pavin N, Pilaš I. CENP-B box and pJa sequence distribution in human alpha satellite higher-order repeats (HOR). *Chromosom Res.* 2006;14:735–53.
41. Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW. Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data. *PLoS Comput Biol.* 2013;9(4). Available from: <http://dx.doi.org/10.1371/journal.pcbi.1003031>.
42. O'Keefe CL, Matera AG. Alpha satellite DNA variant-specific oligoprobes differing by a single base can distinguish chromosome 15 homologs. *Genome Res.* 2000;10:1342–50.
43. Silahtaroglu A, Pfundheller H, Koskkin A, Tommerup N. LNA-modified oligonucleotides are highly efficient as FISH probes. *Cytogenet Genome Res.* 2004;37:32–7.
44. Ollion J, Loll F, Cochenne J, Boudier T, Escudé C. Cell cycle-dependent positioning of individual centromeres in the interphase nucleus of human lymphoblastoid cell lines. *Mol Biol Cell.* 2015;26(13):2550–60.
45. Rosenberg H, Singer M, Rosenberg M. Highly reiterated sequences of SIMIANSIMIANSIMIANSIMIANSIMIAN. *Science.* 1978;200(April):394–402.
46. Rojo V, Martínez-Lage A, Giovannotti M, González-Tizón AM, Cerioni PN, Barucchi VC, et al. Evolutionary dynamics of two satellite DNA families in rock lizards of the genus *Iberolacerta* (Squamata, Lacertidae): different histories but common traits. *Chromosom Res.* 2015;23(3):441–61.
47. Ruiz-ruano FJ, López-león MD, Cabrero J, Camacho JPM. High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Sci Rep.* 2016; 6. Available from: <http://dx.doi.org/10.1038/srep28333>.
48. Sujivattanarat P, Thapana W, Srikulnath K, Hirai Y, Hirai H, Koga A. Higher-order repeat structure in alpha satellite DNA occurs in New World monkeys and is not confined to hominoids. *Sci Rep.* 2015;5:10315. Available from: <http://www.nature.com/doifinder/10.1038/srep10315>.
49. Goldberg IG, Sawhney H, Pluta AF, Warburton PE, Earnshaw WC. Surprising deficiency of CENP-B binding sites in African green monkey alpha-satellite DNA: implications for CENP-B function at centromeres. *Mol Cell Biol.* 1996;16(9):5156–68.
50. Yoda K, Nakamura T, Masumoto H, Suzuki N, Kitagawa K, Nakano M, et al. Centromere Protein B of African Green Monkey Cells : Gene Structure, Cellular Expression, and Centromeric Localization. *Mol Cell Biol.* 1996;16(9):5169–77.
51. Smith GP. Evolution of repeated DNA sequences by unequal crossover. *Science.* 1976;191:528–35.
52. Henikoff S. Near the edge of a chromosome's "black hole". *Trends Genet.* 2002;18(4):165–7.
53. Henikoff JG, Thakur J, Kasinathan S, Henikoff S. A unique chromatin complex occupies young a-satellite arrays of human centromeres. *Sci Adv.* 2015;1:e1400234.
54. Guenatri M, Bailly D, Maison C, Almouzni G. Mouse centric and pericentric satellite repeats form distinct functional heterochromatin. *JCB.* 2004;166(4):493–505.
55. Choo KH, Earle E, McQuillan C. A homologous subfamily of satellite III DNA on human chromosomes 14 and 22. *Nucleic Acids Res.* 1990;18(19):5641–8.
56. Warburton PE, Hasson D, Guillem F, Lescale C, Jin X, Abrusan G. Analysis of the largest tandemly repeated DNA families in the human genome. *BMC Genomics.* 2008;9:533.
57. Pertile MD, Graham AN, Choo KHA, Kalitsis P. Rapid evolution of mouse Y centromere repeat DNA belies recent sequence stability. *Genome Res.* 2009;19(12):2202–13.
58. Mravinač B, Plohl M. Parallelism in evolution of highly repetitive DNAs in sibling species. *Mol Biol Evol.* 2010;27(8):1857–67.
59. Moulin S, Gerbault-Seureau M, Dutrillaux B, Richard FA. Phylogenomics of African guenons. *Chromosom Res.* 2008;16(5):783–99.
60. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. *J Mol Biol.* 1990;215:403–10.
61. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol.* 1970; 48(3):443–53. Available from: <http://www.sciencedirect.com/science/article/pii/0022283670900574>.
62. Ward JH. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc.* 1963;58(301):236–44.
63. Edgar RC. MUSCLE : a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform.* 2004;19:1–19.
64. Gouy M, Guindon S, Gascuel O. SeaView Version 4 : A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. *Mol Biol Evol.* 2010;27(2):221–4.
65. Rice P. The European Molecular Biology Open Software Suite EMBOSS : The European Molecular Biology Open Software Suite. *Trends Genet.* 2000;16(6):2–3.
66. R Core Team. R a Language and Environment for Statistical Computing. <https://www.R-project.org>.
67. Abràmoff MD, Hospitals I, Magalhães PJ, Abràmoff M. Image Processing with ImageJ. *J Biophotonics.* 2004;11(7):36–42.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit



2.2 Diversité et distribution de l'ADN alpha satellite chez *Cercopithecus pogonias*

Deuxième article - Copious alpha satellite DNA diversification events in the *Cercopithecus pogonias* lineage

Lauriane Cacheux^{1,2}, Loïc Ponger¹, Michèle Gerbault-Seureau², François Loll¹, Delphine Gey³, Florence Anne Richard^{2,4} and Christophe Escudé¹

¹Structure et Instabilité des Génomes, U1154 INSERM, UMR7196 CNRS, Sorbonne Universités, Muséum national d'Histoire naturelle, Département Régulations, Développement et Diversité Moléculaire, Paris, France

²Institut de Systématique, Evolution, Biodiversité, UMR7205 CNRS, UPMC, EPHE, Sorbonne Universités, Muséum national d'Histoire naturelle, Département Systématique et Evolution, Paris, France

³Service de Systématique moléculaire, UMS2700 CNRS, Sorbonne Universités, Muséum national d'Histoire naturelle, Paris, France

⁴Université Versailles St-Quentin, Montigny-le-Bretonneux, France

2.2.1 Abstract

Alpha satellite DNA is the main family of tandemly repeated sequences lying in primate centromere regions. Alpha satellite monomers (≈ 170 bp) diversified during evolution, forming well-known families of alpha satellite sequences within great ape genomes. These monomers arrange in homogeneous higher order repeats (HORs) in apes and New World monkeys. A recently elaborated framework revealed the presence of multiple families in the alpha satellite component of an Old World monkey from the Cercopithecini tribe, *Cercopithecus solatus*.

We now report its application on *C. pogonias*, a Cercopithecini species belonging to a different primary lineage and displaying a highly fissioned karyotype ($2n = 72$). XmnI- and HindIII-enzymatic isolations, high-throughput sequencing and computational analysis led to the identification of the C1 to C4 alpha satellite families, previously described in *C. solatus*. Fluorescence in situ hybridizations (FISH) showed that they adopt a rather similar chromosomal distribution in both species. Two additional families, named C5 and C6, were characterized and their distribution shown to be restricted to distinct sets of *C. pogonias* chromosomes. When analyzed

by principal component analysis, C1, C5 and C6 sequences displayed intriguing clusters that were deciphered as signatures for amplification events. Finally, both assembly of XmnI/HindIII sequences and FISH provided evidence for homogeneous HORs in *C. pogonias* centromeres.

This work highlights interspecific diversity of alpha satellite DNA in Cercopithecini, in terms of sequence, organization, and amplification dynamics. Our results suggest a correlation between alpha satellite diversity and the abundance of chromosomes and centromeres in Cercopithecini genomes.

Key words : Alpha satellite DNA, Sequence families, Cercopithecini, Interspecific diversity, Higher order repeats, Centromere evolution

2.2.2 Introduction

Millions of copies of repeated DNA sequences amplified within eukaryotic genomes and display organizational and regulatory functions (Volpe et al., 2002; Bejerano et al., 2006; Lunyak et al., 2007; Sienski et al., 2012; Glauser, 2014; Courac et al., 2016). Most centromere regions rely on large tracts of satellite, tandemly repeated sequences that may take part in the achievement of chromatid cohesion and chromosome segregation during mitosis and meiosis (Wong et al., 2007; Ferri et al., 2009; Chan et al., 2012; Fukagawa and Earnshaw, 2014). In Primates, a particular family, called alpha satellite DNA, spans centromeres and pericentromeres and consists in about 170 bp monomers that display 60 to 100 % sequence identity between them (Rudd and Willard, 2004).

Distinct families of alpha satellite sequences, with specific nucleotide polymorphisms, have been described throughout the Primate order and mostly in great apes genomes (Alexandrov et al., 1993; Alves et al., 1994, 1998; Romanova et al., 1996; Alexandrov et al., 2001; Shepelev et al., 2009; Hayden, 2012; Prakhongcheep et al., 2013; Catacchio et al., 2015). According to the age-gradient hypothesis of Schueler et al. (2001, 2005), alpha satellite families emerge and expand at core centromeres, thereby splitting and displacing older families distally onto each chromosome arm. Consistent with this hypothesis, families spanning human pericentromeres were observed to be heterogeneous (70 to 85 % mean sequence identity) and shared by distant species, whereas human core centromere families would be more homogeneous and restricted to great apes (Baldini et al., 1991; Haaf and Willard, 1997, 1998; Shepelev et al., 2009). By contrast with apes, information gathered on alpha satellite families is limited in monkeys, which is a constraint to a complete comprehension of alpha satellite diversification dynamics in Primates. Nevertheless, a promising framework allowing fast and deep analysis of alpha satellite DNA was recently applied to Cercopithecini (family Cercopithecidae, Old World monkeys) and revealed a number of alpha satellite families that probably formed independently from the ape lineage (Cacheux et al., 2016b).

Alpha satellite families that are conserved between great ape species distribute generally at centromeres of non-homologous chromosomes (Willard and Waye, 1987; Warburton and Willard, 1995). Concerted evolution mechanisms, such as unequal crossing over or sister chromatids exchange, transposition, rolling circle amplification and sequence conversion, would be responsible for such a fast divergence of tandem repeats between homologous loci (Willard, 1991; Rudd et al., 2006; Plohl et al., 2012). This so-called homogenization is seen as mostly intrachromosomal in apes, which display chromosome-specific sets of sequences (Willard and Waye, 1987; Haaf and Willard, 1997), but interchromosomal in monkeys where very similar sequences are usually found on most chromosomes (Alkan et al., 2007; Prakhongcheep et al., 2013; Cacheux et al., 2016b).

Alpha satellite DNA displays two types of organization in primate genomes, which are called monomeric and higher order repeat (HOR) organizations (Schueler and Sullivan, 2006). A monomeric organization is the result of single monomer amplification, and displays arrays of adjacent monomers belonging to the same family. A HOR organization is the outcome of several adjacent monomers amplified together in the same repeat unit and possibly belonging to different families. Complex HORs with repeat units comprising up to more than 30 monomers were characterized at centromeres of apes and New World monkeys (Hayden, 2012; Terada et al., 2013; Koga et al., 2014; Catacchio et al., 2015; Sujiwattanarat et al., 2015). They are generally highly homogeneous, with repeat units from the same array displaying 98 % mean sequence identity (Rudd and Willard, 2004). By contrast, only heterogeneous HORs spanning pericentromeres have been certainly identified in Old World monkeys so far, and the establishment of homogeneous ones at their centromeres is still questioned (Cacheux et al., 2016b).

The *Cercopithecus solatus* and *C. pogonias* species belong to different primary lineages within the Cercopithecini tribe and are separated by 7 million years of evolution (Tosi, 2008; Guschanski et al., 2013). A high number of evolutionary chromosomal fissions occurred specifically along the *C. pogonias* lineage, resulting in the formation of 72 stable chromosomes in its genome versus 60 for *C. solatus* (Dutrillaux et al., 1979, 1988a; Muleris et al., 1984). The alpha satellite component of the *C. solatus* genome was previously analyzed appealing to high-throughput sequencing, bioinformatics and cytogenetics (Cacheux et al., 2016b). Four alpha satellite families, named C1 to C4, were characterized. To pursue the investigation of alpha satellite sequences and organizations in Old World monkeys, we improved this framework on *C. pogonias*. Results were compared to *C. solatus* in order to appreciate alpha satellite evolutionary dynamics in Cercopithecini and investigate the relationship between chromosome/centromere richness and alpha satellite diversity.

2.2.3 Results

Diversity of alpha satellite families in *C. pogonias* monomer and dimer datasets

Alpha satellite monomers and dimers were isolated from the *C. pogonias* genome using the restriction enzyme XmnI, as previously described for *C. solatus* (Cacheux et al., 2016b)(see Methods). Both samples were submitted to high-throughput sequencing on an Ion Torrent sequencing platform ; 210,527 and 273,427 raw sequences were obtained for the monomer and dimer samples, respectively. In silico reorientation and filters were applied to both, and a total of 112,575 alpha satellite sequences were retrieved from the monomer sample (monomer dataset), while 2,433 alpha satellite sequences were retrieved from the dimer sample (dimer dataset)(see Methods).

A principal component analysis (PCA) using the 5-mer nucleotide composition of DNA sequences was applied to *C. pogonias* monomer dataset. Visualization of sequences into the plane formed by the two first components of the PCA revealed a structure that seemed more complex than the one obtained for *C. solatus*, where two main alpha satellite families had been identified (Figure 2.1A). The prediction of *C. solatus* monomer dataset by using the PCA projection of *C. pogonias* monomers (Figure 2.1B) allowed us to hypothesized that the two previously identified C1 (purple) and C2 (pastel green) alpha satellite families coexist in the *C. pogonias* genome with two new families, shown in red and orange on Figure 2.1C and called from now on C5 and C6, respectively. After having assigned all sequences to a family by using a combination of hierarchical clustering analysis (HCA) and linear discriminant analysis (LDA), we confirmed, using phylogenetic trees, that the C1 and C2 families were present in *C. pogonias* and that the C5 and C6 families were distinct (Supplementary figure 2.1). Respective proportion and sequence identity for each family are reported on Table 2.1, and alignment between their consensus sequences is shown on Figure 2.2. The consensus of C5 and C6 differed from that of C1 by two and three single nucleotide variations, respectively. Most C1, C2, C5 and C6 sequences possessed a pJalpha box (Figure 2.2 and Table 2.1).

Dimeric alpha satellite sequences were analyzed separately according to the presence (X dimer dataset ; 65 sequences provided by partial enzymatic digestion) or absence (noX dimer dataset ; 2295 sequences provided by complete enzymatic digestion) of an XmnI restriction site between their component left and right monomers. Three distinct families were identified in *C. pogonias* noX dimer dataset using a PCA/HCA approach : they matched to the families previously identified in *C. solatus* noX dimer dataset, meaning the C2, C3 and C4 families (see Supplementary figures 2.2 and 2.3). C2-C2 and C3-C4 associations were observed to be significant (see Supplementary table 2.1), which suggests a monomeric organization for the C2 family and a HOR organization for the C3 and C4 families. According to their limited number, monomers from *C. pogonias* X dimer dataset were assigned to one of the C1 to C6 alpha satellite family using a LDA trained to recognize these families. These dimeric sequences only displayed significant

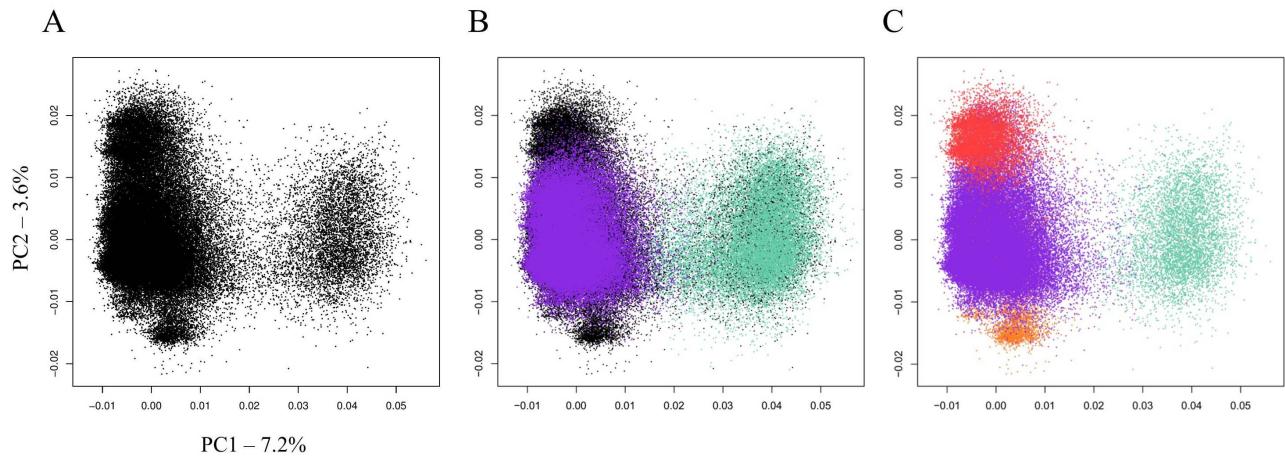


FIGURE 2.1 – Characterization of alpha satellite DNA diversity in *C. pogonias* monomer dataset. (A) PCA projection on principal components 1 and 2 of the normalized 5-mer frequency vectors for all sequences from *C. pogonias* monomer dataset. Each point represents a sequence. (B) Prediction of the C1 (purple) and C2 (pastel green) sequences from *C. solatus* monomer dataset by using the PCA projection of *C. pogonias* monomers. (C) PCA projection of *C. pogonias* monomer dataset with sequences colored according to their assignment to the C1 (purple), C2 (pastel green), C5 (red) or C6 (orange) alpha satellite family, based on hierarchical classification method.

TABLE 2.1 – Sequence features of the C1 to C6 alpha satellite families in *C. pogonias* monomer and dimer datasets

Monomer dataset (%)	Dimer dataset (%)		Mean sequence identity (%)	CENP-B (%)	pJalpha (%)
	Left monomers	Right monomers			
C1	84	0	0	95	0.04
C2	5	93	93	85	0.09
C3	0	7	0	87	0
C4	0	0	7	85	0
C5	7	0	0	95	0.02
C6	4	0	0	98	0.05

NOTE - The percentages of each family in the monomer dataset and in the noX dimer dataset (dimers provided by complete digestion) are displayed. Sequence identities were calculated using a subset of 500 randomly selected sequences within each C1, C2, C5 or C6 monomers, and using the available 165 and 162 sequences for C3 and C4, respectively. The percentages of sequences displaying the CENP-B or pJalpha fixation site were calculated using all the C1, C2, C5 or C6 sequences from the monomer dataset and all the C3 and C4 sequences from the noX dimer dataset.

1

C1	GCTTCTTGAAGGGAAAGATGTAACCTCTGTGAGATGAATTACAGAACACAGAGCAGTTCTCAGAAAGCTTCTTCAGTTTGAA
C2G.....C.....A.....T..
C3G.....C.....C.....T.....A.....TCAN.....T.N
C4	...CNAN...-AT.C..C.....A.....TG...G.C...N.T....A.....N.....T.N
C5N.....C.G.....
C6

87

C1	CGGAAGATATTCCCTTTTCACCATAGCCCTCTATGGGCTTCAAATATCCCTTGCCAATTCCACAAGAACAGCCTTAGCGAAAG
C2	.N.....A.....A.....T.....
C3G.....GG.C.T.....T.....A.A.....A.C.G.....N.AG..CT.AG.....GA..T.AGGC.....A..CA
C4	.T..N.....NNN.....A.....A.....T.....N.
C5
C6A.....AA.....

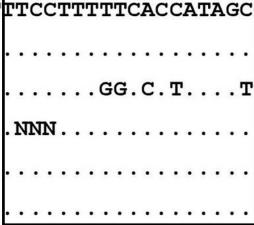


FIGURE 2.2 – Consensus sequences of the C1 to C6 alpha satellite families identified in *C. pogonias* monomer and dimer datasets. The consensus sequences were determined following the alignment of 500 randomly selected sequences within the C1, C2, C5 and C6 monomers, and the alignment of the available 165 and 162 sequences for C3 and C4, respectively. Each position was considered unambiguous if more than 60 % of monomers had the same nucleotide at this position. A point at a position replaces a nucleotide identical to the nucleotide at the homologous position in the C1 consensus. The box shows the fixation site of the pJalpha protein, which is absent from the C3 consensus.

C1-C1 associations (see Supplementary table 2.1), probably because of their limited number. C3 sequences do not possess either pJalpha or CENP-B box in *C. pogonias*, whereas most of C4 sequences possess a pJalpha box (Figure 2.2 and Table 2.1).

Distribution of alpha satellite families on *C. pogonias* chromosomes

Short locked nucleic acid (LNA) modified probes, previously designed to pursue the FISH of the C1 to C4 families on *C. solatus* chromosomes, were used here on *C. pogonias*; however, in silico research revealed that any probe C1 should also target the C5 or C6 family, due to the high sequence similarity between these three families (see Figure 2.2). In fact, our probe C1a targets both C1 and C5 families, and our probe C1b targets both C1 and C6 families. To solve this problem, we designed probes which should target specifically the C5 or C6 family, so we can compare their signals with those of C1a and C1b (see Supplementary figure 2.4 for probe patterns and targets).

Probes C1 produced intense signals at the centromere (primary constriction) of all chromosomes (Figure 2.3A,B,D), except for the Y chromosome and an autosome identified as chromosome 6; both chromosomes even remained unlabeled by probe Cx that targets a conserved site within the C1 to C6 families (see arrows on Figure 2.3A and Supplementary figures 2.4 and 2.5). Thus, all centromeres but two would be colonized here by the C1, and/or C5, and/or C6 families. Probes C2 provided signals that were located at the pericentromeres (around the primary constriction) of most chromosomes (Figure 2.3A,B and Supplementary figure 2.6). Probes C4 provided slight signals at the pericentromeres of several chromosomes (Supplementary figure 2.7), which was in accordance with their hybridization pattern on *C. solatus* chromosomes. By contrast, probe C3a strongly labelled centromeres on all acrocentric chromosomes, whereas probe C3b provided no detectable signal. As we know from our sequence analysis that C3 and C4 sequences are associated in the *C. pogonias* genome, probe C3a probably does not reveal only C3 sequences, and should target an alpha satellite family that we did not gathered using our enzymatic digestion approach. On the contrary, C3b may target actual C3 sequences; however, the C3 family would not display enough repeats in *C. pogonias* pericentromeres to be detected in this experiment.

Probe C5a produced intense signals on 11 chromosomes, i.e. 5 pairs of autosomes and the X chromosome (Figure 2.3C,D and Supplementary figure 2.8). Interestingly, the C5a label was pericentromeric on several chromosome pairs, compared to the C1a label (see arrows on Figure 2.3D). Probe C6a produced intense signals at the centromere of all 12 pairs of acrocentrics, slightly lighter signals at the centromere of one pair of submetacentrics, and weak signals at the centromere of four other meta- or submetacentrics (Figure 2.3C and Supplementary figure 2.9). We noticed that probes C5a and C6a produced signals on different chromosomes but two (see arrows on Figure 2.3C).

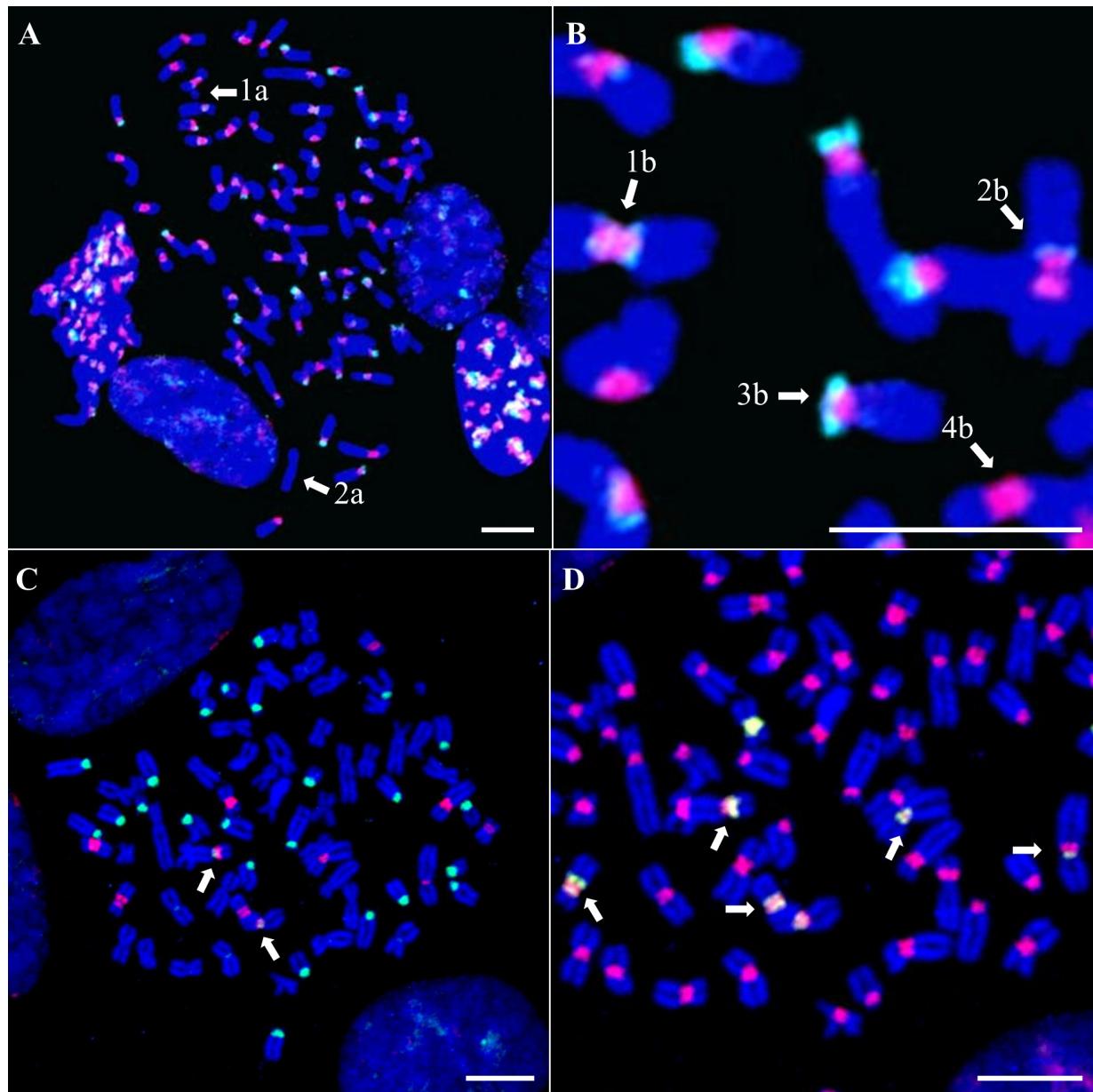


FIGURE 2.3 – FISH analysis of the C1, C2, C5 and C6 alpha satellite families on *C. pogonias* chromosomes. *Cercopithecus pogonias* metaphase chromosomes are colored in blue. (A,B) Probes C1a, C1b and C2b are hybridized simultaneously. (A) Hybridization of probes C1a and C1b (red) and probe C2b (green). 1a and 2a : unlabeled chromosomes (the Y chromosome and one autosome, respectively; see Supplementary figure 2.5). (B) Focus on image (A) showing in details the different types of distribution of the C2b signals. 1b : both pericentromeric regions, 2b : one pericentromeric region toward the long arm, 3b : one pericentromeric region toward the short arm of an acrocentric chromosome, 4b : no signal. (C, D) Probes C1a, C5a and C6a are hybridized simultaneously. (C) Hybridization of probe C5a (red) and probe C6a (green). Arrows : two chromosomes where both probes produce signals. (D) Focus on the metaphase shown in (C) but with hybridization of probe C1a (red) and probe C5a (green). Arrows : pericentromeric hybridization of probe C5a on several chromosomes. Scale bar = 10 µm.

Finally, probes C5a and C6a were also used by FISH on *C. solatus* chromosomes as a negative control. Probe C6a did not provide any signal and probe C5a provided signals at the centromere of eight chromosomes ; however, their intensity decreased substantially when probes were washed at 68 °C after the hybridization step, instead of the usual 63 °C (Supplementary figure 2.10). By contrast, signals provided by probe C5a on *C. pogonias* chromosomes were resistant to such temperature. This phenomenon likely reveals a nonspecific hybridization of probe C5a to *C. solatus* chromosomes, suggesting that, as expected, both C5 and C6 families are absent from the *C. solatus* genome.

Amplification event signatures in *C. pogonias* monomer dataset

Some sequences were previously noticed to be strictly repeated in high copy number in *C. solatus* monomer dataset ; we searched for such repeated sequences in *C. pogonias* monomer dataset. Indeed, abundant repeated sequences were found within the C1, C5 and C6 families ; we decided to investigate further the 20 most abundant repeats (Table 2.2). The most abundant sequence 1 (2983 repeats) matched to the consensus of the C1 family, as observed for *C. solatus*, while sequences 5 (455 repeats) matched to the consensus of C6. The consensus of C5 displays an ambiguous G or T nucleotide at position 28 (see Figure 2.2) ; sequence 11 (116 repeats) matched to the C5 consensus with a 28G. All other repeated sequences were similar to the C1 or C6 consensus but with single nucleotide variations and/or deletions. Among them, five sequences (4, 6, 9, 17 and 18) were likely to correspond to artefacts as they displayed deletions within homopolymer tracts and were associated to strong biases for read orientation. All other repeated sequences were obtained in both sequencing orientations and may correspond to sequences that are present with a high abundance in the *C. pogonias* genome. Some of them still displayed a slight bias for read orientation (see for example sequences 10, 13, 16 and 19) ; their strands may have been sequenced with differential efficiency by the Ion torrent technology. Except for sequences 1 and 14, all the repeated sequences found in *C. pogonias* were distinct from those previously observed in *C. solatus*. By contrast, the two sequence artifacts 4 and 6 were also observed in *C. solatus* monomer dataset.

When the PCA of *C. pogonias* monomer dataset was projected with high point transparency, unexpected comet-like clusters were observed within the C1, C5 and C6 families (Figure 2.4). Interestingly, the repeated sequences 1, 5 and 11, matching to the C1, C6 and C5 (28G) consensus, were found to be positioned at the left extremity of such comet-like clusters (see yellow, red and green spots on Figure 2.4A, respectively). Two comet-like clusters were easily visualized within the C5 family ; a search for a second repeated C5 sequence that could position on the second cluster lead us to identify a sequence matching to the C5 consensus with a 28T (53 repeats)(see blue spot on Figure 2.4A). We colored then, on the PCA projection, all C5 sequences displaying a 28G or a 28T (see green and blue sequences on Figure 2.4B, respectively), and observed that both groups were differentially revealed all along the comet-like clusters. Same observations were made about the position of the repeated C1 sequences 2, 3 and 19 (see

TABLE 2.2 – Analysis of alpha satellite sequences found in high copy number in *C. pogonias* monomer dataset

Id	Sequence	Number	Forward (%)
1	Consensus C1	2983	46
2	C158G	848	48
3	C116T	568	41
4	C114Del	508	1*
5	Consensus C6	455	34
6	T101Del	323	98*
7	C2A-G17Del	250	66
8	C2A-G17Del-C158G	208	70
9	C158G-C114Del	145	0*
10	A3741T-G104A-C158G	136	15
11	Consensus C5 (28G)	116	44
12	C116T-C158G	112	46
13	C2A	103	73
14	T121A	100	43
15	C137A-C158G	100	51
16	A3741T-G104A	100	24
17	Consensus C6-C114Del	89	1*
18	C2A-G17Del-C114Del	81	1*
19	T38G	77	29
20	A110G	76	56

NOTE - The sequences are numbered according to the “Id” column. The “Sequence” column indicates how each sequence differs from the consensus sequence of the C1 family, using standard notations, except when they are identical to the consensus sequence of the C5 or C6 family (Consensus C5 or Consensus C6) or when they differ from the consensus sequence of the C6 family (Consensus C6-C114Del). The “Number” column displays the number of strict copies of the sequence in the monomer dataset. The “Forward” column displays the percentage of reads obtained in the forward orientation (i.e. the orientation of our reference sequence). Strong biases for read orientation are shown with an asterix (*).

blue, green and red spots on Figure 2.4C, respectively) and all C1 sequences displaying the C158G of sequence 2, the C116T of sequence 3 or the T38G of sequence 19 (see blue, green and red sequences on Figure 2.4D,E, respectively). Thus, comet-like clusters are likely to be the signatures of amplification events of one particular sequence, followed by the divergence of part of the generated copies. Such clusters were not apparent on the PCA projection of *C. solatus* monomer dataset; however, some clusters appeared in this dataset when predicted using the PCA projection of *C. pogonias* monomers, and some repeated C1 sequences of *C. solatus* were also found to be positioned at the left extremity of these clusters (Supplementary figure 2.11).

Finally, we designed different LNA probes to compete and achieve the required level of specificity to target by FISH the repeated sequence 2 on its C158G variation (Figure 2.5). Probe C158G did not hybridize on all chromosomes and provided particularly strong signals at the centromere of all 12 pairs of acrocentrics, suggesting peculiar homogenization and amplification of this sequence on and between those chromosomes. This probe also provided slighter signals at the very core centromere of a few other chromosomes (see arrows on Figure 2.5B), where lower repeated sequences displaying a C158G might be revealed, like sequences 10, 12 or 15.

Organization of alpha satellite families in the *C. pogonias* genome

In order to obtain overlapping alpha satellite sequences and conduct further studies on the relative organization of alpha satellite families in the *C. pogonias* genome, we isolated alpha satellite monomers using the HindIII enzyme, which also results in a ladder pattern, and performed high-throughput sequencing on the same Ion torrent platform as above. A total of 84,485 alpha satellite monomers were recovered and analyzed with a similar approach as described above. Four distinct groups of HindIII sequences were visualized by PCA and compared to the C1 to C6 families by prediction using PCA (Figure 2.6A,B). The good overlap that was observed between three of those groups and the C1, C2 and C5 families led us to call them C1', C2' and C5' (shown in purple, pastel green and red on Figure 2.6C, respectively). The strict identity of the consensus sequences between C1 and C1', C2 and C2', and C5 and C5', except for a phase shift, suggests that these monomers are the results of regularly spaced digestion of tandemly repeated sequences from the same family by the two different enzymes (Supplementary figure 2.12). Thus, the C1, C2 and C5 families would display a monomeric organization within the *C. pogonias* genome.

The fourth group of HindIII sequences was of similar size as C6, and they both occupied a close but slightly shifted position with respect to each other on the PCA projection (Figure 2.6B); this group was called C6' (shown in blue on Figure 2.6C). Comparison of the consensus sequences of C6 and C6' showed that they were identical in their overlapping 106 bp fragment but, unlike C1 and C1', C2 and C2' or C5 and C5', their non-overlapping 66 bp fragments were not : they differed by a substitution and an indel (Supplementary figure 2.12). The polymorphisms observed in the C6' consensus were not found in any of the C1 to C5 consensus

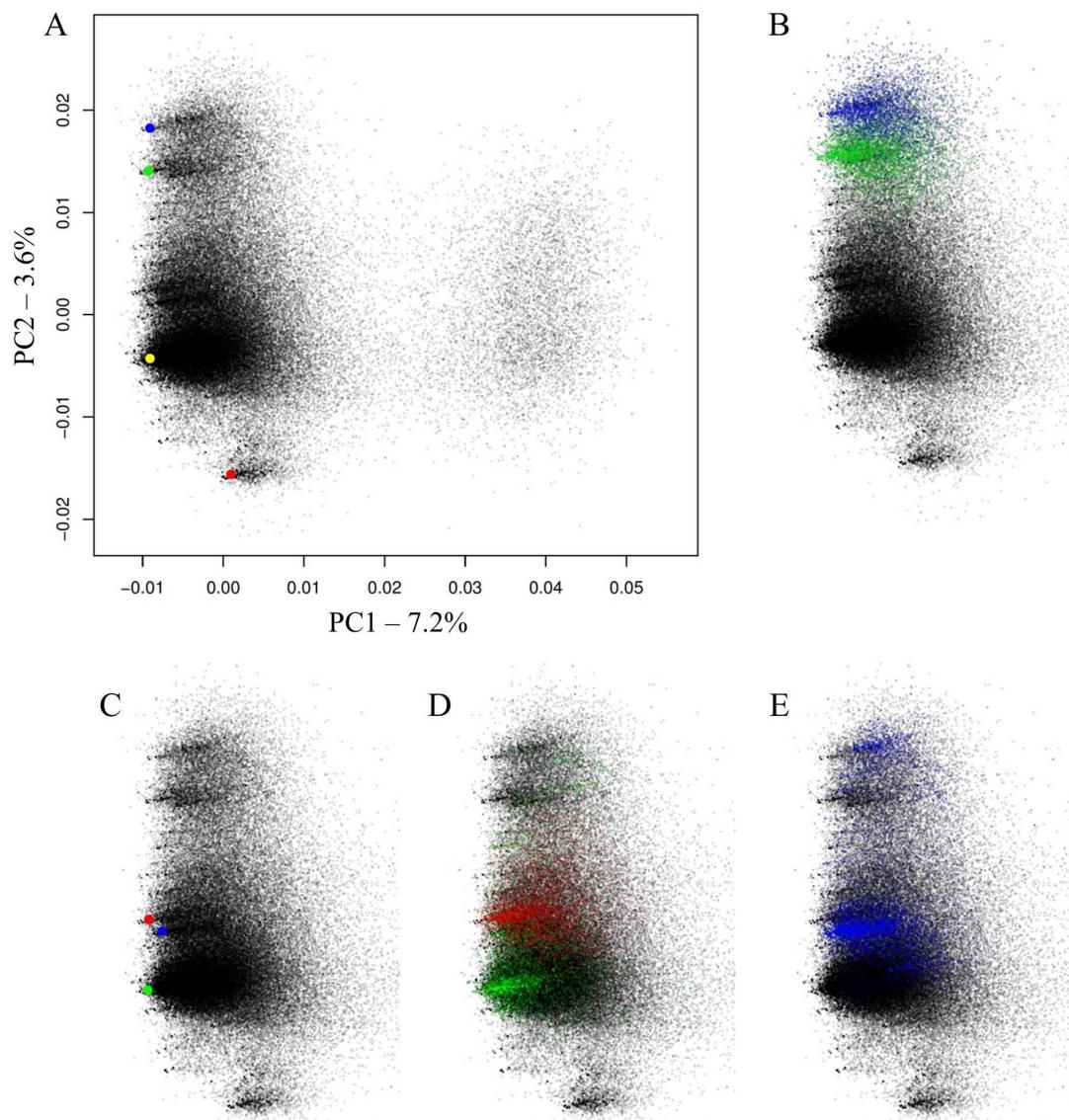


FIGURE 2.4 – Characterization of comet-like alpha satellite clusters on the PCA projection of *C. pogonias* monomer dataset. PCA projection on principal components 1 and 2 of the normalized 5-mer frequency vectors for all sequences is represented in all subfigures; only the left part containing the C1, C5 and C6 families has been kept for subfigures (B-E). (A) Repeated sequences matching to C1 consensus, C5 consensus (28G), C5 consensus (28T) and C6 consensus are spotted in yellow, green, blue or red, respectively. (B) All C5 sequences displaying a 28G or a 28T are colored in green or blue, respectively. (C) Repeated sequences 2, 3 and 19 are spotted in blue, green or red, respectively. (D,E) All C1 sequences displaying the C158G of sequence 2, the C116T of sequence 3 or the T38G of sequence 19 are colored in blue, green or red, respectively.

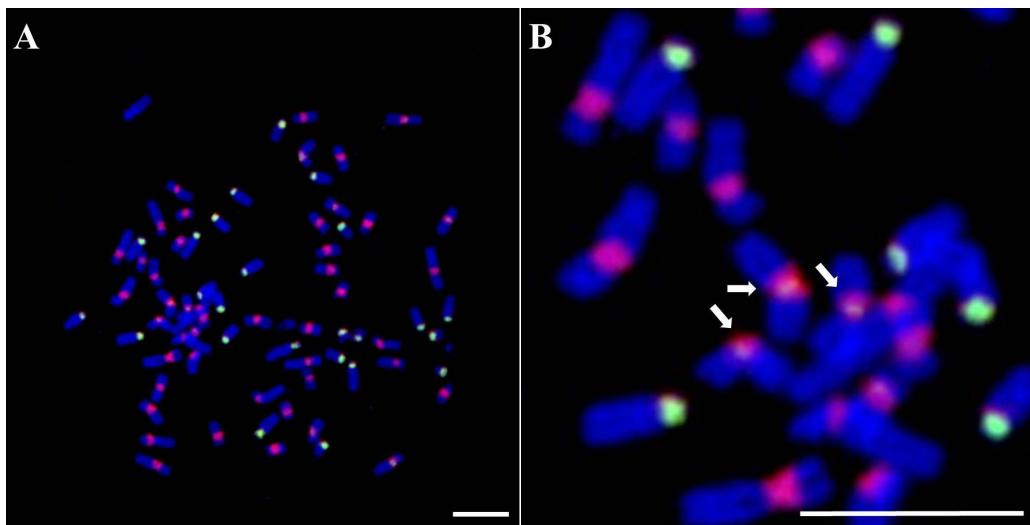


FIGURE 2.5 – FISH analysis of the high copy numbered alpha satellite sequence 2 (C158G) on *C. pogonias* chromosomes. Probe C158G and the competitor probe C158co, which prevents it to hybridize on the global C1 family, are hybridized simultaneously to *C. pogonias* metaphase chromosomes, which are colored in blue. (A) Hybridization of probe C158G (green) and probe C158co (red). (B) Focus on image (A) showing in details the distribution pattern of probe C158G with strong signals at the centromere of acrocentrics and slighter signals (arrows) at the core centromere of several other chromosomes. Scale bar = 10 µm.

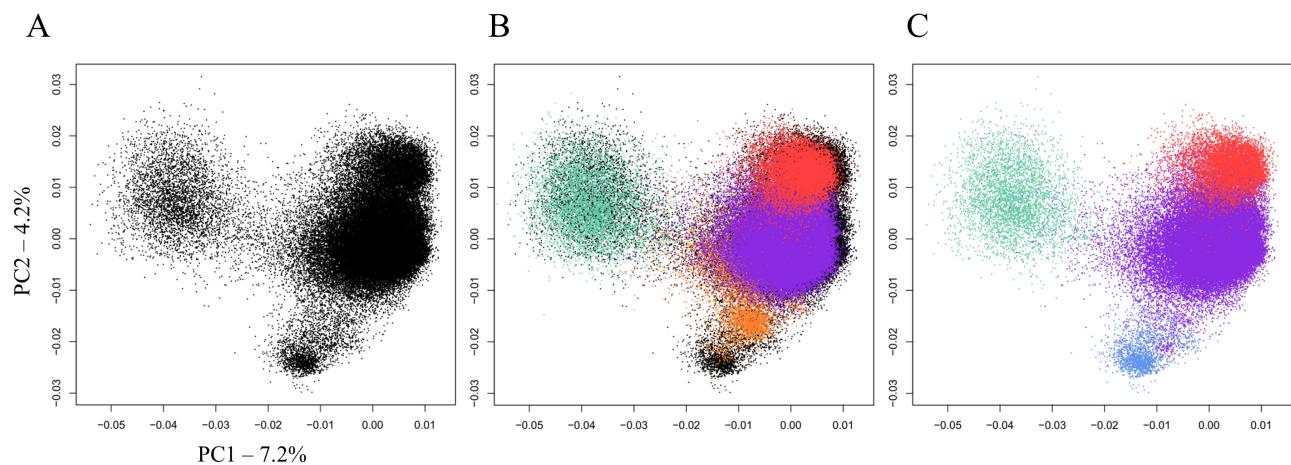


FIGURE 2.6 – Comparison of alpha satellite diversity between *C. pogonias* XmnI and HindIII monomers. (A) PCA projection of the normalized 5-mer frequency vectors for all HindIII monomers. Each point represents a sequence. (B) Prediction of XmnI monomers by using HindIII monomer PCA projection. XmnI sequences are colored according to their assignment to the C1 (purple), C2 (pastel green), C5 (red) or C6 (orange) alpha satellite family. (C) HindIII monomer PCA projection, with sequences colored according to their similarity to the C1 (C1', purple), C2 (C2', pastel green), or C5 (C5', red) family. One group of HindIII sequences is close to the C6 family but slightly shifted; it is colored in blue (C6').

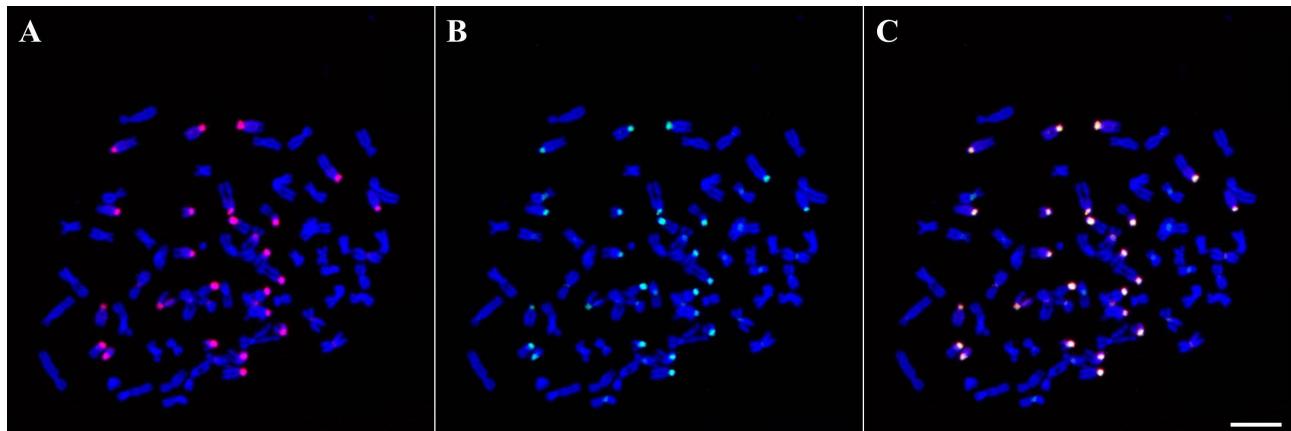


FIGURE 2.7 – FISH analysis of the C6' HindIII sequences relatively to the C6 alpha satellite family. Probes C6a and C6'a are hybridized simultaneously to *C. pogonias* metaphase chromosomes, which are colored in blue. (A) Hybridization of probe C6a is shown in red. (B) Hybridization of probe C6'a is shown in green. (C) Combined signals from (A) and (B). Scale bar = 10 µm.

either. These observations suggest that C6 monomers are not adjacent and tandemly repeated in the *C. pogonias* genome; they may be instead involved in a HOR organization. In fact, the C6 family is likely to be associated to a new family whose consensus begins with the last 66 bp of the C6' consensus. The sequence constituting the second part of the consensus of this new family, as well as the global composition of the HOR unit, remains to be elucidated. Furthermore, FISH of both C6 and C6' sequences showed a colocalization of their respective signals at the centromere of all 12 pairs of acrocentrics and a few meta- or submetacentrics (Figure 2.7 and Supplementary figure 2.4), which is in accordance with an overlap between C6 and C6' sequences in the *C. pogonias* genome.

2.2.4 Discussion

Interspecific diversity of alpha satellite DNA in Cercopithecini

The concerted evolution of alpha satellite DNA predicts accumulation of species-specific sequence variations in short evolutionary periods (Dover, 1986). This phenomenon, typical of multigene family dynamics, would involve fast homogenization mechanisms driving considerable alpha satellite divergence even among closely related primate species (Waye and Willard, 1989; Rudd et al., 2006; Catacchio et al., 2015). The presence of the C5 and C6 families in the *C. pogonias* genome and their absence in *C. solatus* are consistent with this view, as they probably formed after the divergence of their respective lineages, i.e. in a few million years of evolution. This scenario is reinforced by the high homogeneity of these families (95 and 98 % mean sequence identity, respectively) and by the strong intensity of their FISH signals on res-

tricted subsets of *C. pogonias* chromosomes, which further reveals interchromosomal diversity of alpha satellite DNA in Old World monkeys.

By contrast, the C1, C2, C3 and C4 families are shared between species. C2, C3 and C4 are quite heterogeneous (85, 87 and 85 % mean sequence identity, respectively) and display a pericentromeric distribution in *C. solatus* as in *C. pogonias*, being consistent with the age-gradient hypothesis (Schueler et al., 2001, 2005; Shepelev et al., 2009). Contrary to what was observed in *C. solatus*, C3 and C4 do not distribute on the Y chromosome in *C. pogonias*; this feature may be a specificity of the *C. solatus* genome. The C1 family is more homogeneous (95 % mean sequence identity) and distributes at centromeres in both species, whereas one may have expected it to be also displaced toward pericentromeres, especially in the *C. pogonias* lineage wherein considerable alpha satellite diversification was noticed. This feature implies that alpha satellite families can be strongly conserved during the course of evolution, retaining high sequence homogeneity, which raises once again the question of potential functional constraints applied on alpha satellite DNA (Romanova et al., 1996; Plohl et al., 2008; Ugarković, 2009). Furthermore, the pericentromeric distribution of C5 compared to C1 on several chromosomes suggests an ongoing competition between these families; C1 sequences would have won on some centromeres and reintegrated core regions wherein C5 had established. In this view, the age-gradient hypothesis might be slightly reassessed, as old families could be localized in more proximal centromeric regions than new families. Nevertheless, this idea relies on two assumptions : C5 is indeed newer than C1, and our C1 probes did not reveal, at centromeres, some other families that we would have not gathered using our enzymatic digestion approach.

Distinct alpha satellite amplification events along Cercopithecini lineages

Our understanding of alpha satellite diversification dynamics suffers from the difficulty conferred to the study of tandem DNA repeats and from the subsequent scarce information available for most primates (Eichler et al., 2004; Rudd and Willard, 2004). Our new framework allows digging into Old World monkey centromeres, and the description of high copy numbered identical alpha satellite sequences within and between genomes is an additional step towards a comprehensive view of such dynamics. Indeed, repeated sequences matching to the C1 consensus were noticed in *C. solatus* as in *C. pogonias*: they should be some remains of the amplification event that led to the formation of the C1 family. Such features were also observed for the C5 and C6 families in *C. pogonias* and can be similarly interpreted. Additionally, these repeated sequences are associated, on monomer PCA projections, to comet-like clusters that may be interpreted as signatures for their initial amplification and subsequent accumulation of mutations.

Further amplification events can be inferred from secondary comet-like clusters of C1 sequences, where are positioned other repeated sequences displaying single nucleotide variations compared to the C1 consensus. Such repeated sequences, generally specific to *C. solatus* or *C. pogonias*, may be interpreted as the lineage-dependent remains of subfamily formations within the C1

family. In the same way, the repeated sequences matching the C5 consensus and displaying a 28G were likely amplified from a mutated copy of the C1 consensus, but two single nucleotide variations appear to be sufficient to discriminate C5 as a different family on PCA projections and phylogenetic trees. The repeated sequences matching the C5 consensus and displaying a 28T may have then amplified from a single-nucleotide mutated copy of the C5 (28G) consensus to form a subfamily within C5. On the other hand, no repeated sequences were detected within the C6 family except for the ones matching the C6 consensus ; the C6 family may have formed recently by the amplification of a three-nucleotide mutated copy of the C1 consensus, and did not undergo such diversification events yet.

By contrast, the C2 family does not display high copy numbered sequences nor comet-like clusters ; the remains of the initial amplification of this likely old family may have progressively mutated and disappeared from genomes. These observations also imply that no C2 sequence amplified recently ; the C2 family may have lost its capacity to homogenize.

HOR organizations at Cercopithecini centromeres

While homogeneous alpha satellite HORs have long been considered to be specific to hominoid centromeres (Rosenberg et al., 1978; Alves et al., 1994; Cellamare et al., 2009; Terada et al., 2013), recent discoveries provided the existence of such organizations in New World monkeys (Sujiwattanarat et al., 2015). By contrast, only heterogeneous pericentromeric HORs have been certainly identified in Old World monkeys and more precisely in Cercopithecini (Cacheux et al., 2016b). In the present work, the use of a high quantity of sequences provided in two different enzymatic phases, XmnI and HindIII, allowed reliable alpha satellite assembly attempts that provided evidence for the association of the homogeneous C6 family with at least one other family into HORs. Only part of the consensus sequence of this second family was obtained here, but was sufficient to target both families by FISH. As expected, their respective signals were seen to colocalize, and were revealed at centromeres of *C. pogonias* acrocentric chromosomes. Interestingly, FISH signals for the C1 family were also detected at acrocentric centromeres, which suggests that C1 sequences may be part of the concerned HORs. More precisely, acrocentric centromeres showed strong signals for a particular repeated C1 sequence (sequence 2, displaying a C158G) ; this sequence may have been included and amplified within these homogeneous HORs.

2.2.5 Conclusion

Alpha satellite DNA diversified in families of sequences during Cercopithecini evolution, some of them being shared between the *C. solatus* and *C. pogonias* species (C1, C2, C3 and C4 families) and others, possibly younger, being specific to *C. pogonias* (C5 and C6 families). The

shared C1 family displays several subfamilies of sequences that are still distinct between species and may have amplified after the divergence of the *C. solatus* and *C. pogonias* lineages, along with the C5 and C6 families. Alpha satellite families adopt distinct distribution patterns on *C. pogonias* chromosomes that are largely in accordance with their relative homogeneity and the age-gradient hypothesis. Furthermore, the C5 and C6 families are present on restricted subsets of chromosomes : the current view about an absence of alpha satellite interchromosomal diversity in Old World monkeys definitely needs to be modified. Finally, evidence for homogeneous HORs, including the C6 family, was provided at *C. pogonias* centromeres and allows extending the presence of such organizations to Old World monkey genomes.

2.2.6 Methods

DNA collection and metaphase preparations

Fibroblast cell samples of *Cercopithecus pogonias* (ID : 2001-027, male sample) from the Collection of cryopreserved living tissues and cells of vertebrates (RBCell collection, Muséum national d'Histoire naturelle, Paris) were used for DNA extraction. DNA was extracted using the Omega Bioteck Tissue DNA Kit. Fibroblast cell samples of this same specimen were used for metaphase preparations. Cell cultures and metaphase preparations were achieved according to [Moulin et al. \(2008\)](#).

Alpha satellite DNA isolation and sequencing

XmnI or HindIII were used to digest *C. pogonias* DNA in vitro. 10 µg of *C. pogonias* genomic DNA were digested for 6 h at 37 °C with 70 units of XmnI or HindIII activity (New England Biolabs) in a total volume of 35 µL. The enzymes were both inactivated for 20 min at 65 °C. Both samples were loaded on a 1 % agarose gel after addition of 7 µL loading buffer (50 % glycerol) and electrophoresis was performed in 0.5X Tris-borate-EDTA buffer, at room temperature for 3 h at 100 V. The gel was briefly stained with ethidium bromide and then imaged by UV transillumination. Bands corresponding to alpha satellite XmnI and HindIII monomers (\approx 170 bp) and XmnI dimers (\approx 340 bp) were cut and DNA was extracted from the gel with the Omega Bioteck Gel extraction kit and resuspended in 100 µl of elution buffer. About 250 ng were obtained for the 170 bp samples for both XmnI and HindIII digestions, and about 125 ng were obtained for the XmnI 340 bp sample.

Sequencing was performed on a PGM sequencing platform (Ion Torrent technology) using the 400 bp sequencing kit. The HindIII 170 bp sample was blunted according to the Quick Blunting Kit (E1201S, NEB). Three libraries were generated using 50 ng of the three blunt digest samples and the Ion Plus Fragment Library Kit (4471252, Life Technologies) and tagged

with Ion Xpress barcode adapters (4471250, Life Technologies). After purification (1.8X) with Ampure XP Beads (A63880, Agencourt Technology), the libraries were quantitated using a Sybr Green qPCR assay (SsoAdvanced supermix, Biorad) based on a custom *Escherichia coli* reference library. After a dilution of each library down to 26 pM, 0.22 fmol for the 170 bp libraries and 0.44 fmol for the 340 bp library were pooled as templates for the clonal amplification on Ion Sphere particles during the emulsion PCR, performed on a One Touch2 emPCR robot according to the Ion PGM Template OT2 400 Kit user guide (4479878, Life Technologies). The amplification products were loaded onto an Ion 316v2 chip (4483324, Life Technologies), and subsequently sequenced according to the Ion PGM Sequencing 400 Kit user guide (4482002, Life Technologies). After standard filtration of the raw reads (polyclonal and low quality removal), the Ion Torrent sequencing yielded 210,527 sequences for the Xmnl 170 bp sample, 273,427 sequences for the Xmnl 340 bp sample, and 166,099 sequences for the HindIII 170 bp sample. They were deposited in the NIH Short Read Archive (SRA accession numbers SRX1959818, SRX1959817 and SRX1959815).

Alpha satellite sequence filtering

All Xmnl sequences with an average Phred score lower than 25, a length outside the range 162-182 bp for monomers and 324-364 bp for dimers, and sequences without the Xmnl digested sites at the extremities (5'-NNTTC ... GAANN-3') were not considered for further analysis. Alpha satellite sequences were identified with a BLAST search against a reference alpha satellite sequence from *Chlorocebus aethiops* (AM23721) ([Altschul et al., 1990](#)). Using default BLAST parameters, all sequences exhibiting a hit longer than 80 bp for monomers and 160 bp for dimers were considered as alpha satellite sequences and conserved for the following analysis. All sequences were then reoriented if necessary in order to match the orientation of the reference alpha satellite sequence. The orientation information was preserved for investigations regarding reading biases.

Processing of Xmnl dimeric sequences was performed as follows. When an Xmnl site was present in the middle of these sequences, it was used for separating both monomers, providing the so-called left and right monomers located on the 5' side and on the 3' side of the sequence, respectively. Dimers that did not contain any Xmnl site in the middle were aligned against a synthetic sequence formed by two consecutive copies of the reference sequence using the Needleman-Wunsch algorithm ([Needleman and Wunsch, 1970](#)) to identify the monomer limits and split them into left and right monomers according to the same rule as described above. All pairs with at least one monomer outside the 162-182 bp range were discarded. Pairing information was conserved to study association between left and right monomers.

All HindIII sequences with an average Phred score lower than 25, a length outside the range 166-186 bp (the blunting step added 4 nucleotides to the classic monomer length), and sequences without the HindIII digested or blunted sites at the extremities (5'-ATGC ... ATGC-3') were

not considered for further analysis. Alpha satellite sequences were identified with the same BLAST search as above. All sequences were then reoriented if necessary in order to match the orientation of the reference alpha satellite sequence. The four supplementary nucleotides added to the HindIII monomers during the blunting step (3'ATGC) were discarded.

Alpha satellite sequence characterization

Monomeric sequences were compared using their 5-mer composition in order to identify putative alpha satellite families without direct alignment. For each set of monomers, XmnI or HindIII, the 5-mer frequency table was analyzed using a principal component analysis (PCA) to reduce the space complexity and enable data visualization on the first factorial planes. Sequences were classified into groups by using a hierarchical clustering method (HCA) based on the Ward criterion (Ward Jr, 1963) applied to the Euclidean distances calculated from the 100 first principal components of the PCA. Because of the size of the monomer datasets, direct classification of the sequences using HCA was not possible. Instead, HCA was applied on 2500 randomly selected sequences from each monomer dataset which were used to train a linear discriminant model. This model has been finally used to classify all the other monomers. The XmnI dimer dataset was analyzed in two different ways : monomers extracted from dimers without an XmnI site were classified by using an HCA based on a PCA, and monomers extracted from dimers with an XmnI site have been classified by using a linear discriminant analysis trained to recognize the C1 to C6 families.

Because of the size of the datasets, the phylogenetic trees, the consensus sequences and the sequence distance analysis were conducted with different subsets of randomly selected sequences (see Results). The selected sequences were aligned using MUSCLE (Edgar, 2004) and analyzed with SeaView (Gouy et al., 2010). CENP-B and pJalpha boxes were searched in all sequences with the patterns TTCGTTGGAARCGGG and TTCCCTTTYCACCRTAG, respectively (Rosandić et al., 2006), by using the program Fuzznuc (Rice et al., 2000) and allowing 2 mismatches. All statistical analyses were conducted with R (R Core Team, 2013). Our R scripts and other programs are available upon request.

Oligonucleotide probes

Short oligonucleotide probes (18 or 19 nucleotides) were designed in order to target specifically the different alpha satellite families identified in *C. pogonias*, by systematic prediction of binding frequencies based on the sequencing results. Sequences and binding frequencies are available in Supplementary figure 2.4, which also provides details about the positions of locked nucleic acid (LNA) modifications in the probes. These positions were selected based on previous experience in order to achieve a good binding affinity and specificity (Ollion et al., 2015; Cacheux et al., 2016b). When possible, we selected probes that were perfectly complementary

to more than 20 % of the sequences from the target group and to less than 3 % of the sequences from the other groups. Supplementary figure 2.4 also provides the expected binding frequencies if hybridization is possible despite the presence of one mismatch between the probe and its targets. To target the high copy numbered sequence 2 (C158G), we designed two LNA-modified probes (LNA are written in lower case and classic nucleotides are written in upper case) : probe C158G (5'CaCaAgAaGAgCcTtAgC3', 3'Biotin) and probe C158co (5'CaCaA-gAaCAgCcTtAgC3', 3'Digoxigenin) which is complementary to the C1 consensus sequence. LNA-modified probes were purchased from Eurogentec (Seraing, Belgium).

FISH experiments

FISH were performed on metaphase chromosome preparations. Hybridization solutions were prepared by diluting the oligonucleotide probes to a final concentration of 0.1 μ M in a hybridization solution consisting of 2X SSC pH 6.3, 50 % deionized formamide, 1X Denhardt solution, 10 % dextran sulfate, and 0.1 % SDS. 20 μ L of the hybridization solution were deposited on each slide and covered with a coverslip. The slides were then heated for 3 min at 70 °C and hybridized for 1 h at 37 °C in a Thermobrite apparatus (Leica Biosystems). Then, each slide was washed twice in 2X SSC at 63 °C. Preparations were then incubated in blocking solution (4 % bovine serum albumin (BSA), 1X PBS, 0.05 % Tween 20) for 30 min at 37 °C to reduce nonspecific binding. Then, depending on the combination of probes, the following antibodies were used for subsequent revelations : Alexa 488-conjugated streptavidin (1 :200 ; Life Technologies), Cy5-conjugated streptavidin (1 :200 ; Caltag Laboratories), FITC-conjugated sheep anti-digoxigenin (1 :200 ; Roche), and Rhodamine-conjugated sheep anti-digoxigenin (1 :200 ; Roche). All antibodies were diluted in blocking solution containing 1X PBS, 0.05 % Tween 20, and 4 % BSA. Antibody incubation lasted for 30 min at 37 °C. All washings were performed in 2X SSC, 0.05 % Tween 20. Chromosomes were counterstained with DAPI (4',6-diamidino-2-phenylindole) by pipetting 40 μ L of a 5 μ g/mL solution onto the slides, incubating for 5 min and then briefly washing in 1X PBS. Slides were mounted by adding a drop of Vectashield Antifade Mounting Medium (Vector Laboratories) and covering with a coverslip.

Image acquisition and analysis

Metaphases were imaged using an Axio Observer Z1 epifluorescent inverted microscope (Zeiss) coupled to an ORCA R2 cooled CDD camera (Hamamatsu). The Axio Observer Z1 was equipped with a Plan-Apochromat 63 \times 1.4 NA oil-immersion objective and the following filters set : 49 shift free for DAPI (G365 / FT395 / BP445/50), 38 HE shift free for FITC/Alexa488 (BP470/40 / FT495 / BP525/50), homemade sets for Rhodamine (BP546/10 / FF555 / BP 583/22) and for Cy5 (BP643/20 / FF660 / BP684/24). The light source was LED illumination (wavelengths : 365nm, 470nm or 625nm) except for Rhodamine, for which a metal halide

lamp HXP120 was preferred. Immersion oil of refractive index 1.518 at 23 °C was used. Color-combined images were reconstructed using ImageJ ([Abràmoff et al., 2004](#)).

2.2.7 Additional files

Supplementary figure 2.1 : Phylogenetic analysis of alpha satellite DNA diversity in *C. pogonias* monomer dataset

Supplementary figure 2.2 : Characterization of alpha satellite diversity in *C. pogonias* dimer dataset

Supplementary figure 2.3 : Interspecific comparison of alpha satellite diversity in dimer datasets

Supplementary figure 2.4 : LNA-modified probes used to target the C1 to C6 families and the C6' sequences on *C. pogonias* chromosomes

Supplementary figure 2.5 : Hybridization pattern of probe Cx on *C. pogonias* chromosomes

Supplementary figure 2.6 : Distribution pattern of the C2 alpha satellite family on *C. pogonias* chromosomes

Supplementary figure 2.7 : FISH analysis of the C4 alpha satellite family on *C. pogonias* chromosomes

Supplementary figure 2.8 : Distribution pattern of the C5 alpha satellite family on *C. pogonias* chromosomes

Supplementary figure 2.9 : Distribution pattern of the C6 alpha satellite family on *C. pogonias* chromosomes

Supplementary figure 2.10 : FISH analysis of the C5 alpha satellite family on *C. solatus* and *C. pogonias* chromosomes

Supplementary figure 2.11 : Characterization of comet-like clusters in *C. solatus* monomer dataset

Supplementary figure 2.12 : Consensus sequence comparison between XmnI and HindIII monomer datasets

Supplementary table 2.1 : Alpha satellite family associations in *C. pogonias* dimer dataset

2.3 Histoire évolutive de l'ADN alpha satellite chez les Cercopithèques

Troisième article - Evolutionary history of alpha satellite DNA in Cercopithecini

Lauriane Cacheux^{1,2}, Violaine Nicolas², Bernard Dutrillaux², Loïc Ponger¹, Michèle Gerbault-Seureau², Florence Anne Richard^{2,3} and Christophe Escudé¹

¹Structure et Instabilité des Génomes, U1154 INSERM, UMR7196 CNRS, Sorbonne Universités, Muséum national d'Histoire naturelle, Département Régulations, Développement et Diversité Moléculaire, Paris, France

²Institut de Systématique, Evolution, Biodiversité, UMR 7205 CNRS, UPMC, EPHE, Sorbonne Universités, Muséum national d'Histoire naturelle, Département Systématique et Evolution, Paris, France

³Université Versailles St-Quentin, Montigny-le-Bretonneux, France

2.3.1 Abstract

Available data on alpha satellite DNA, the main repeated DNA element of primate centromeres, remain mostly limited to great apes. Among Old World monkeys, Cercopithecini, which have a quite unusual chromosomal evolution by multiple fissions and thus new centromere formations, constitute a promising model to study alpha satellite diversity and dynamics.

We previously characterized six families of alpha satellite sequences, named C1 to C6, in the genome of two *Cercopithecus* species. We now compare their distribution by fluorescence in situ hybridization on the chromosomes of 13 Cercopithecini species, one Papionini species and one Colobinae species. All alpha satellite families were observed to display inter-species variations, being present or absent from genomes, or spanning centromeres or pericentromeres depending on the considered species. Evolutionary new centromeres, in species with highly fissioned karyotypes, appear to be spanned by the most homogeneous family identified.

The interpretation of our results in the light of a molecular phylogeny, independently reconstructed, allowed us to propose an evolutionary scenario for alpha satellite DNA in Cercopithecini. The emergence of each alpha satellite family, their interspecific transfers and the evolution of their distribution pattern are discussed. Our results strongly suggest that alpha satellite

diversification and chromosome rearrangement dynamics are intimately connected. This work provided information about Cercopithecini relationships and thus encourage the integration of alpha satellite DNA into the study of primate evolutionary history.

Keywords : Alpha satellite DNA, Cercopithecini, Centromere, Cytogenetics, Fluorescence in situ hybridization, Genome evolution

2.3.2 Introduction

Centromere regions control chromatid cohesion and segregation during cell divisions in eukaryotes. In Primates, the DNA underlying centromeres and pericentromeres is generally made of large tracts of homologous AT-rich sequences (or monomers) of about 170 bp in length. It corresponds to the tandemly repeated alpha satellite DNA, initially identified through its sedimentation particularities ([Willard, 1991](#); [Rudd et al., 2006](#)).

Multiple cloning, sequencing and cytogenetic studies showed that alpha satellite DNA displayed sequence diversity in the human genome, with monomers being 60 to 100 % identical. Seventeen human families of alpha satellite sequences were characterized, some of them being shared with other primate species ([Willard and Waye, 1987](#); [Alexandrov et al., 1988](#); [Lee et al., 1997](#); [Shepelev et al., 2009](#)). For long, such analysis did not extend beyond great ape genomes ([Archidiacono et al., 1995](#); [Haaf and Willard, 1997](#); [Alkan et al., 2007](#); [Koga et al., 2014](#); [Catacchio et al., 2015](#)). Nevertheless, alpha satellite DNA was recently investigated in Old World monkeys from the Cercopithecini tribe, appealing to high-throughput sequencing, computational and cytogenetic analysis of alpha satellite sequences isolated from the *Cercopithecus solatus* and *C. pogonias* genomes ([Cacheux et al., 2016a,b](#)). Six new families, termed C1 to C6, were identified and characterized ; further use of new sequencing technologies should provide a more comprehensive overview of alpha satellite DNA diversity in Primates.

Alpha satellite families were observed to distribute on specific sets of chromosomes within genomes ; interestingly, families shared between the closely related great ape species generally span non homologous chromosomes, suggesting a fast turnover of sequences residing in centromeres ([Archidiacono et al., 1995](#); [Rudd et al., 2006](#)). The so-called concerted evolution of alpha satellites would rely on mechanisms of non-reciprocal transfer occurring within or between chromosomes, such as unequal crossover, gene conversion, rolling circle replication and reinsertion, and transposon-mediated exchange ([Plohl et al., 2012](#)). By contrast, families may be conserved at centromeres over long evolutionary periods in Old World monkeys ([Cacheux et al., 2016a](#)).

The repetitive nature of alpha satellites and their high sequence homogeneity in centromeres make them difficult if not impossible to assemble ([Eichler et al., 2004](#); [Rudd and Willard, 2004](#)). However, pericentromeres are known to be more heterogeneous and their successful assembly

on specific human chromosomes provided additional clues about alpha satellite evolutionary dynamics (Schueler et al., 2001, 2005; Rudd and Willard, 2004; She et al., 2004; Shepelev et al., 2009). The distribution pattern of alpha satellite families found in these regions led indeed to propose an original evolutionary model, also called the age-gradient hypothesis : during the course of evolution, new families of alpha satellite sequences would expand at the core centromere, thereby splitting and displacing older families distally onto each chromosome arm. The investigation of alpha satellite family distribution on *C. solatus* and *C. pogonias* chromosomes, though not using DNA repeat assembly but highly specific fluorescence in situ hybridizations (FISH), largely supports this hypothesis (Cacheux et al., 2016a,b).

During the course of evolution, chromosome rearrangements shape genomes in many ways, including by the emergence of centromeres which functionally replace older centromeres or provide a centromeric function to acentric chromosomes derived from chromosomal fission events (Montefalcone et al., 1999; Ventura et al., 2001; Rocchi et al., 2012). While these so-called evolutionary new centromeres (ENCs) would not seed on specific DNA sequences, they may slowly acquire large arrays of satellites (Ventura et al., 2007; Wade et al., 2009; Piras et al., 2010; Locke et al., 2011). Alpha satellite DNA dynamics in the context of ENC maturation is still unclear, and the combination of classical and molecular cytogenetics with massive sequencing may be needed to understand further the mechanisms of centromere formation and evolution (Stanyon et al., 2008; Rocchi et al., 2012).

The Cercopithecini tribe is composed of at least 35 species that diverged ten million years ago (Wilson and Reeder, 2005; Guschanski et al., 2013). The Cercopithecini phylogeny is still discussed as strong discrepancies exist between nuclear and mitochondrial markers, which are possibly due to repeated hybridization and extensive backcrossing among ancestors (Sineo, 2012; Guschanski et al., 2013). Notwithstanding, two primary lineages are usually recognized and known as the terrestrial and arboreal clades (Tosi et al., 2004; Tosi, 2008; Xing et al., 2007; Perelman et al., 2011). The evolution of Cercopithecini karyotypes appeared to be mainly due to non-centromeric chromosomal fissions (i.e. chromosomal fissions with a breakpoint outside the centromere) associated to ENC emergences (Dutrillaux et al., 1979, 1981; Stanyon et al., 2004; Moulin et al., 2008), which led to increase their diploid chromosome number from 48 to 72. The alpha satellite diversity that has been recently characterized in *C. solatus* and *C. pogonias* further supports Cercopithecini as a promising model to progress in the understanding of alpha satellite evolutionary dynamics, especially during ENC maturation.

We provide here the largest integration of alpha satellite DNA into comparative cytogenetic studies by analyzing the FISH pattern of probes targeting the C1 to C6 families on the chromosomes of 13 Cercopithecini, one Papionini and one Colobinae species. The results were interpreted in the light of a new Cercopithecini phylogeny, which allowed us to infer an evolutionary history of alpha satellite DNA in this tribe. This history involves family births and deaths, displacements on centromeric regions, amplifications on evolutionary new centromeres and inter-lineage transfers. This study revealed an intimate connection between the diversifica-

tion history of alpha satellite DNA and chromosome rearrangement dynamics in Cercopithecini.

2.3.3 Results

Cercopithecini phylogenetic relationships

We reconstructed the most extensive nuclear phylogeny proposed for Cercopithecini until now by using 22 species and 9 molecular markers. This phylogeny gave strong support to the debated existence of the terrestrial and arboreal clades, and allowed to propose the existence of a basal clade comprising the genera *Allenopithecus* and *Miopithecus* (Figure 3.1). Within the terrestrial clade, nodes corresponding to the *Chlorocebus* genus and to the *Cercopithecus lhoesti* group (*C. lhoesti*, *C. preussi*, *C. solatus*) (Groves, 2001), sometimes called the *Allochrocebus* genus, were well supported. The same applied within the arboreal clade for nodes corresponding to the *C. diana* group (*C. diana*, *C. roloway*), *C. mona* group (*C. mona*, *C. wolfi*, *C. pogonias*, *C. campbelli*) and *C. cephushispanus* group (*C. cephushispanus*, *C. petaurista*, *C. erythrotis*, *C. ascanius*), but not for the one corresponding to the *C. mitis* group (*C. mitis*, *C. nictitans*, *C. albogularis*) (BI support = 0.74). However, the node corresponding to both *C. cephushispanus* and *C. mitis* groups was well supported. Species with the same number of chromosomes tended to cluster together, and the arboreal clade contained all species with highly fissioned karyotypes (highest chromosome numbers); even though the diploid chromosome number varied greatly, from 58 to 72, between arboreal clade species.

Alpha satellite global diversity and distribution in Cercopithecini genomes

Red-colored Cercopithecini species on Figure 3.1 correspond to those involved in the present alpha satellite study; representative species in terms of variety of karyotypes were chosen in the three main clades. The genomes of two outgroup species, one macaque and one colobe, were also investigated. The C1 to C6 alpha satellite families were searched by FISH using previously designed short locked nucleic acid (LNA) modified probes (Cacheux et al., 2016a,b)(see Supplementary tables 3.1 and 3.2); results are summarized in Table 3.1.

Cercopithecini terrestrial clade

Data on alpha satellite families present in *C. solatus* ($2n = 60$) being already available (Cacheux et al., 2016a,b), two other species from the terrestrial clade, *Erythrocebus patas* ($2n = 54$) and *C. lhoesti* ($2n = 60$), were studied here (Figure 3.1).

In *C. lhoesti*, FISH signals for the C1 and C2 families were observed at most centromeres (primary constrictions) and pericentromeres (around the primary constrictions), respectively

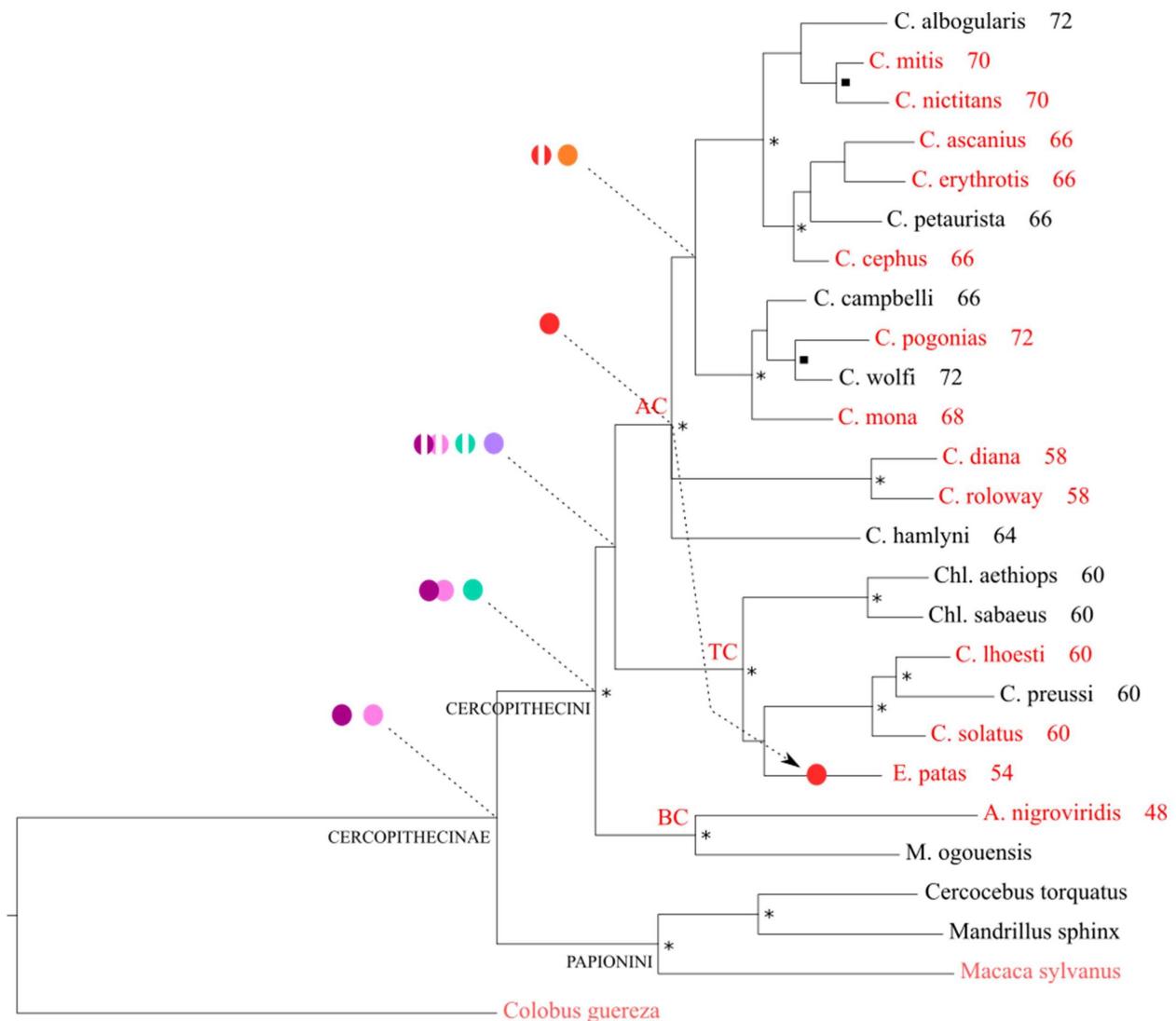


FIGURE 3.1 – Cercopithecini phylogenetic relationships (Bayesian inference) and alpha satellite DNA evolutionary dynamics. The phylogeny was produced using nine nuclear markers (ABCA1, BRCA2, CFTR, DENND5A, ERC2, LRPPRC-169, SRY, TTR and ZFX). Nodes with BI support of 1 are labeled with a star and nodes with BI support ≥ 0.98 and < 1 are labeled with a square. Numbers associated to species names represent their diploid chromosome number. Cercopithecini species and representatives species for outgroups involved in the present alpha satellite study are colored in red. The five Cercopithecini genera are represented on the phylogeny : C : *Cercopithecus*, Chl : *Chlorocebus*, E : *Erythrocebus*, A : *Allenopithecus*, M : *Miopithecus*. AC : Arboreal clade, TC : Terrestrial clade, BC : Basal clade. The hypothesized emergence of the C1 to C6 alpha satellite families was inferred from the present study results and is displayed with circles (C1 : purple, C2 : pastel green, C3 : dark pink, C4 : light pink, C5 : red, C6 : orange). The displacement of some families toward pericentromeres is displayed with broken circles. Arrow : hypothesized genetic transfer.

TABLE 3.1 – Chromosomal distribution of the C1 to C6 alpha satellite families in Cercopithecini

	Species	C1	C2	C3-C4	C5	C6
AC	<i>C. mitis</i>	+	c	+	pc	nd
	<i>C. nictitans</i>	+	c	+	pc	+
	<i>C. ascanius</i>	+	c	+	pc	+
	<i>C. erythrotis</i>	+	c	+	pc	+
	<i>C. cephush</i>	+	c	+	pc	+
	<i>C. pogonias</i>	+	c	+	pc	+
	<i>C. mona</i>	+	c	+	pc	nd
	<i>C. diana</i>	+	c	+	pc	+
	<i>C. roloway</i>	+	c	+	pc	+
TC	<i>C. lhoesti</i>	+	c	+	pc	(+)
	<i>C. solatus</i>	+	c	+	pc	(+)
	<i>E. patas</i>	+	c	+	pc	+
BC	<i>A. nigroviridis</i>	(+)	+	c	(+)	-
Out	<i>M. sylvanus</i>	-	(+)	na	(+)	(+)
	<i>C. angolensis</i>	-	-	-	(+)	-

NOTE - AC : arboreal clade. TC : terrestrial clade. BC : basal clade. Out : outgroups. +/- : presence/absence. (+) : non-specific signal (see Results). c/pc : centromeres/pericentromeres (global distribution tendency). nd : non-detected. na : detected, but non-associated. Results as regards *C. solatus* and *C. pogonias* are included according to Cacheux et al. (2016a,b).

(Figure 3.2A). Strong C2 signals were detected on the short arm of acrocentric chromosomes. The distribution of C1 and C2 signals was roughly similar in *E. patas* (Figure 3.2B) but with additional particularities : C2 probes hybridized at some telomeric and interstitial regions of submetacentric chromosomes (see arrow 1 and 2 on Figure 3.2D) ; C1 and C2 signals colocalized on one chromosome pair (see arrow 3 on Figure 3.2D) ; and especially large C1 signals were observed on some chromosomes (see arrow 4 on Figure 3.2D). C3 and C4 probes provided few if any signals on *C. lhoesti* and *E. patas* pericentromeres (Supplementary figure 3.1).

C5 probes provided signals on *C. lhoesti* chromosomes, but their intensity decreased substantially when probes were washed at 68 °C after the hybridization step, instead of the usual 63 °C (Supplementary figure 3.2). This phenomenon is likely to reveal a nonspecific hybridization of C5 probes on *C. lhoesti* chromosomes, which suggests that the C5 family is absent from the *C. lhoesti* genome. At difference, temperature-resisting signals were provided by C5 probes at the centromeres of 9 chromosome pairs of *E. patas* (Figure 3.2C and Supplementary figure 3.2). Finally, C6 probes did not provide any signals on *C. lhoesti* and *E. patas* chromosomes.

Cercopithecini arboreal clade

Data on alpha satellite families present in *C. pogonias* (2n = 72) being already available (Cacheux et al., 2016a), eight other species from the arboreal clade were studied here : *C. roloway* (2n = 58), *C. diana* (2n = 58), *C. mona* (2n = 68), *C. cephush* (2n = 66), *C. erythrotis* (2n =

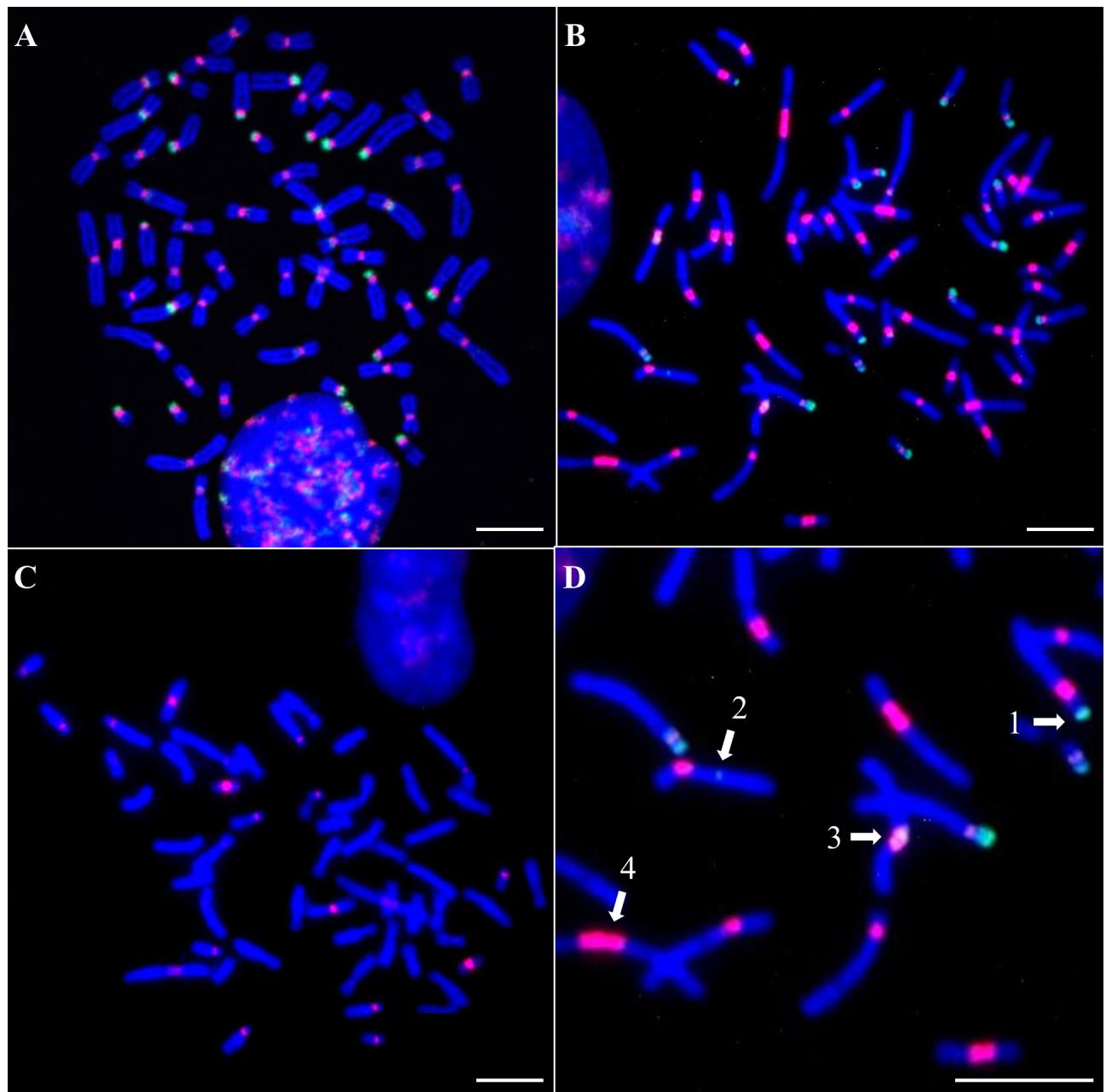


FIGURE 3.2 – FISH analysis of the C1, C2 and C5 alpha satellite families in species from the terrestrial clade. Metaphase chromosomes are colored in blue. (A) Hybridization of probes C1b (red) and C2b (green) on *C. lhoesti* chromosomes. (B) Hybridization of probes C1a (red) and C2a (green) on *E. patas* chromosomes. (C) Hybridization of probe C5a (red) on *E. patas* chromosomes. (D) Focus on image (B). 1 : C2a labels telomeric regions on some submetacentric chromosomes, 2 : C2a labels some interstitial regions, 3 : C1a and C2a signals overlap on some centromeric regions, 4 : C1 signals extend on long arrays on some chromosomes. Scale bar = 10 µm.

66), *C. ascanius* ($2n = 66$), *C. nictitans* ($2n = 70$) and *C. mitis* ($2n = 70$) (Figure 3.1). Figures 3.3 and 3.4 show alpha satellite distribution for *C. roloway* and *C. cephush* as representative examples.

For all arboreal clade species, C1 and C2 signals were detected at most centromeres and pericentromeres, respectively (Figure 3.3). No signal was observed on any Y chromosome (see arrow 1 on Figure 3.3B); a single other chromosome displayed no signal in *C. cephush* exclusively (see arrow 2 on Figure 3.3B). The C2 family was detected at some telomeric regions in all species, and sometimes at interstitial positions (see arrows on Figure 3.3C,E). Very occasionally, C2 signals were located at centromeres and were surrounded by pericentromeric C1 signals (see arrows on Figure 3.3D). C3 and C4 probes provided few if any signals on pericentromeres.

C5 signals were observed on the chromosomes of all arboreal clade species, but with different distributions on centromeric regions. In the species possessing the least fissioned karyotypes (*C. roloway* and *C. diana*), C5 signals were revealed at the centromere of 6 chromosome pairs and did not extend toward pericentromeres (Figure 3.4A,D). At difference, in species with highly fissioned karyotypes (*C. mona*, *C. cephush*, *C. erythrotis*, *C. ascanius*, *C. nictitans* and *C. mitis*), C5 signals had a pericentromeric distribution on several chromosome pairs (Figure 3.4B,E,F). All C5 signals were persistent after stringent washing conditions. C6 signals were only observed in species with highly fissioned karyotypes (mentioned above); especially strong signals were observed at the centromere of their acrocentric chromosomes (Figure 3.4C).

Cercopithecini basal clade, Papionini and Colobinae

The representative species of the basal clade in the present study is *Allenopithecus nigroviridis* ($2n = 48$). FISH signals were observed with C1 probes on several chromosomes; however, our two probes C1a and C1b, which target different sites on C1 sequences, provided non overlapping signals (Supplementary figure 3.3). This feature suggests that the sequences detected by these probes on *A. nigroviridis* chromosomes do not belong to the C1 family. Consequently, this family may lack in the *A. nigroviridis* genome. At difference with other Cercopithecini species, C2 signals were observed at the centromere of all chromosomes, with the exception of the Y chromosome (Figure 3.5A,D). In addition, C2 signals were detected at the extremity of acrocentric chromosome short arms (see arrows on Figure 3.5D). C3 and C4 probes also provided clear signals at the centromeres of all chromosomes, with the exception of the Y chromosome, and at the extremity of acrocentric chromosome short arms (Figure 3.5B,C,E). A unique chromosome was labelled by C3 but not C4 probes (see arrow 2 on Figure 3.5E). Slight signals were observed with C5 probes; however, the intensity of these signals decreased substantially when probes were washed at 68°C instead of 63°C , which suggests the absence of the C5 family in the *A. nigroviridis* genome. No signal was detectable for the C6 family.

Rather similar results were obtained when searching for the presence of the C1 and C5 families in the outgroup *Macaca sylvanus*. In addition, the two probes C2a and C2b, which target different sites on C2 sequences, provided non overlapping signals (Supplementary figure 3.4).

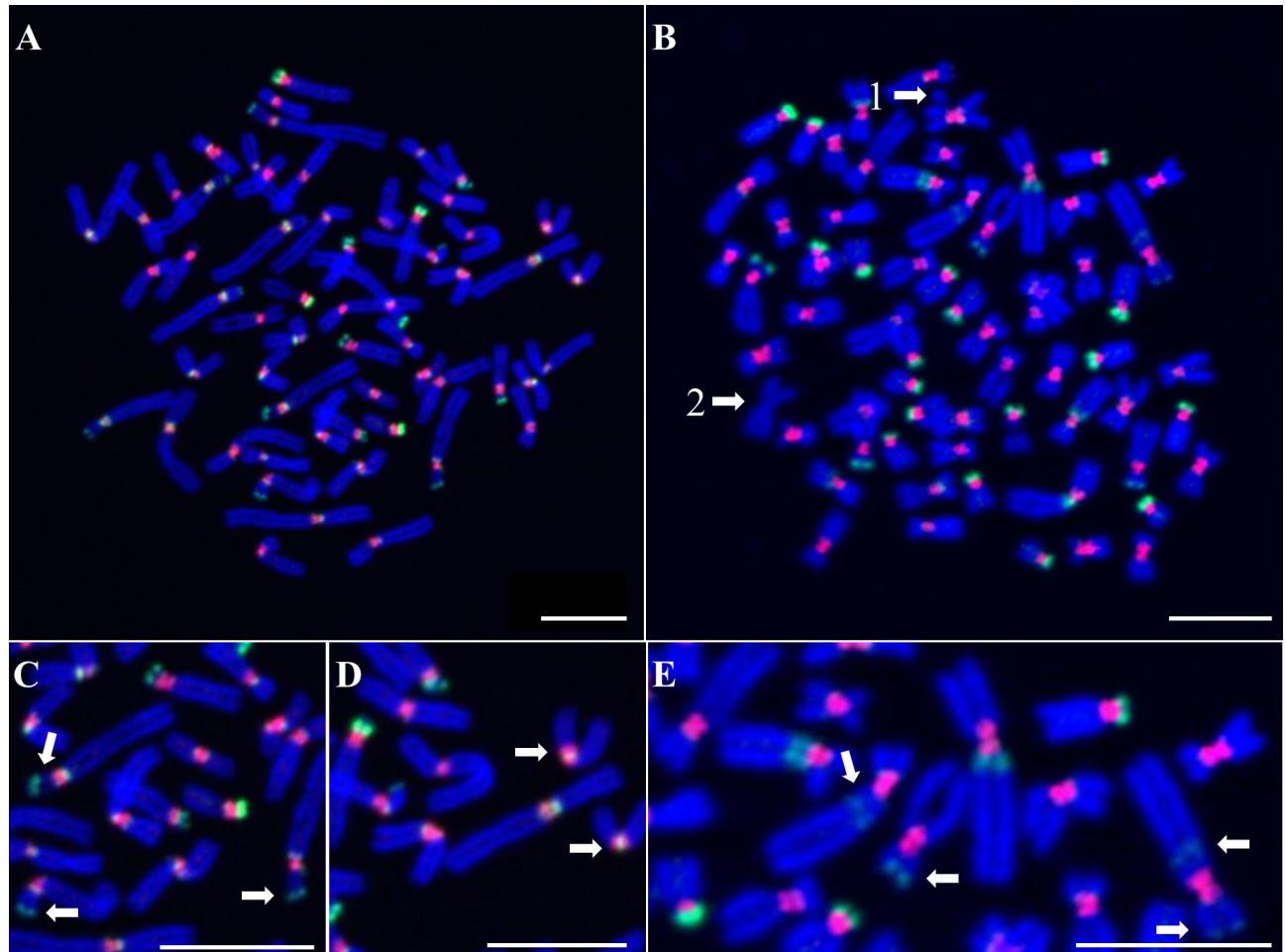


FIGURE 3.3 – FISH analysis of the C1 and C2 alpha satellite families in species from the arboreal clade. Metaphase chromosomes are colored in blue. Hybridization of probes C1b (red) and C2b (green) on (A) *C. roloway* female chromosomes and (B) *C. cebus* male chromosomes. Arrow 1 points to the unlabeled Y chromosome. Arrow 2 points to an additional unlabeled chromosome. (C) Focus on image (A). Arrows point to C2 signals at telomeric regions. (D) Focus on image (A). Arrows point to chromosomes where C1 signals are revealed around C2 signals on centromeric regions. (E) Focus on image (B). Arrows point to C2 signals at telomeric regions or inside chromosome long arms. Scale bar = 10 µm.

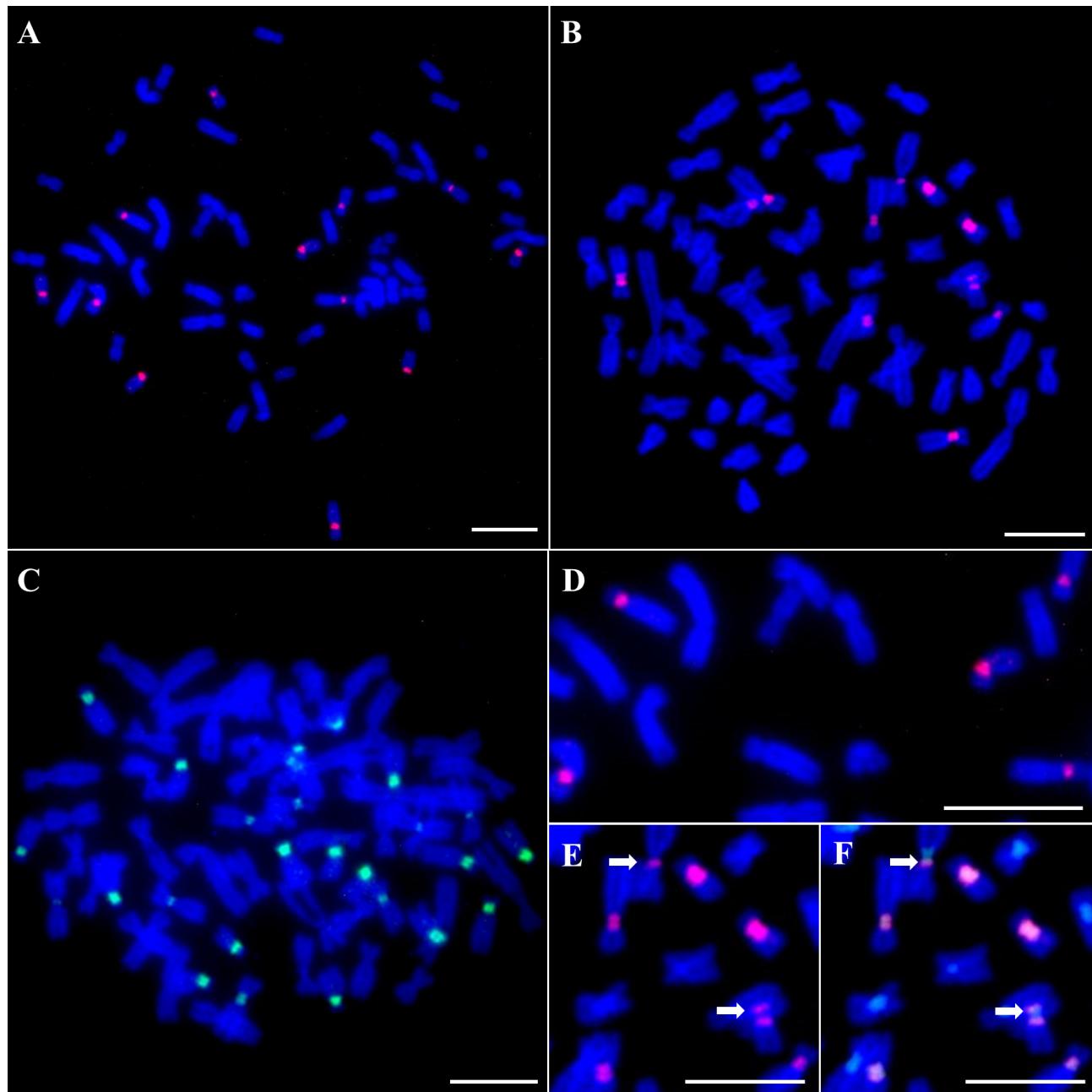


FIGURE 3.4 – FISH analysis of the C5 and C6 alpha satellite families in species from the arboreal clade. Metaphase chromosomes are colored in blue. (A) Hybridization of probe C5a (red) on *C. roloway* chromosomes. (B) Hybridization of probes C5a (red) and C1a (not shown) on *C. cephush* chromosomes. (C) Hybridization of probe C6a (green) on *C. cephush* chromosomes. (D) Focus on image (A). (E) Focus on image (B). Arrows point to chromosomes where C5 signals are revealed at pericentromeres. (F) Same focus as (E) with hybridization of probe C1a shown in green. Scale bar = 10 µm.

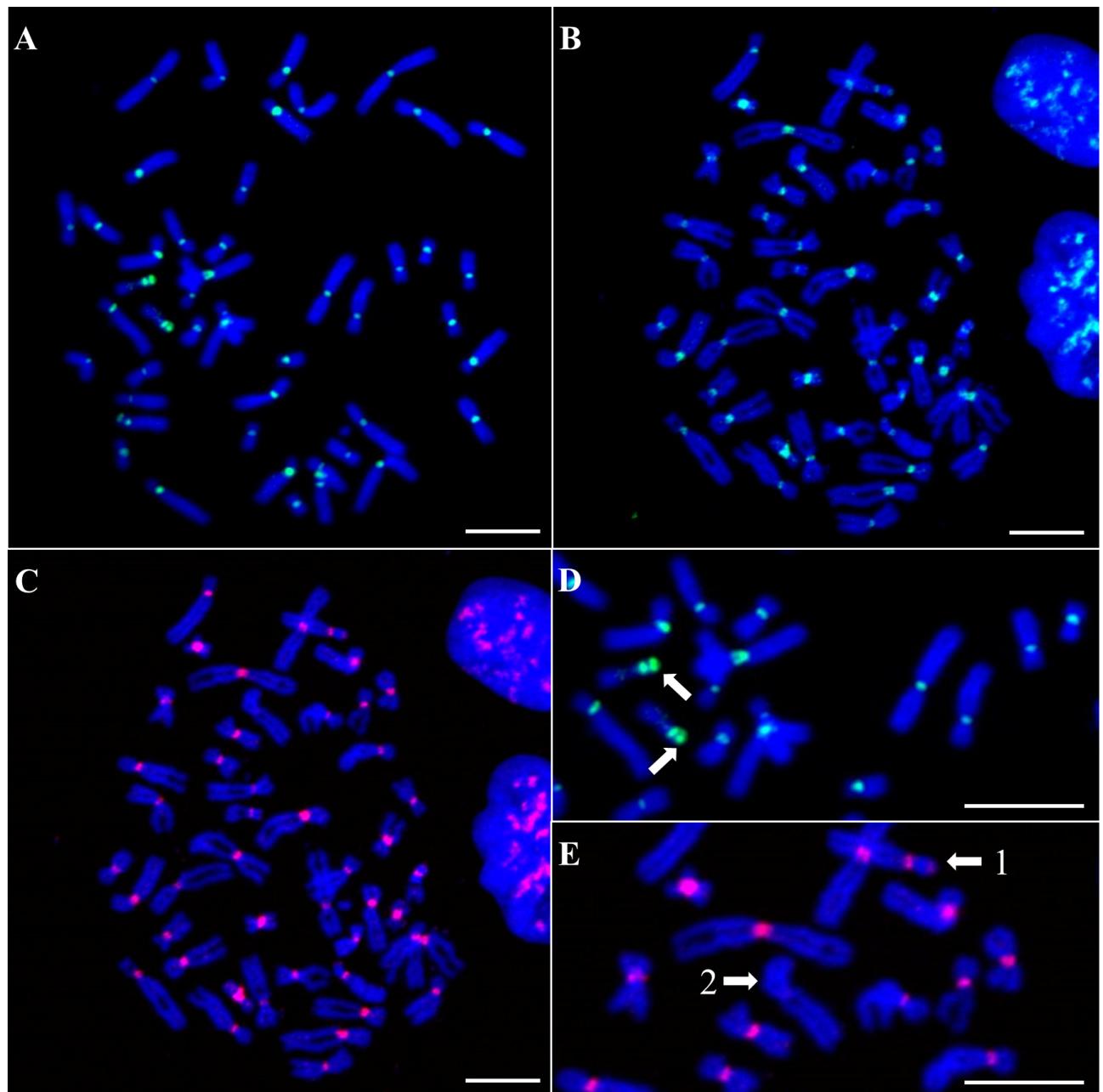


FIGURE 3.5 – FISH analysis of the C2, C3 and C4 alpha satellite families in a species from the basal clade. *Allenopithecus nigroviridis* metaphase chromosomes are colored in blue. (A) Hybridization of probes C1b (not shown) and C2b (green). (B,C) Hybridization of probes C3a (green) and C4a (red). (D) Focus on image (A). Arrows point to labelled extremities of acrocentric chromosome short arms. (E) Focus on image (C). 1 : labelled extremity of an acrocentric chromosome short arm. 2 : unlabeled chromosome. Scale bar = 10 µm.

Consequently, the C2 family may lack in the *M. sylvanus* genome. C3 and C4 probes provided clear signals on centromeric regions (Supplementary figure 3.5) ; however, their respective signals did not clearly overlap, which suggests that the C3 and C4 families are not associated into dimeric sequences in this species. Finally, clear C6 signals were observed on all chromosomes. Considering that the C6 family was revealed in the Cercopithecini arboreal clade species with highly-fissioned karyotype only, i. e. the most derived, these C6 signals on *M. sylvanus* chromosomes are likely to reveal another alpha satellite family that would possess a nucleotide site detected by C6 probes.

In the species *Colobus angolensis*, which represents the second outgroup of this study, signals were detected with C5 probes only. However, these signals decreased substantially when probes were washed at 68 °C instead of 63 °C, which suggests that none of the C1 to C6 families are present in the *C. angolensis* genome.

Alpha satellite detailed diversity and distribution on Cercopithecini chromosomes

We finally compared alpha satellite diversity on homologous chromosomes between species, to evaluate both conservation and variation of alpha satellite family distribution through evolution. For this study, six phylogenetically dispersed species that are representative for the diversity of karyotypes and global alpha satellite distributions in Cercopithecini were selected : *A. nigroviridis* (ANI), *E. patas* (EPA), *C. solatus* (CSO), *C. roloway* (CRO), *C. cephus* (CCE) and *C. pogonias* (CPO). This comparative study was conducted for the three alpha satellite families C2, C5 and C6. C1, which displayed few interspecific variations, and C3/C4, which were hardly detectable, have been discarded. Figure 3.6 is a schematic version of Supplementary figure 3.6 that shows the alignment of homologous chromosomes between the six studied species and the presumed ancestral chromosomes of Cercopithecidae, i.e. Old World monkeys (ANC), based on R banding patterns and chromosome painting (Dutrillaux et al., 1981; Moulin et al., 2008). C2, C5 and C6 distribution patterns were pictured on chromosomes according to the achievement of karyotypes following FISH experiments (see Supplementary figures 3.7, 3.8 and 3.9). As C2 was observable on most centromeric regions in greater or lesser amounts, only centro- or pericentromeric loci displaying strong C2 signals were pictured here ; by contrast, all telomeric or interstitial loci displaying C2 signals were pictured.

Inter-species homologous chromosomes

C2 probes gave strong FISH signals on the short arms of all acrocentric chromosomes in all the studied species (Figure 3.6 and Supplementary figure 3.7). In some instance, large amounts of C2 sequences were detected at centromeric regions of homologous chromosomes ; see for example homologs to ANC4 : ANI2, EPA2, CRO1 and CCE1 on Figure 3.6. C2 sequences were also observed at telomeric regions of homologous chromosomes, as on homologs to ANC9 : EPA10, CRO9 and CCE4.

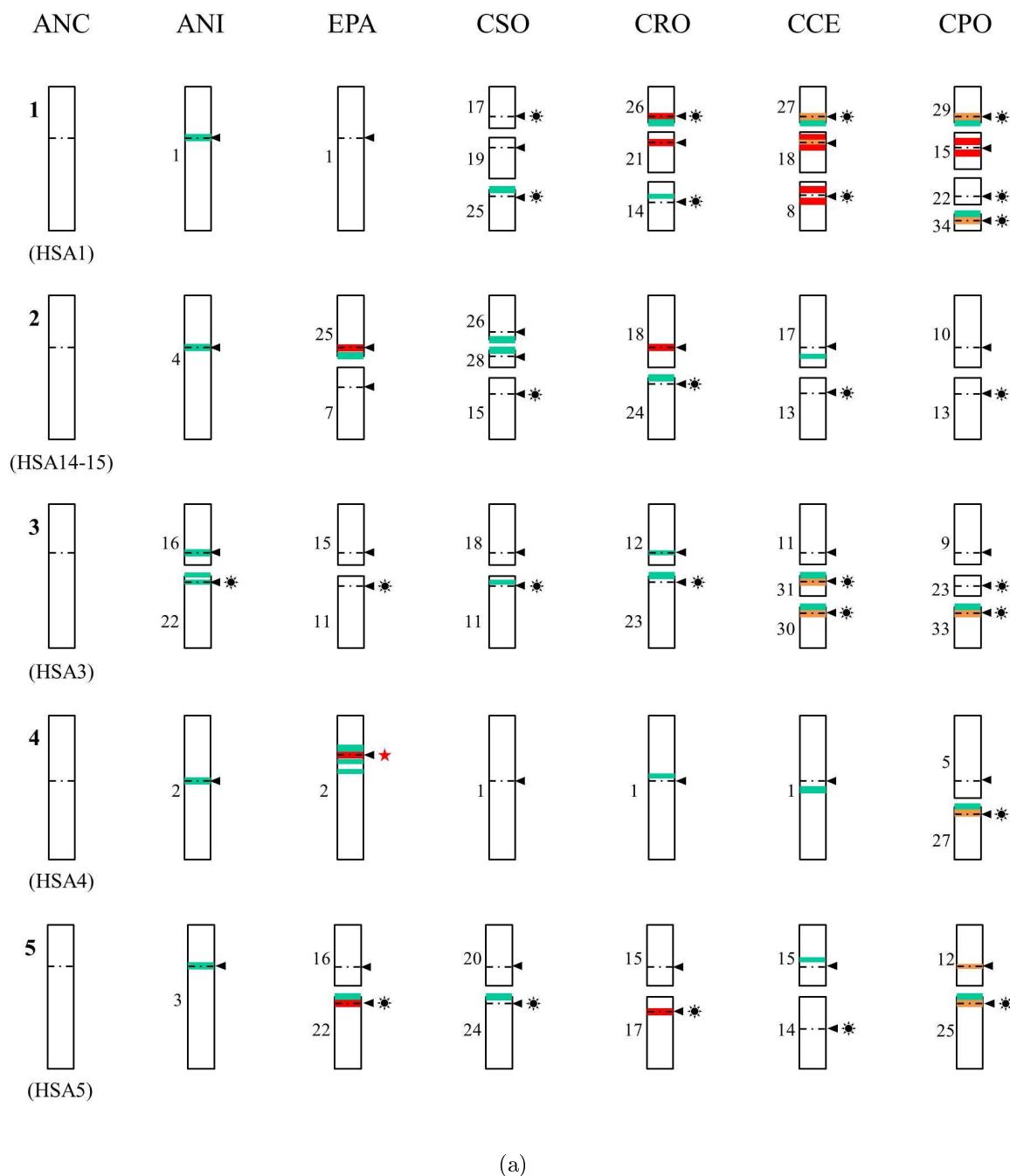
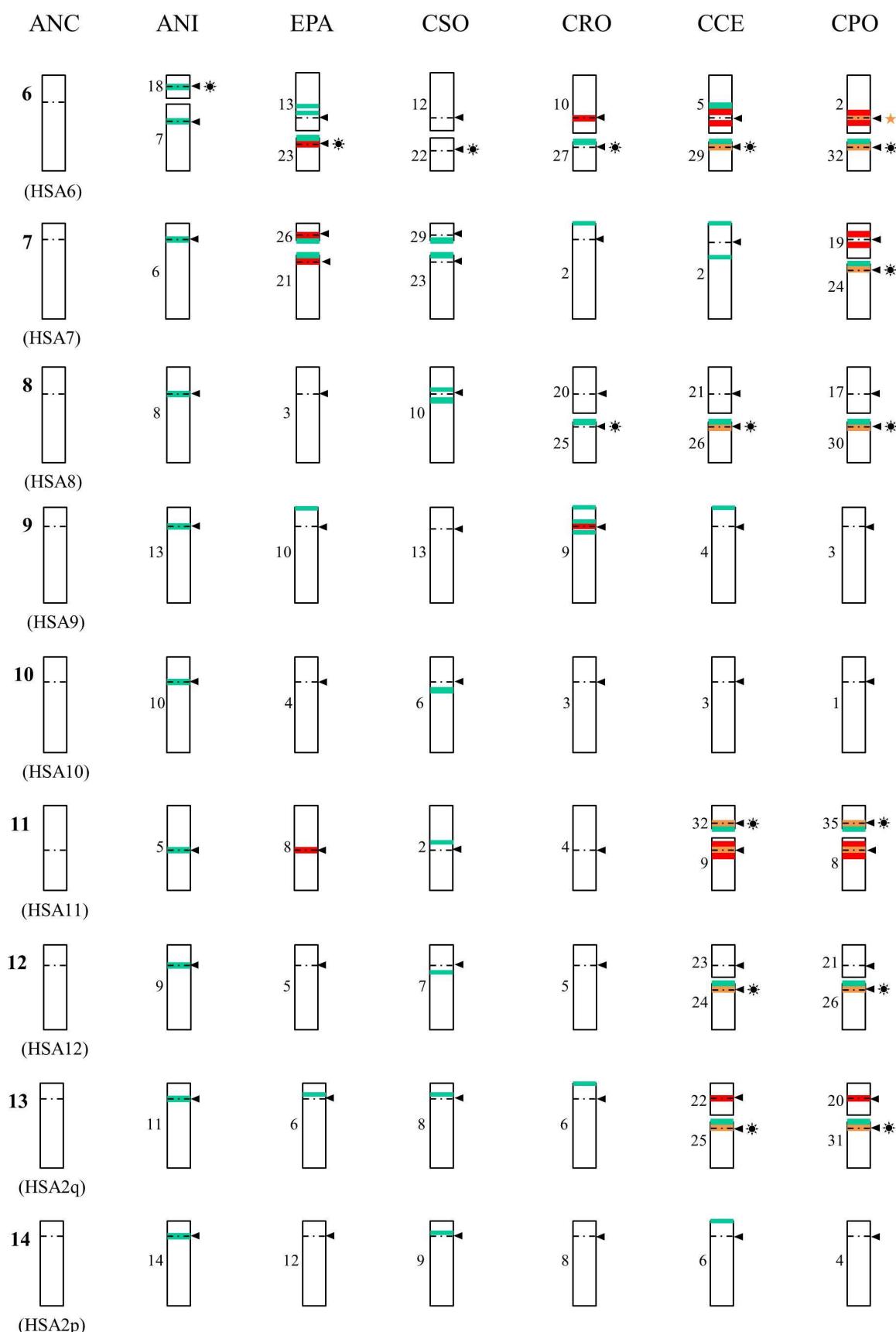
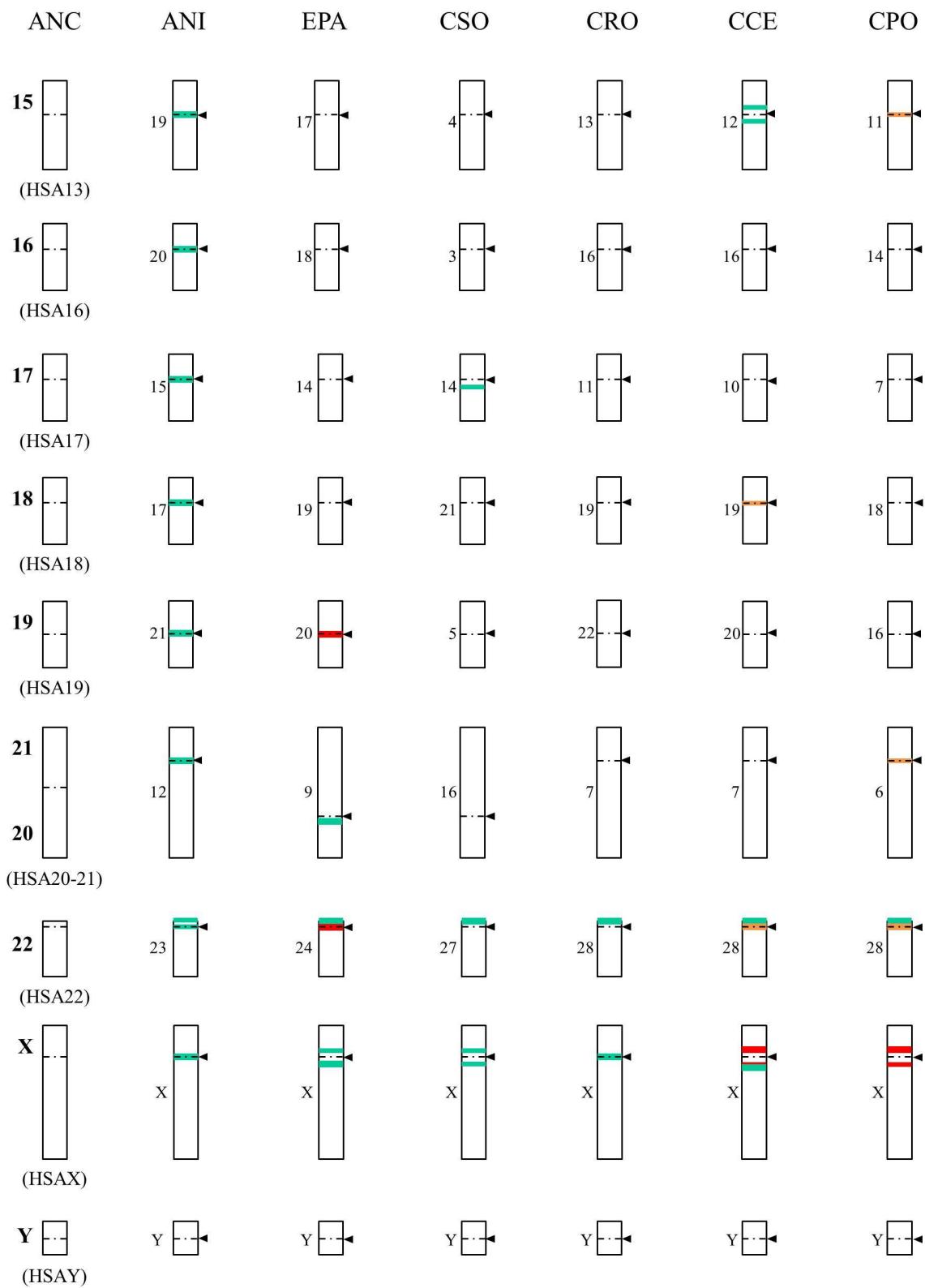


FIGURE 3.6 – Schema of alpha satellite distribution on homologous chromosomes and centromeres. Presumed ancestral chromosomes of Cercopithecidae, i.e. Old World monkeys (ANC), are aligned with their homologs in *A. nigroviridis* (ANI), *E. patas* (EPA), *C. solatus* (CSO), *C. roloway* (CRO), *C. cephus* (CCE) and *C. pogonias* (CPO) (Dutrillaux et al., 1981; Moulin et al., 2008). Chromosome numbers in reference karyotypes are displayed on their left side (Dutrillaux et al., 1978, 1979, 1988a; Moulin et al., 2008). Homologies with human chromosomes (HSA) are mentioned. Arrows and dotted lines point to centromere positions. Distribution is shown in pastel green for C2, red for C5 and orange for C6 FISH signals. Two different widths were used according to the intensity of the signals. For C2, only the strongest signals observed at centromeric regions are pictured. Black stars stand for putative ENCs that emerged after Cercopithecidae divergence, in relation with fission events. Colored stars indicate heterozygosity, as C5 (red) or C6 (orange) signals were provided on a single chromosome of the concerned pairs.



(b)



(c)

C5 signals were also detected at centromeric regions of homologous chromosomes, as homologs to ANC1 : CRO21, CCE18 and CPO15 (Figure 3.6 and Supplementary figure 3.8). At difference, some chromosomes displayed C5 signals in unique species, as homolog to ANC9 : CRO9. C5 signals were observed at the centromere of all acrocentric chromosomes in *E. patas*, and of a single acrocentric pair in *C. roloway* (ANC1 : CRO26). Finally, one should notice that C5 signals were observed on a single chromosome from the EPA2 pair (ANC4) regarding the studied specimen.

C6 signals were detected at the centromere of all acrocentric chromosomes of *C. cephush* and *C. pogonias* (Figure 3.6 and Supplementary figure 3.9). Many of them are homologous between the two species. In addition to that, C6 signals were noticed on two homologous submetacentric chromosomes : CCE9 and CPO8 (ANC11). Weaker C6 signals were occasionally observed on non-homologous meta- or submetacentric chromosomes, as CPO12 (ANC5), CPO11 (ANC15) and CCE19 (ANC18). Finally, C6 signals were detected on a single chromosome from the CPO2 pair (ANC6) regarding the studied specimen.

Inter-species homologous centromeres

As mentioned above, FISH signals provided by C2 and C5 probes were observed at centromeres or pericentromeres depending on the considered species. We can see here that strong C2 signals were provided at the centromere of *A. nigroviridis* chromosomes and at the pericentromeres of some homologs in other species, as for homologs to ANC4 : ANI2, EPA2, CRO1 and CCE1. Similarly, C5 signals were observed at the centromere of *E. patas* and *C. roloway* chromosomes and at the pericentromeres of some homologs; see for example homologs to ANC6 : CRO10, CCE5 and CPO2 or to ANC11 : EPA8, CCE9 and CPO8.

Evolutionary new centromeres

Chromosomes that possess a putative ENC are specified with black stars on Figure 3.6 ; according to banding pattern, these chromosomes are indeed homologous to presumed ancestor chromosome arms and may be the result of non-centromeric fissions of ancestral chromosomes (Dutrillaux et al., 1979; Moulin et al., 2008)(see Supplementary figure 3.6). Strong FISH signals were observed with C2 probes at the centromeric regions of a large proportion of these putative ENCs : 2/2 for *A. nigroviridis*, 2/3 for *E. patas*, 3/6 for *C. solatus*, 6/7 for *C. roloway*, 8/11 for *C. cephush* and 11/14 for *C. pogonias*. C5 signals were occasionally detected on putative ENCs : 2/3 for *E. patas*, 2/7 for *C. roloway* and 1/11 for *C. cephush*. Finally, we can notice that C6 signals were observed on most *C. cephush* and *C. pogonias* putative ENCs : 8/11 and 11/14, respectively.

2.3.4 Discussion

Our previous results (Cacheux et al., 2016a,b) and those presented here enable us to infer the presence and chromosomal distribution of the C1 to C6 alpha satellite families in 13 Cercopithecini species and 2 other Old World monkey species. These results, combined with the phylogeny of Cercopithecini, allow us to decipher the evolutionary history of alpha satellite DNA in this tribe (see Figure 3.1).

Tree-based alpha satellite evolutionary history in Cercopithecini

The molecular phylogeny we propose for the Cercopithecini tribe is consistent with previous phylogenies based on nuclear (Tosi et al., 2004; Tosi, 2008; Xing et al., 2007; Perelman et al., 2011), chromosomal (Dutrillaux et al., 1979, 1981; Moulin et al., 2008), morphological (Cardini and Elton, 2008) and vocalization data (Gautier, 1988), in displaying a terrestrial and an arboreal clades. It also displays a new "basal clade" comprising *A. nigroviridis* and *Miopithecus ogouensis*. All nodes that are essential to reconstruct alpha satellite evolutionary history in Cercopithecini are well supported but two : the node corresponding to the common ancestor of the terrestrial and arboreal clades (BI support = 0.70), and the node corresponding to the common ancestor of the *C. mona/mitis/cephus* groups (BI support = 0.63). However, these nodes were previously found in the phylogeny based on chromosomes (Moulin et al., 2008) and they are supported by the species distribution of alpha satellite families (see below). The monophyly of the clade grouping the Cercopithecini and the Papionini tribes is not tested here, but was largely documented elsewhere (Page and Goodman, 2001; Rodrigues et al., 2009; Perelman et al., 2011).

Dynamics of alpha satellite family emergences in Cercopithecini

Sequencing experiments and in silico analyzes previously showed that the C3 and C4 families were associated into dimeric sequences in the *C. solatus* and *C. pogonias* genomes (Cacheux et al., 2016a,b). Both families were detected by FISH and shown to colocalize at *C. solatus* pericentromeres, which supported their association within genomes. By contrast, C3 probes were probably not sufficiently efficient to detect the C3 family on *C. pogonias* chromosomes, and only the C4 family was revealed at *C. pogonias* pericentromeres. Similarly, in the present study only slight FISH signals, if any, were detected for the C3 and C4 families at the pericentromeres of terrestrial and arboreal clade species. By contrast, clear signals were detected for both families at centromeres of the basal clade species *A. nigroviridis*, and at centromeric regions of the Papionini *M. sylvanus*. C3 and C4 signals colocalized on *A. nigroviridis* chromosomes but showed non-overlapping patterns in *M. sylvanus*. Consequently and according to the age-gradient hypothesis (Schueler et al., 2001, 2005; Shepelev et al., 2009), we propose that

the C3 and C4 families emerged in Cercopithecinae (Cercopithecini and Papionini) ancestors and amplified being associated in Cercopithecini ancestors. Conserved at centromeres in basal clade species, C3 and C4 would have been displaced toward pericentromeres in the common ancestors of terrestrial and arboreal clade species. These old families may tend to disappear in both terrestrial and arboreal clades, displaying not enough repeats in some genomes to be detected using FISH experiments.

The C2 family was detected at centromeres of the basal clade species *A. nigroviridis* and at pericentromeres of all terrestrial and arboreal clade species. Therefore, this family has probably emerged in Cercopithecini ancestors and been displaced toward pericentromeres in the common ancestors of both terrestrial and arboreal clade species.

The C1 family was not observed in *A. nigroviridis*, but C1 signals were revealed at almost every centromere of all terrestrial and arboreal clade species ; this species distribution, combined with the locations of C2, C3 and C4 on Cercopithecini centromeric regions, gives further support to the existence of a clade comprising terrestrial and arboreal clade species and to an external position of the *A. nigroviridis* lineage. C1 may have emerged in the common ancestors of terrestrial and arboreal clade species, rapidly colonized genomes, and been conserved at centromeric position since then. As regards, alpha satellite families could remain active for long periods during evolution, retaining centromeric position and high sequence homogeneity (95 % mean sequence identity here).

The species and chromosomal distribution of the C5 family is more complex to interpret. This family was observed in all arboreal clade species and one terrestrial clade species, *E. patas*. Possibly, the C5 family emerged in the common ancestors of terrestrial and arboreal clade species and got extinct in the terrestrial lineage leading to *C. solatus* and *C. lhoesti*; but the disappearance of an entire alpha satellite family in such a short period, without leaving a trace at pericentromeres, is hardly convincing. Furthermore, ancient hybridizations between arboreal clade species and *E. patas* ancestors were previously suggested by discrepancies between phylogenies (Tosi, 2008; Perelman et al., 2011; Guschanski et al., 2013). Consequently, the C5 family may have emerged in arboreal clade ancestors and been transferred to *E. patas* ancestors following genetic introgressions. Interestingly, the C5 family adopts a centromeric position in *C. diana* and *C. roloway* and a pericentromeric one in all other arboreal clade species. Thus, this family may have been displaced toward pericentromeres in the common ancestors of the *C. mona/mitis/cephus* groups, after its emergence at centromeric position in the arboreal clade ancestors. The intriguing pericentromeric distribution of the C5 family compared to C1 in most species leads to consider a slight reassessment of the age-gradient hypothesis, as older families could reintegrate centromere core regions after having been displaced toward pericentromeres by younger families (Cacheux et al., 2016a).

Finally, the C6 family was exclusively detected at centromeric position in species from the *C. mona/mitis/cephus* groups ; this species distribution, combined with the locations of C5

on arboreal clade species centromeric regions, gives further support to the existence of a clade comprising the species from the *C. mona/mitis/cephus* groups. C6 would be the last-born family of our study, with a recent emergence in the common ancestors of the *C. mona/mitis/cephus* groups. Interestingly, the dynamics of family emergences proposed here are in accordance with the mean sequence identities of each family, in the likely hypothesis that sequence identity is higher in young than old families (C3 : 87 %, C4 : 85 %, C2 : 85 %, C1 : 95 %, C5 : 95 %, C6 : 98 %)(Cacheux et al., 2016a).

Alpha satellite DNA diversity on evolutionary new centromeres

Determination of ENC satellite composition remains unresolved by genome assembly attempts (Rocchi et al., 2012). The increase of chromosome number during Cercopithecini evolution, associated with non-centromeric chromosome fissions, implies the formation of numerous ENCs in their genomes. As predicted by Stanyon et al. (2012), the combination of high throughput sequencing with classical and molecular cytogenetics allowed us to propose a molecular characterization of the putative ENCs of several Cercopithecini species. Especially, within the highly fissioned genomes of the arboreal clade species *C. pogonias* and *C. cephushis*, most ENCs appear to be spanned by the newly-emerged C6 family. Interestingly, all those ENCs are located on acrocentric chromosomes, echoing the hypothesized role of centromeric protein recruitment at DNA double-strand breaks in the formation of centromeres (Zeitlin et al., 2009). Only one homologous pair of acrocentric chromosomes would not possess ENCs in *C. cephushis* and *C. pogonias* (CPO28 and CCE28) ; one should notice that C6 is still detected at the centromere of these chromosomes, which stands once more for peculiar DNA exchanges between the centromeric regions of acrocentric chromosomes (Choo et al., 1990; Warburton et al., 2008; Cacheux et al., 2016a,b). This idea is strengthened by the surprising observation of the old C2 family on the short arm of all acrocentrics ; this family may have started to colonize acrocentric ENC pericentromeres following exchanges with the older pericentromeres of CPO28 and CCE28.

Centromere DNA evolution and chromosome rearrangement dynamics

Mechanisms driving the progressive proximal expansion of alpha satellite DNA at centromeres remain largely unclear, and the impact of chromosome rearrangements on the dynamics of such expansion has not been investigated so far. Since the divergence of Cercopithecini, chromosomes of basal clade species such as *A. nigroviridis* ($2n = 48$) did not undergo many evolutionary fissions, and chromosomes of terrestrial clade species went through fissions that increased moderately their diploid number ($2n = 54$ to 60)(Dutrillaux et al., 1979; Moulin et al., 2008). By contrast, arboreal clade species possess highly rearranged karyotypes, in particular in the *C. mona/mitis/cephus* groups ($2n = 66$ to 72) and at difference with the *C. diana* group ($2n = 58$). Thus, the successive formations of C1 (in terrestrial/arboreal clade ancestors), C5 (in

arboreal clade ancestors), and C6 (in ancestors of the *C. mona/mitis/cephus* groups) are correlated with Cercopithecini chromosome rearrangement dynamics. Similarly, the pericentromeric displacement of C2, C3 and C4 (in terrestrial/arboreal clade ancestors) and C5 (in ancestors of the *C. mona/mitis/cephus* groups) sign periods of large karyotype rearrangements.

This intriguing parallel brings new insights about centromere evolution : alpha satellite proximal expansion at centromeres may be driven by chromosome rearrangements, at least in Cercopithecini. Transposable elements could take part in this phenomenon, as they can integrate into proximal centromeric domains and are known to be active under stress conditions (Capy et al., 2000; Schueler et al., 2001, 2005; Miga et al., 2014) : their transpositional activity and insertion at centromeres could be promoted by some rearrangement-induced cellular stress. The subsequent decreased in centromeric sequence homogeneity and integrity could affect centromere function (Ugarković, 2009) and newly amplified arrays of "clean" alpha satellites could be selected within centromeric regions.

2.3.5 Conclusion

During the course of the Cercopithecini evolution, alpha satellite DNA diversified in families of sequences. These families emerged successively along lineages and adopt distinct distributions on centromeric regions that are largely in accordance with the age-gradient hypothesis. Alpha satellite DNA diversification and displacement from centromeres to pericentromeres appear to be correlated with chromosome rearrangement dynamics and ENC formations. Most of the putative ENCs present in highly fissioned karyotype species are spanned by the youngest family characterized. In addition, the species distribution and location of alpha satellite families on centromeric regions provided information regarding Cercopithecini relationships. Further integration of satellite DNA into extensive comparative studies, using deep sequencing combined to classic and molecular cytogenetics, should keep improving our understanding of the evolutionary dynamics of centromere tandem repeats.

2.3.6 Methods

Phylogenetic reconstruction

Sequences of nine nuclear genes, classically employed in primate phylogenies (ABCA1, BRCA2, CFTR, DENND5A, ERC2, LRPPRC-169, SRY, TTR and ZFX) (Perelman et al., 2011), were used to reconstruct phylogenetic relationships between 22 Cercopithecini species (*Miopithecus ogouensis*, *Allenopithecus nigroviridis*, *Erythrocebus patas*, *Chlorocebus sabaeus*, *C. aethiops*, *Cercopithecus solatus*, *C. preussi*, *C. lhoesti*, *C. hamlyni*, *C. roloway*, *C. diana*, *C. mona*, *C. wolffii*, *C. pogonias*, *C. campbelli*, *C. cephushus*, *C. petaurista*, *C. erythrotis*, *C. ascanius*, *C. nictitans*,

C. mitis and *C. albogularis*). When not retrieved from Genbank, these sequences were originally obtained using fibroblast samples provided by the Collection of cryopreserved living tissues and cells of vertebrates (RBCell collection, Muséum national d'Histoire naturelle, Paris). DNA was extracted from these samples using QIAGEN QIAamp DNA Mini Kit. Genes were amplified using PCR primers available in Supplementary table 3.3. The PCR consisted of 40 cycles : 30 s at 94 °C, 40 s at 57 °C (ABCA1) or 63 °C (ZFX) or 65 °C (BRCA2, LRPPRC-169, TTR) or 67 °C (CFTR, DENND5A, ERC2) and 1 min 20 s at 72 °C. The double-stranded PCR products were sequenced at Eurofins Genomics (Germany). Genbank accession numbers for sequences used here are available in Supplementary table 3.4.

Sequences were aligned using MUSCLE (Edgar, 2004). The phylogenetic tree was constructed using Bayesian Markov chain Monte Carlo (MCMC) phylogenetic analysis with MRBAYES 3.1 (Huelsenbeck et al., 2001). The computer program PartitionFinder (Lanfear et al., 2012) was used to evaluate the fit of 56 nested models of nucleotide substitution to the data. The models chosen by PartitionFinder according to the Bayesian Information Criterion were then used in the Bayesian analysis. Three heated chains and a single cold chain were employed in the Bayesian analysis, and runs were initiated with random trees. Two independent runs were conducted with 10 million generations per run ; trees (and parameters) were sampled every 1000 generations. Our phylogenetic tree was rooted with 4 species closely related to Cercopithecini (Perelman et al., 2011; Guschanski et al., 2013) : 3 Papionini species (*Macaca sylvanus*, *Mandrillus sphinx* and *Cercocebus torquatus*) and 1 Colobinae species (*Colobus guereza*).

Cell cultures, metaphase preparations and chromosome banding

Fibroblast samples from one Papionini species (*M. sylvanus*), one Colobinae species (*Colobus angolensis*) and 13 Cercopithecini species (*A. nigroviridis*, *E. patas*, *C. solatus*, *C. lhoesti*, *C. roloway*, *C. diana*, *C. mona*, *C. pogonias*, *C. cephicus*, *C. erythrotis*, *C. ascanius*, *C. nictitans* and *C. mitis*) were obtained from the Collection of cryopreserved living tissues and cells of vertebrates (RBCell collection, Muséum national d'Histoire naturelle, Paris). IDs of the specimens are available in Supplementary table 3.5. Cells were grown at 37 °C in D-MEM (Gibco), supplemented with 10 % SVF (Dutscher). FUdR (fluorodeoxyuridine, Sigma ; final concentration 0.06 mg/mL) was used overnight to synchronize cultures. BrdU (5'-bromodeoxyuridine, Sigma ; final concentration 0.02 mg/mL), added for the last 8 h of culture, was incorporated in late-replicating DNA. Colchicine (0.04 mg/mL) was added for the last 3 h to stop the cell cycle at metaphase. Hypotonic shock with FBS (fetal bovine serum, 1/6) and KCl (187 mg/mL) was used for 15 min. Cells were then fixed in a solution of acetic acid/ethanol (25 %/75 %), spread on cold slides, dried and stored at -20 °C.

FISH experiments

FISH were performed on metaphase chromosome preparations with short oligonucleotide probes designed to target specifically the C1 to C6 alpha satellite families identified in the *C. solatus* and/or *C. pogonias* genomes ([Cacheux et al., 2016a,b](#)). Probe sequences are available in Supplementary table 3.1, which also provides details about the positions of locked nucleic acid (LNA) modifications in the probes. Supplementary table 3.2 resumes which probes were used according to species. LNA-modified probes were purchased from Eurogentec (Seraing, Belgium).

Hybridization solutions were prepared by diluting the oligonucleotide probes to a final concentration of 0.1 µM in a hybridization solution consisting of 2X SSC pH 6.3, 50 % deionized formamide, 1X Denhardt solution, 10 % dextran sulfate, and 0.1 % SDS. 20 µL of the hybridization solution were deposited on each slide and covered with a coverslip. The slides were then heated for 3 min at 70 °C and hybridized for 1 h at 37 °C in a Thermobrite apparatus (Leica Biosystems). Then, each slide was washed twice in 2X SSC at 63 °C or occasionally at 68 °C (see Results). Preparations were then incubated in blocking solution (4 % bovine serum albumin (BSA), 1X PBS, 0.05 % Tween 20) for 30 min at 37 °C to reduce nonspecific binding. Then, depending on the combination of probes, the following antibodies were used for subsequent revelations : Alexa 488-conjugated streptavidin (1 :200 ; Life Technologies), Cy5-conjugated streptavidin (1 :200 ; Caltag Laboratories), FITCconjugated sheep anti-digoxigenin (1 :200 ; Roche), and Rhodamine-conjugated sheep anti-digoxigenin (1 :200 ; Roche). All antibodies were diluted in blocking solution containing 1X PBS, 0.05 % Tween 20, and 4 % BSA. Antibody incubation lasted for 30 min at 37 °C. All washings were performed in 2X SSC, 0.05 % Tween 20. Chromosomes were counterstained in blue with DAPI (4',6-diamidino-2-phenylindole, 5 µg/mL) and slides were mounted by adding a drop of Vectashield Antifade Mounting Medium (Vector Laboratories), or in red with PI (propidium iodide, 1 µg/mL) and slides were mounted by adding a drop of home-made PPD11 (p-phenylenediamine-11) when RBP chromosome banding was needed for further chromosome identification.

Image acquisition and analysis

Metaphases were imaged using an Axio Observer Z1 epifluorescent inverted microscope (Zeiss) coupled to an ORCA R2 cooled CDD camera (Hamamatsu). The Axio Observer Z1 was equipped with a Plan-Apochromat 63× 1.4 NA oilimmersion objective and the following filters set : 49 shift free for DAPI (G365 / FT395 / BP 445/50), 38 HE shift free for FITC/Alexa488 (BP470/40 / FT495 / BP525/50), homemade sets for Rhodamine (BP546/10 / FF555 / BP 583/22) and for Cy5 (BP643/20 / FF660 / BP684/24). The light source was LED illumination (wavelengths : 365 nm, 470 nm or 625 nm) except for Rhodamine, for which a metal halide lamp HXP120 was preferred. Immersion oil of refractive index 1.518 at 23 °C was used. Color-combined images were reconstructed using ImageJ ([Abràmoff et al., 2004](#)). When chromosome identification was

needed, metaphases were imaged using an epifluorescent microscope (Microphot-FXA, Nikon) and images were captured using a cooled CCD camera (ProgRes MFcool, Jenoplak). Metaphases were karyotyped using the Isis 5.3 software (Metasystems, Altussheim, Germany).

2.3.7 Additional files

Supplementary figure 3.1 : FISH analysis of the C3 and C4 alpha satellite families in species from the terrestrial clade

Supplementary figure 3.2 : FISH analysis of the C5 alpha satellite family in species from the terrestrial clade

Supplementary figure 3.3 : FISH analysis of the C1 alpha satellite family on *A. nigroviridis* chromosomes

Supplementary figure 3.4 : FISH analysis of the C2 alpha satellite family on *M. sylvanus* chromosomes

Supplementary figure 3.5 : FISH analysis of the C3 and C4 alpha satellite families on *M. sylvanus* chromosomes

Supplementary figure 3.6 : Alignment of homologous chromosomes between Cercopithecini species

Supplementary figure 3.7 : Distribution pattern of the C2 alpha satellite family on Cercopithecini chromosomes

Supplementary figure 3.8 : Distribution pattern of the C5 alpha satellite family on Cercopithecini chromosomes

Supplementary figure 3.9 : Distribution pattern of the C6 alpha satellite family on *C. cephushispidus* chromosomes

Supplementary table 3.1 : Patterns and labels of oligonucleotide probes used by FISH

Supplementary table 3.2 : Oligonucleotide probes used by FISH according to species

Supplementary table 3.3 : PCR primers used for phylogenetic studies

Supplementary table 3.4 : Genbank sequences included in phylogenetic studies

Supplementary table 3.5 : RBCell collection specimens included in cytogenetic studies

Chapitre 3

Discussion et conclusion

3.1 Méthodologie d'étude de l'ADN alpha satellite

3.1.1 Isolement, séquençage et analyse bioinformatique

Nos travaux ont permis la mise au point d'une nouvelle approche, simple et efficace, pour étudier l'ADN alpha satellite. Celle-ci repose sur l'isolement de séquences alpha satellites à partir d'un génome, puis sur leur soumission aux nouvelles technologies de séquençage ; des dizaines de milliers de séquences alpha satellites peuvent ainsi être obtenues de manière rapide et peu coûteuse. L'utilisation d'une enzyme de restriction pour l'isolement des séquences et de la technologie Ion Torrent, capable de séquencer jusqu'à 400 paires de bases contigües, permet d'obtenir des monomères dans leur intégralité et dans la même phase enzymatique. Ceux-ci peuvent ainsi être alignés sur la totalité de leur longueur et facilement comparés entre eux, notamment via la génération subséquente d'arbres phylogénétiques. Cette génération d'arbres, classiquement utilisée pour analyser la diversité des séquences alpha satellites, ne peut cependant porter que sur un nombre limité de quelques centaines de séquences et n'est pas appropriée à l'analyse primaire de gros jeux de données. Ce problème peut être résolu par l'utilisation parallèle d'analyses en composantes principales, ou ACP, reposant sur la composition en k-mers nucléotidiques des séquences ; elles sont capables d'intégrer des dizaines de milliers de monomères tout en permettant une visualisation aisée de leur structuration en familles. Le clustering des familles observées et la projection du résultat de ce clustering sur un arbre phylogénétique facilitent la lecture de l'arbre ; ils permettent également d'évaluer l'accord entre ACP et phylogénie quant à l'existence de familles d'alpha satellites au sein du jeu de données analysé.

Cette approche méthodologique présente en revanche plusieurs biais, dont un biais de sélection quant aux monomères séquencés : ceux ne portant pas le site de restriction de l'enzyme employée ne pourront être isolés. Par ailleurs, des monomères situés dans la même phase enzymatique ne

peuvent être assemblés suite au séquençage, ce qui est un frein à l'étude de l'organisation des familles d'alpha satellites. A ces deux problèmes, notre étude apporte plusieurs solutions. D'une part, des dimères d'alpha satellites peuvent être séquencés, bien qu'une moindre proportion de séquences dimériques puisse apparemment être récupérée comparativement aux séquences monomériques. Il est possible que les séquences dimériques forment des structures secondaires particulières affectant l'amplification clonale préalable au séquençage Ion Torrent, voire le séquençage lui-même. Les dimères obtenus permettent cependant l'étude de deux monomères consécutifs et donnent ainsi des informations sur l'organisation, monomérique ou en HORs, des familles ; mais ils permettent également la caractérisation de familles associées en HORs et ne présentant pas de site de restriction entre elles, et donc absentes des jeux de données monomériques, comme observé pour les familles C3 et C4 (Sections 2.1.3 et 2.2.3).

D'autre part, il est possible d'étudier l'organisation des familles identifiées en réitérant la méthode, mais en utilisant une seconde enzyme de restriction afin de récupérer des monomères chevauchants les premiers. Les monomères issus de phases enzymatiques différentes sont difficilement comparables par phylogénie, puisqu'alors seules les parties chevauchantes s'alignent et sont utilisées ; en revanche, de tels monomères sont aisément comparables par ACP puisque ce sont leurs compositions nucléotidiques globales qui sont alors utilisées. Ainsi, les monomères d'une famille adoptant une organisation monomérique vont co-localiser sur la projection de telles analyses, quelle que soit leur phase enzymatique, comme observé pour les familles C1, C2 et C5 (Section 2.2.3). Au contraire, l'implication d'une famille dans une organisation en HORs peut être inférée lorsqu'un cluster de séquences d'une taille égale mais obtenu dans une phase enzymatique différente se localise à proximité de celle-ci sur la projection, comme observé pour la famille C6. Dans les deux cas, la comparaison des consensus des familles identifiées dans une première phase avec les consensus de leurs clusters chevauchants identifiés dans la seconde phase est une étape importante pour appuyer l'organisation inférée. Enfin, l'obtention de séquences alpha satellites via une seconde enzyme permet également d'identifier des familles non digérées par la première, et ainsi d'étendre la connaissance de la diversité de l'ADN alpha satellite au sein d'un génome ; ici, les séquences C6' ont révélé la présence d'une famille supplémentaire dans le génome de *C. pogonias*, qui reste cependant à caractériser lorsque l'on se place dans le référentiel de phase Xmnl (Section 2.2.3).

3.1.2 Analyse cytogénétique

Notre approche méthodologique offre une vision globale de la diversité de l'ADN alpha satellite au sein d'un génome, et laisse donc espérer la possibilité d'un ciblage par FISH spécifique des familles d'alpha satellites identifiées. Cette révélation est importante : elle permet de montrer que les séquences d'une même famille sont présentes en abondance dans une voire des régions localisées du génome, et font donc possiblement partie d'organisations monomériques ou en HORs. Elles seraient donc issues d'évènements d'amplification depuis une même séquence ancestrale : la similarité observée *in silico* traduirait bien une proximité phylogénétique, non une

convergence entre séquences.

De courtes sondes de 18 ou 19 oligonucléotides, constituées d'une alternance de nucléotides classiques et de nucléotides modifiés LNA, sont tout à fait capables de s'hybrider à l'ADN ; ne portant qu'un fluorophore chacune, elles ne peuvent révéler que des séquences présentes en abondance dans une ou des régions localisées du génome (Ollion et al., 2015). Nos travaux montrent que ces sondes s'avèrent tout à fait aptes à discriminer des ensembles de séquences homologues ne différant, au niveau du site ciblé, que de deux nucléotides en moyenne. Des sondes compétitrices peuvent cependant être utilisées pour exclure l'hybridation aspécifique d'une sonde sur un ensemble de séquences ; des ensembles de séquences homologues ne différant que d'un nucléotide en moyenne, au niveau du site ciblé, peuvent alors être discriminés.

Nous avons mentionné plus avant que notre méthode de caractérisation des familles d'alpha satellites était possiblement non exhaustive, les monomères dénués du site de restriction de l'enzyme employée ne pouvant être récupérés ; ce biais trouve sa continuité dans l'analyse cytogénétique des familles. En effet, des sondes discriminantes ne peuvent être dessinées qu'à partir de l'analyse des familles connues ; les familles non recueillies par l'isolement enzymatique ne peuvent être prises en compte. Il est ainsi possible que certaines sondes, dessinées pour cibler spécifiquement une famille identifiée, ciblent également d'autres familles absentes de nos jeux de données *in silico*, comme observé pour la sonde C3a (Section 2.2.3). A cet inconvénient, une solution est apportée : dans la mesure du possible, deux sondes peuvent être utilisées pour cibler la même famille sur des sites différents. Une colocalisation de leurs signaux respectifs va en faveur de la détection spécifique de la famille en question, tandis qu'un non chevauchement de leurs signaux traduit la détection d'au moins deux familles, différentes entre elles et différentes de la famille ciblée, comme observé avec les sondes C1a et C1b chez *A. nigroviridis* (Section 2.3.3). Si de telles sondes révèlent en revanche des signaux respectifs se chevauchant partiellement, on peut envisager que la famille ciblée se distribue sur les loci de chevauchement. Enfin, des familles non identifiées mais également non ciblées par les sondes utilisées dans le cadre d'une étude pourraient être révélées par FISH, si le profil d'hybridation d'une sonde pan-alpha satellites, comme la sonde Cx (Section 2.1.3), s'avérait plus large que le profil d'hybridation combiné des sondes ciblant les familles caractérisées.

3.2 Diversité et évolution de l'ADN alpha satellite chez les Cercopithèques

3.2.1 Pluralité de familles et d'organisations

L'ADN alpha satellite de *C. solatus* et *C. pogonias* a été analysé, et les résultats de ces analyses interprétés en prenant les précautions nécessaires liées aux différents biais mentionnés plus avant

(Sections 2.1.3 et 2.2.3). Nos travaux ont ainsi révélé la présence de plusieurs familles d'alpha satellites distinctes au sein de leurs génomes, remettant en question l'idée admise d'une diversité limitée de l'ADN alpha satellite chez les cercopithèques (Goldberg et al., 1996; Yoda et al., 1996; Alexandrov et al., 2001). De même, la détection de familles ou sous-familles sur certains chromosomes seulement, chez l'une comme l'autre des espèces, appelle à une réévaluation de l'idée admise quant à la prévalence de l'homogénéisation interchromosomique des alpha satellites chez les singes de l'Ancien et du Nouveau Monde ; celle-ci y était pensée suffisamment intense pour annihiler la diversité interchromosomique de l'ADN alpha satellite (Alkan et al., 2007).

Nous avons par ailleurs montré que certaines de ces familles étaient impliquées dans des organisations en HORs d'ordre 2 ou plus, tandis que l'absence de telles organisations était acceptée chez les cercopithèques (Goldberg et al., 1996; Yoda et al., 1996; Alexandrov et al., 2001). La première, constituée des familles C3 et C4, est hétérogène et localisée dans les péricentromères de toutes les espèces de cercopithèques étudiées, à l'exception de *A. nigroviridis* (Section 2.3.3). Une telle HOR, dont les caractéristiques tranchent avec celles communément admises pour ce type d'organisation, a précédemment été observée chez l'homme seulement et sur un unique chromosome (Shepelev et al., 2009). Les HORs hétérogènes péricentromériques seraient en fait plus largement distribuées chez les Primates qu'envisagé jusqu'ici. La seconde HOR, impliquant la famille C6, est plus proche des HORs classiques puisqu'a priori très homogène et reposant sur les centromères d'un nombre restreint de chromosomes (Section 2.2.3). Il sera intéressant de caractériser cette HOR dans son intégralité, étant donné qu'elle est la première de ce type à être identifiée chez les singes de l'Ancien Monde.

Au-delà de cette diversité intragénomique de l'ADN alpha satellite, notre étude a également montré une diversité interspécifique de cet élément chez les cercopithèques. Certaines familles d'alpha satellites sont conservées chez toutes les espèces étudiées, tandis que d'autres familles y sont différemment distribuées (Sections 2.2.3 et 2.3.3). De même, l'HOR homogène détectée chez *C. pogonias* serait présente dans un nombre limité d'espèces de la lignée arboricole, et l'existence d'une telle organisation dans le génome des autres espèces reste à déterminer. Enfin, l'observation d'hétérozygotes pour les familles C5 et C6, sur certaines paires de chromosomes, suggère également une diversité intraspécifique de l'ADN alpha satellite chez les cercopithèques (Section 2.3.3) ; celle-ci serait intéressante à étudier dans le prolongement de nos travaux, bien qu'une telle étude puisse souffrir de l'accès difficile aux cellules vivantes de plus de quelques individus de la même espèce.

3.2.2 Dynamique d'émergence et d'extinction

L'ensemble de nos travaux ont permis d'inférer, au cours de l'évolution des cercopithèques, l'émergence des familles d'alpha satellites caractérisées. Les familles C1, C2, C5 et C6 seraient apparues successivement depuis la divergence de ce clade, i.e. durant les dix derniers millions d'années (Guschanski et al., 2013). Les familles C3 et C4 seraient en revanche apparues il y a

douze millions d'années, lors de la divergence des Cercopithecinae (Perelman et al., 2011). Les familles les plus anciennes (C2, C3, C4) sont retrouvées sur les péricentromères de la plupart des espèces et sont hétérogènes en séquences chez *C. solatus* comme *C. pogonias*; tandis que les familles plus récentes (C1, C5 et C6) sont plus homogènes et se distribuent généralement sur les centromères. L'hypothèse du gradient d'âge de Schueler et al. (2001, 2005), proposée à partir de l'analyse de quelques péricentromères humains, se trouve appuyée par notre étude à grande échelle; les familles d'alpha satellites récentes apparaissent occuper les centromères suite au déplacement des familles plus anciennes vers les péricentromères. Chez *A. nigroviridis*, qui ne possède pas les familles C1, C5 et C6, les anciennes familles sont restées lovées au niveau des centromères (Section 2.3.3); il serait intéressant de séquencer et d'analyser l'ADN alpha satellite de cette espèce afin de déterminer si elles y sont plus homogènes que chez les espèces où elles ont été déplacées vers les péricentromères. De même, l'on peut se demander si la famille C5 ne serait pas plus homogène chez *C. diana* et *C. roloway* que chez *C. pogonias*, puisque sa distribution apparaît plus centromérique chez ces deux premières espèces.

Par ailleurs, nos travaux ont mis en évidence que les anciennes familles d'alpha satellites C3 et C4 seraient probablement en voie d'extinction dans les lignées terrestre et arboricole des cercopithèques. Difficilement décelables par FISH, elles ne seraient plus représentées que par un faible nombre de copies qui pourraient totalement disparaître si de nouvelles familles venaient encore à émerger dans les génomes. Les mécanismes responsables de ce phénomène d'érosion des séquences satellites sur les bords des péricentromères restent à caractériser. Les familles C3 et C4 ne semblent en revanche pas menacées chez *A. nigroviridis* (Section 2.3.3); les familles d'alpha satellites n'auraient pas une durée de vie équivalente quelle que soit la lignée évolutive considérée. De plus, certains loci, comme les péricentromères des bras courts des acrocentriques, peuvent assurer la persistance de certaines familles; la famille C2 par exemple y apparaît très amplifiée dans des espèces, comme *C. pogonias*, où elle serait pourtant en voie d'extinction sur les péricentromères des méta- et submétacentriques (Section 2.2.3).

D'après notre étude, l'évolution des familles d'alpha satellites dépendrait en fait largement de l'évolution des chromosomes, tout du moins chez les cercopithèques (Section 2.3.3). D'une part, les familles C2, C3, C4 et C5 semblent en effet être repoussées vers les péricentromères à mesure que les chromosomes fissionnent et que les nouveaux centromères évolutifs émergent dans les génomes; l'existence de ces nouveaux centromères évolutifs, ici inférés comme tels grâce à la comparaison du profil en bandes de chromosomes homologues, reste cependant à confirmer via de nouvelles méthodes comme la BAC-FISH (Ventura et al., 2001). D'autre part, les familles C1, C5 et C6 se forment à mesure que ces remaniements se produisent et la famille C6 se distribue presque exclusivement sur les présumés nouveaux centromères évolutifs des espèces aux caryotypes les plus fissionnés; il est donc possible qu'au-delà de leurs déplacements sur les régions centromériques, la formation elle-même des familles d'alpha satellites soit entraînée par l'évolution des chromosomes. Il faut cependant garder à l'esprit que de nombreuses autres familles que C1 à C6 sont éventuellement apparues chez les cercopithèques, et qu'une caractérisation plus complète des familles d'alpha satellites au sein de ce clade, notamment via

l'application de notre méthode à d'autres espèces que *C. solatus* et *C. pogonias*, permettra d'y obtenir une vision plus exhaustive de la dynamique évolutive de l'ADN alpha satellite et des facteurs attachés à cette évolution.

3.2.3 Histoire évolutive des Cercopithèques

Tandis que les cercopithèques se sont révélés être d'excellents modèles pour étudier la dynamique de diversification de l'ADN alpha satellite, celui-ci est à son tour apparu comme un caractère de choix pour mieux comprendre l'évolution de ce clade (Section 2.3.3). D'une part, la distribution des familles d'alpha satellites entre espèces et sur les régions centromériques supporte la monophylie controversée de différents groupes ([Tosi, 2008](#); [Guschanski et al., 2013](#)). Ainsi, l'absence de la famille C1 chez *A. nigroviridis* et sa présence chez les autres espèces, corrélatées à la distribution centromérique et péricentromérique des familles C2, C3 et C4, confortent l'existence d'un clade regroupant les espèces des lignées terrestre et arboricole ainsi qu'une position de la lignée d'*A. nigroviridis* à l'extérieur de ce clade. La présence de la famille C6 et la distribution péricentromérique de la famille C5 supportent par ailleurs l'existence d'un clade interne au clade arboricole, comprenant les espèces des groupes de *C. mona*, de *C. mitis* et de *C. cephush*.

D'autre part, la distribution de la famille C5 dans un ensemble paraphylétique d'espèces, comprenant les espèces du clade arboricole et une unique espèce du clade terrestre, va en faveur d'une hypothèse précédemment émise quant à des introgressions génétiques ancestrales allant du clade arboricole vers le clade terrestre ([Guschanski et al., 2013](#)). De nouvelles caractérisations de l'ADN alpha satellite seraient tout à fait appropriées pour continuer à éclairer les relations de parenté chez les cercopithèques. [Shepelev et al. \(2009\)](#) présentaient les centromères comme des outils phylogénétiques prometteurs ; de plus amples intégrations de l'ADN alpha satellite dans l'étude de l'origine des espèces pourraient en effet révéler, grâce à sa dynamique et ses mécanismes particuliers de diversification, des informations tout à fait inédites sur l'histoire évolutive des Primates.

3.3 Conclusion

Nos travaux ont permis de proposer une nouvelle approche méthodologique, rapide et peu coûteuse, pour étudier la diversité de l'ADN alpha satellite au sein d'un génome et plus généralement des séquences d'ADN répétées en tandem. Le séquençage nouvelle génération ainsi que l'analyse bioinformatique et cytogénétique d'alpha satellites ont mis en évidence la présence de plusieurs familles de séquences alpha satellites dans les génomes des cercopithèques, apparues successivement au cours de l'évolution de ce clade et adoptant des organisations monomériques comme des organisations en répétitions d'ordre supérieur. Certaines de ces familles

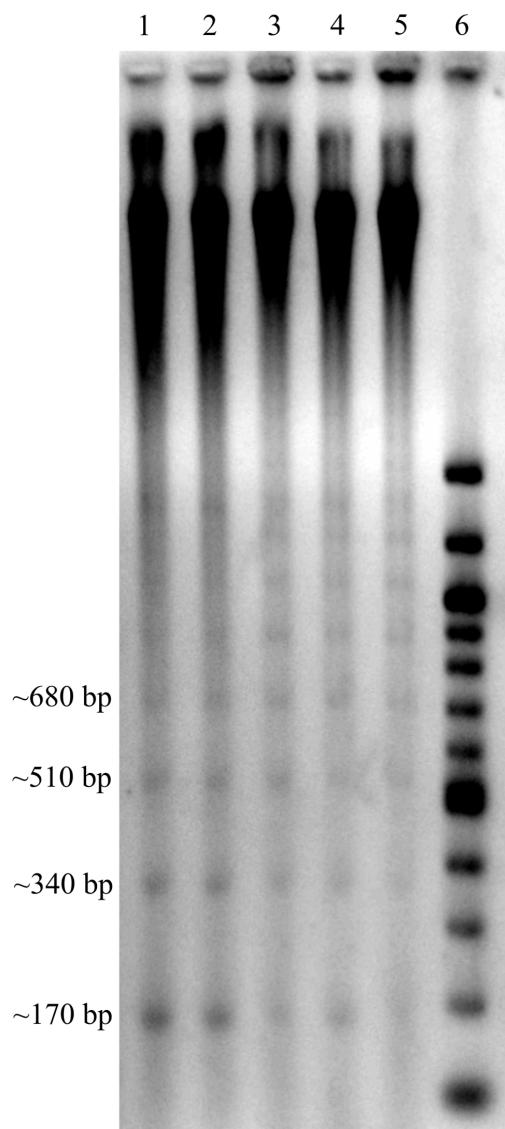
se distribuent différemment entre chromosomes, révélant l'existence d'une diversité inter-chromosomique de l'ADN alpha satellite chez les singes de l'Ancien Monde. Leurs positions sur les régions centromériques vont par ailleurs en faveur de l'hypothèse du gradient d'âge des alpha satellites, selon laquelle les familles se forment aux centromères en déplaçant les familles préexistantes vers les péricentromères. L'ADN alpha satellite et les chromosomes évolueraient enfin de manière concertée chez les cercopithèques, le premier se diversifiant et se déplaçant sur les régions centromériques à mesure que les seconds se fissinent et voient l'émergence de nouveaux centromères évolutifs. L'ADN alpha satellite a par ailleurs apporté des informations nouvelles quant aux relations de parenté chez les cercopithèques, invitant à l'intégration de cette famille de séquences dans l'étude de l'histoire évolutive des Primates.

Chapitre 4

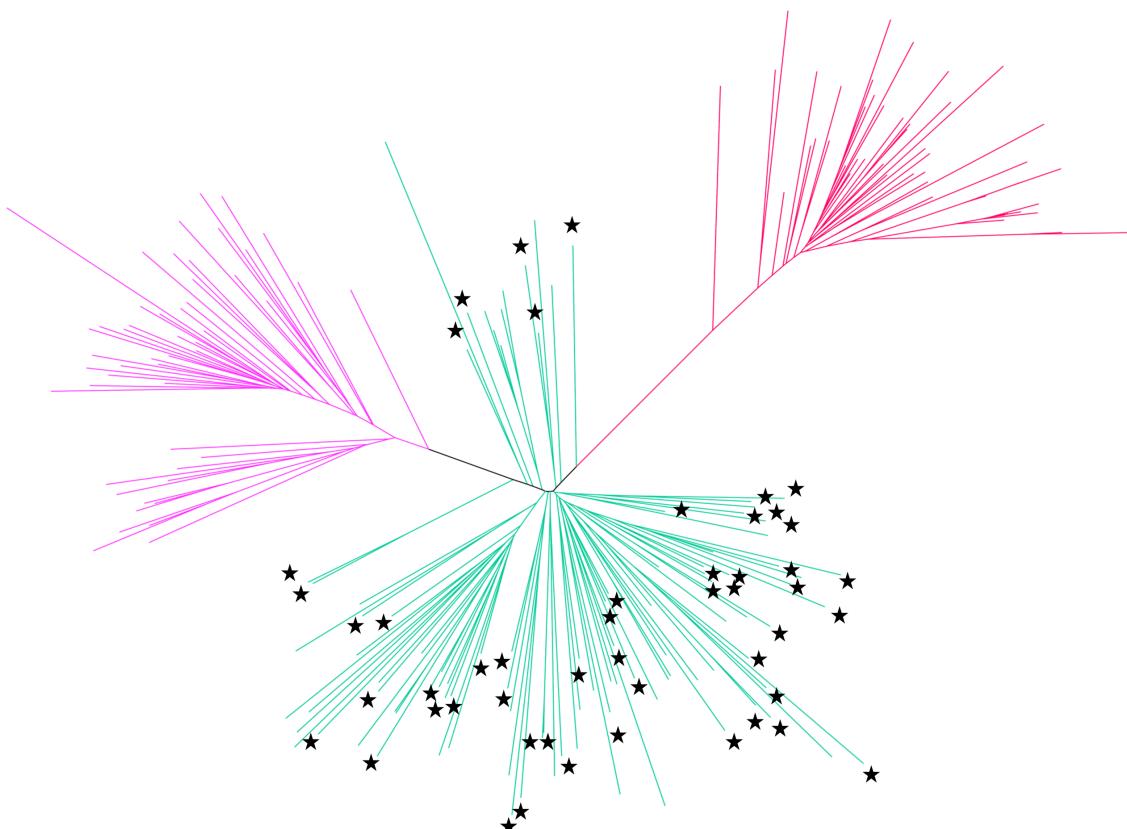
Matériel supplémentaire

4.1 Matériel supplémentaire : premier article

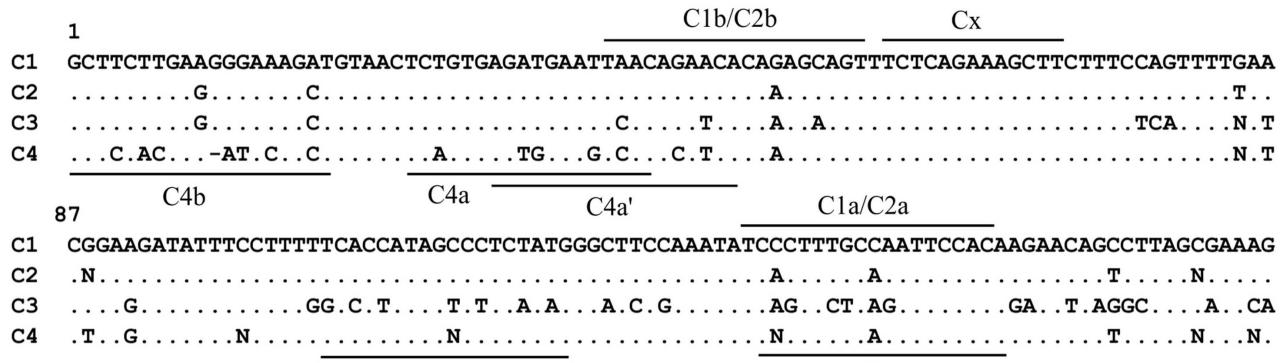
4.1.1 Figures supplémentaires



SUPPLEMENTARY FIGURE 1.1 – Migration profiles of *C. solatus* genomic DNA digested with **XmnI.** *Cercopithecus solatus* genomic DNA was digested for 1 h at 37°C with the enzyme XmnI (New England Biolabs) in a total volume of 10 µL. The enzyme was inactivated for 20 min at 65 °C. The samples were loaded on a 1.5 % agarose gel after addition of 2 µL loading buffer (50 % glycerol) and electrophoresis was performed in 0.5X Tris-borate-EDTA buffer, at room temperature for 2 h 45 min at 100 V. 1 : 0.5 µg of genomic DNA digested with 20 units of XmnI activity ; 2 : 0.5 µg of genomic DNA digested with 10 units of XmnI activity ; 3 : 0.5 µg of genomic DNA digested with 5 units of XmnI activity ; 4 : 0.5 µg of genomic DNA digested with 1.5 units of XmnI activity ; 5 : 0.5 µg of genomic DNA digested with 1 unit of XmnI activity ; 6 : 100 bp DNA Ladder (New England Biolabs).



SUPPLEMENTARY FIGURE 1.2 – Phylogenetic tree for left and right alpha satellite monomers from *C. solatus* dimer dataset. The phylogenetic tree (Neighbor-joining method, K2P model) was generated following a MUSCLE alignment ([Edgar, 2004](#)) between 50 randomly selected C3 (dark pink) monomers, 50 C4 (light pink) monomers, 50 C2 (pastel green) left monomers and 50 C2 (pastel green) right monomers. C2 right monomers are pointed out with a black star so we can observe that they mix well with the C2 left monomers. Only dimers without the XmnI site (noX dataset) were considered in this analysis.

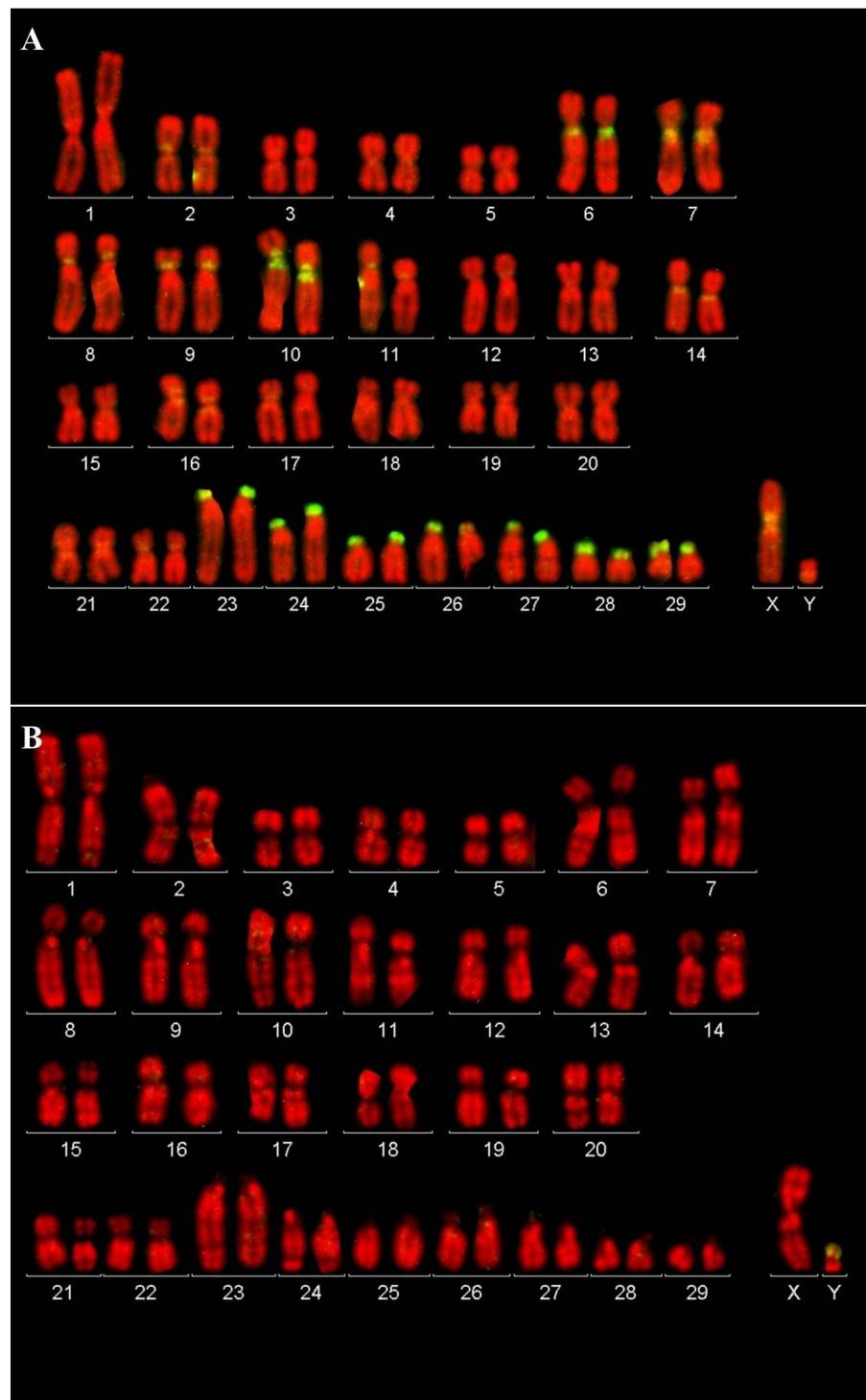


(a)

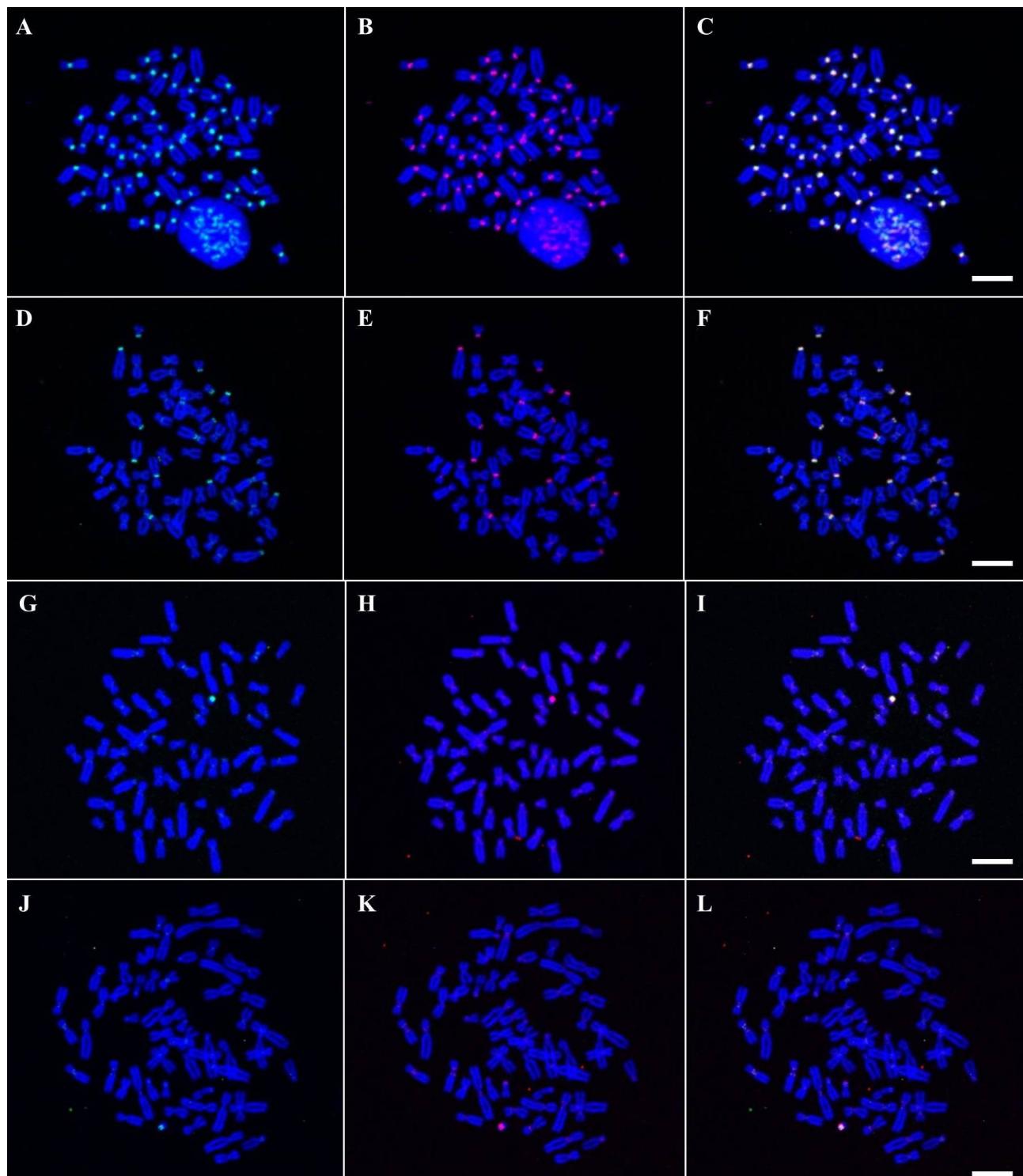
Probe	Pattern (5'-3')	Label	Mismatch	C1 (%)	C2	C3	C4
C1a	TcCCtTtGcCaAtTcCAC	3'Cy3	0	61	-	-	2
			1	86	6	-	16
C1b	AcTgCtCtGtGtTcTGtTa	3'Digoxygenin	0	60	1	-	-
			1	89	28	2	-
C2a	TcACtTtGcAaAtTcCAC	5'AlexaFluor488	0	-	17	-	8
			1	2	51	-	34
C2b	AcTgCtTtGtGtTcTGtTa	5'Cy5	0	1	24	2	-
			1	62	56	2	-
C3a	CaGtTcTcAGaTtCcAcA	3'Digoxygenin	0	-	-	28	-
			1	-	-	53	-
C3b	GcCcTaTaGtCtTcAaAg	3'Biotin	0	-	-	17	-
			1	-	-	56	-
C4a	GtGaCtTcCACtTcAcTgA	3'Biotin	0	-	-	-	34
			1	-	-	-	69
C4a'	TgAtGtGtGaCtTcCACt	3'Digoxygenin	0	-	-	-	29
			1	-	-	-	54
C4b	aGcTGtATtTcGTgGaGc	3'Biotin	0	-	-	-	13
			1	-	-	-	32
Cx	tctcagaaaggctt	3'Biotin	0	77	40	51	40
			1	93	74	79	75

(b)

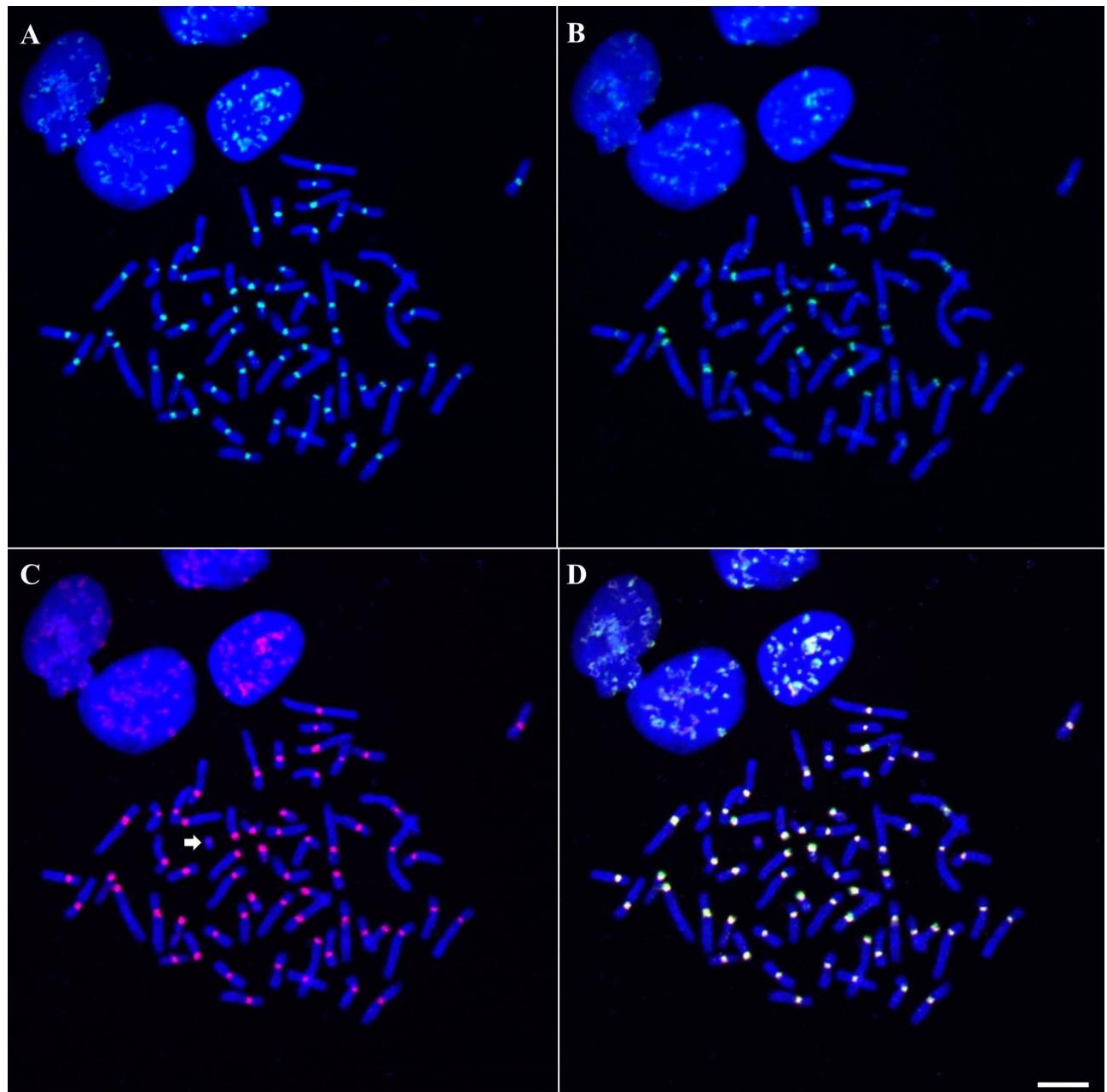
SUPPLEMENTARY FIGURE 1.3 – LNA-modified probes used to target the C1 to C4 alpha satellite families on *C. solatus* chromosomes. (A) Probe binding sites on C1, C2, C3 and C4 consensus. (B) Probe sequences and labels, associated to the percentage of targeted sequences within the C1 to C4 families. LNA are written in lower case and classic nucleotides are written in upper case. Percentage calculations have been made using all C1 and C2 sequences from the monomer dataset and all C3 and C4 sequences from the (noX) dimer dataset, when no mismatch or one mismatch was present. The sign “-“ means 0.



SUPPLEMENTARY FIGURE 1.4 – Distribution pattern of the C2 and C3 alpha satellite families on *C. solatus* chromosomes. Metaphase chromosomes are counterstained with propidium iodide (2 µg/mL) and shown here in red, while the green colors stand for probes targeting (A) the C2 family (C2a) or (B) the C3 family (C3a). C2 signals on the long arm of chromosome 2 (right) and 11 (left) are due to chromosome superposition in the original metaphase. RBG chromosome banding techniques were employed as in Moulin et al. (2008). Observation was performed with an epifluorescent microscope (Microphot-FXA, Nikon) and images were captured using a cooled CCD camera (ProgRes MFcool, Jenoplik). The metaphases were karyotyped using the Isis 5.3 software (Metasystems, Altusshausen, Germany) according to Dutrillaux et al. (1988a).



SUPPLEMENTARY FIGURE 1.5 – Comparison of FISH signals from different probes targeting identical alpha satellite families on *C. solatus* chromosomes. Probe mixtures were hybridized as indicated to *C. solatus* metaphase chromosomes, which are colored in blue. Probe signal is shown in green (*A, D, G and J*) or red (*B, E, H and K*). Images shown in (*C, F, I and L*) are superpositions of the two previous images on the same lane. Probe mixtures are the following ones : (*A,B,C*) C1a (green), C1b (red) and C2b (not shown), (*D,E,F*) C2a (green), C2b (red) and C1b (not shown), (*G, H, I*) C3a (green) and C3b (red), (*K,L,M*) C4a' (green) and C4b (red). Scale bar = 10 µm.



SUPPLEMENTARY FIGURE 1.6 – Comparison of the hybridization pattern of probe Cx with those of probes targeting the C1 to C4 alpha satellite families on *C. solatus* chromosomes. Probes Cx, C1b and C2b were hybridized simultaneously to *C. solatus* metaphase chromosomes, which are colored in blue. The displayed signals correspond to (A) probe C1b (shown in green), (B) probe C2b (shown in green), (C) probe Cx (shown in red), and (D) the overlap of the three signals with the same colors. Cx displays colocalization with C1b and C2b and also labels the single chromosome they do not label (arrow), which is the single chromosome strongly labeled by probes C3a/b and C4a/b. Therefore, the Cx signal overlaps with the signals of probes targeting the C1 to C4 families without extending beyond their global signal. Scale bar = 10 μm .

4.1.2 Tables supplémentaires

SUPPLEMENTARY TABLE 1.1 – Filtering steps from *C. solatus* raw data to alpha satellite monomer and dimer datasets

Filters	Monomers	Dimers
None	204,990	353,683
1. Quality	136,360	155,900
2. Extremity	119,134	81,693
3. Length	104,962	18,364
4. Alpha satellite	100,713	3,568

NOTE - The numbers of sequences fitting with the filtering criteria are given (see Methods).

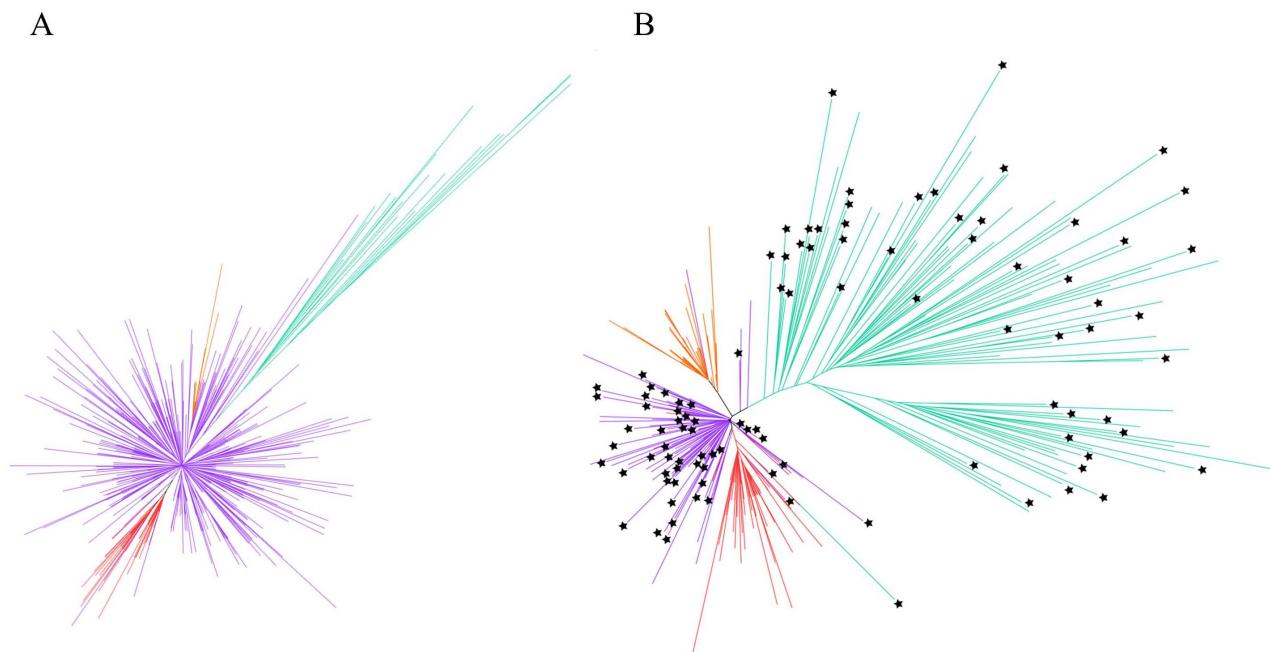
SUPPLEMENTARY TABLE 1.2 – Alpha satellite family associations in *C. solatus* dimer dataset

A noX dataset						B X dataset					
Left	Right					Left	Right				
	C1	C2	C3	C4	Total		C1	C2	C3	C4	Total
C1	-	-	-	-	0	C1	76	8	-	0	84
C2	-	2,296	-	3	2,299	C2	2	1,003	-	4	1,009
C3	-	0	-	109	109	C3	-	-	-	-	0
C4	-	-	-	-	0	C4	0	2	-	0	2
Total	0	2,296	0	112	2,408	Total	78	1,013	0	4	1,095

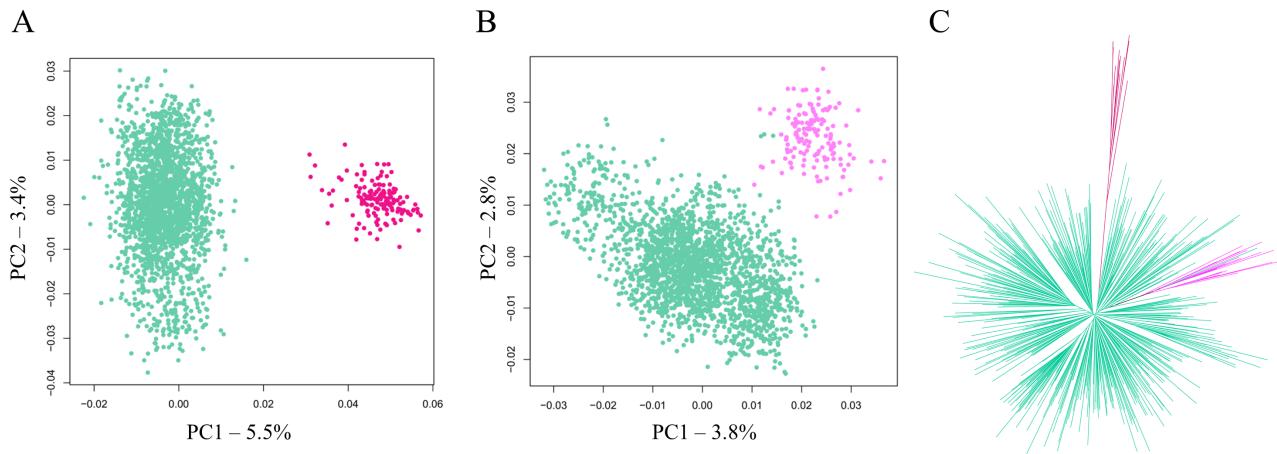
NOTE - The numbers of dimers are displayed at the intersection between left and right monomers. (A) For dimers without the XmnI site (noX), left and right monomers are not randomly associated : C2-C2 and C3-C4 over-representations are significant (Pearson's Chi-squared test ; P-value < 10^{-15}). (B) For dimers with the XmnI site (X), left and right monomers are not randomly associated : C1-C1 and C2-C2 over-representations are significant (P-value < 10^{-4}). The sign "-" means non-relevant associations, i.e. when the considered families are absent in the left or right monomers.

4.2 Matériel supplémentaire : deuxième article

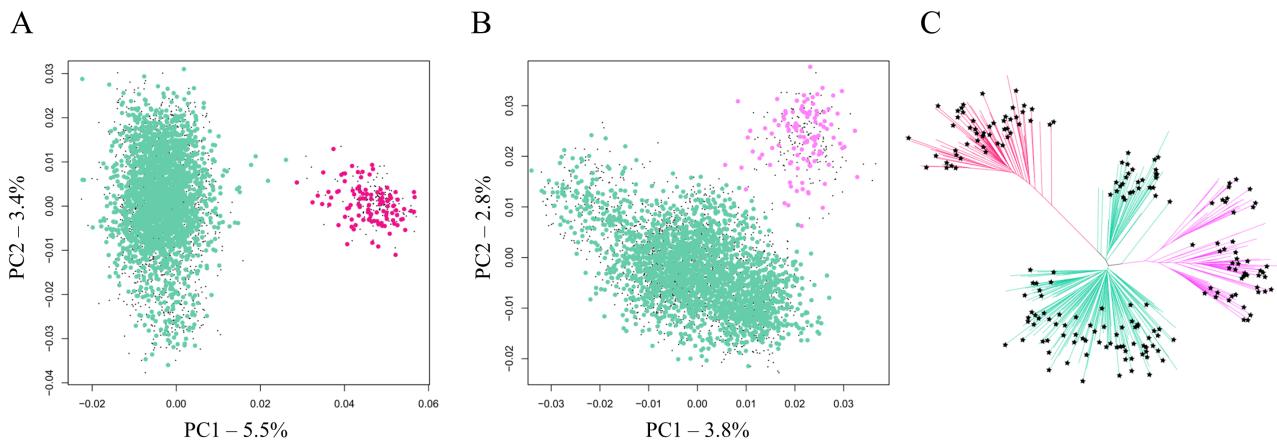
4.2.1 Figures supplémentaires



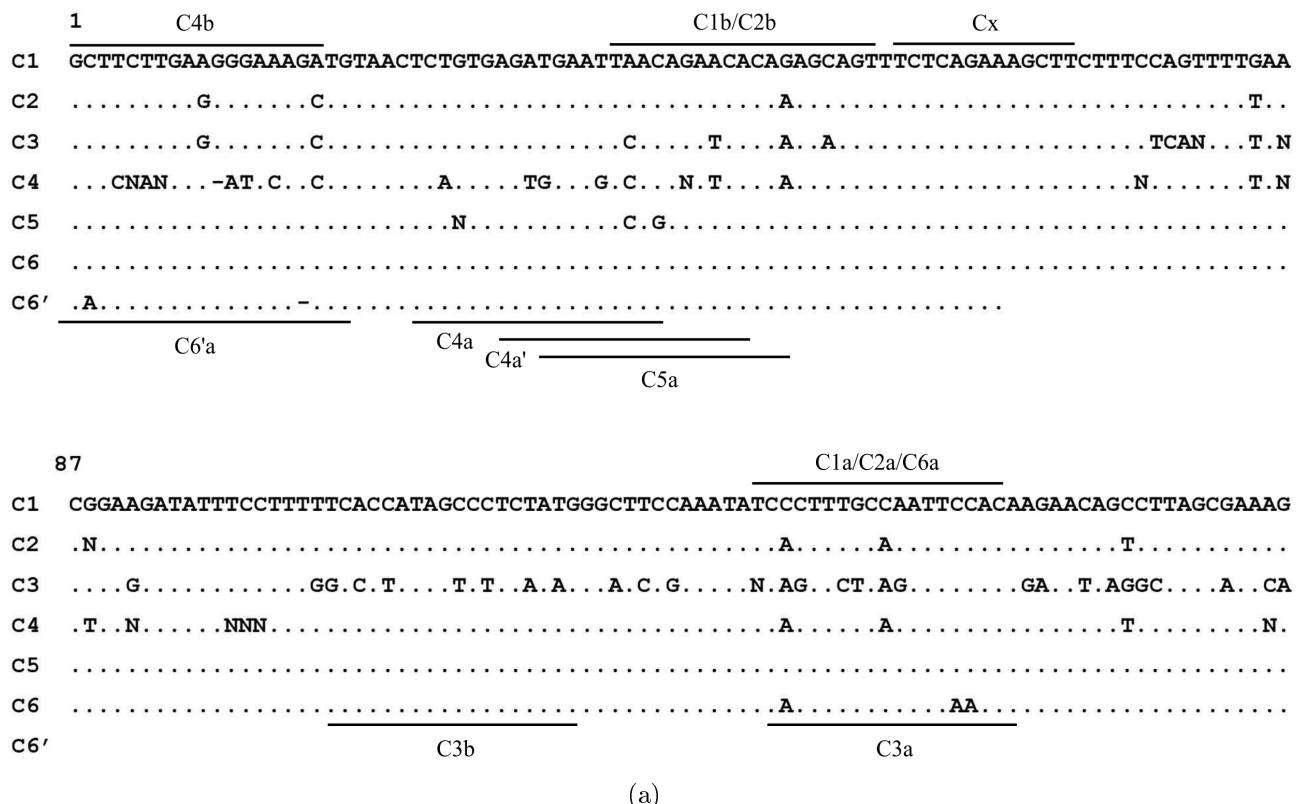
SUPPLEMENTARY FIGURE 2.1 – Phylogenetic analysis of alpha satellite DNA diversity in *C. pogonias* monomer dataset. (A) Phylogenetic tree (Neighbor-joining method, K2P model) for 500 randomly selected sequences, colored according to their assignment to the C1 (purple), C2 (pastel green), C5 (red) or C6 (orange) alpha satellite family based on hierarchical classification method. (B) Phylogenetic tree (Neighbor-joining method, K2P model) for 50 randomly selected sequences within *C. solatus* C1 and C2 sequences (labeled with arrows) and within *C. pogonias* C1, C2, C5 and C6 sequences. The color code matches the one described for (A). MUSCLE ([Edgar, 2004](#)) was used for alignment.



SUPPLEMENTARY FIGURE 2.2 – Characterization of alpha satellite diversity in *C. pogonias* dimer dataset. PCA projections of the normalized 5-mer frequency vectors are shown for (A) the left monomers and (B) the right monomers. Each point represents a sequence and has been colored according to its assignment to the C2 (pastel green), C3 (dark pink) or C4 (light pink) alpha satellite family based on hierarchical classification method. (C) Phylogenetic tree (Neighbor-joining method, K2P model) for 250 randomly selected left monomers and 250 randomly selected right monomers. The color code matches the one described for (A) and (B). Only dimers without the Xmnl site (noX dataset) were considered in this analysis. MUSCLE was used for alignment.



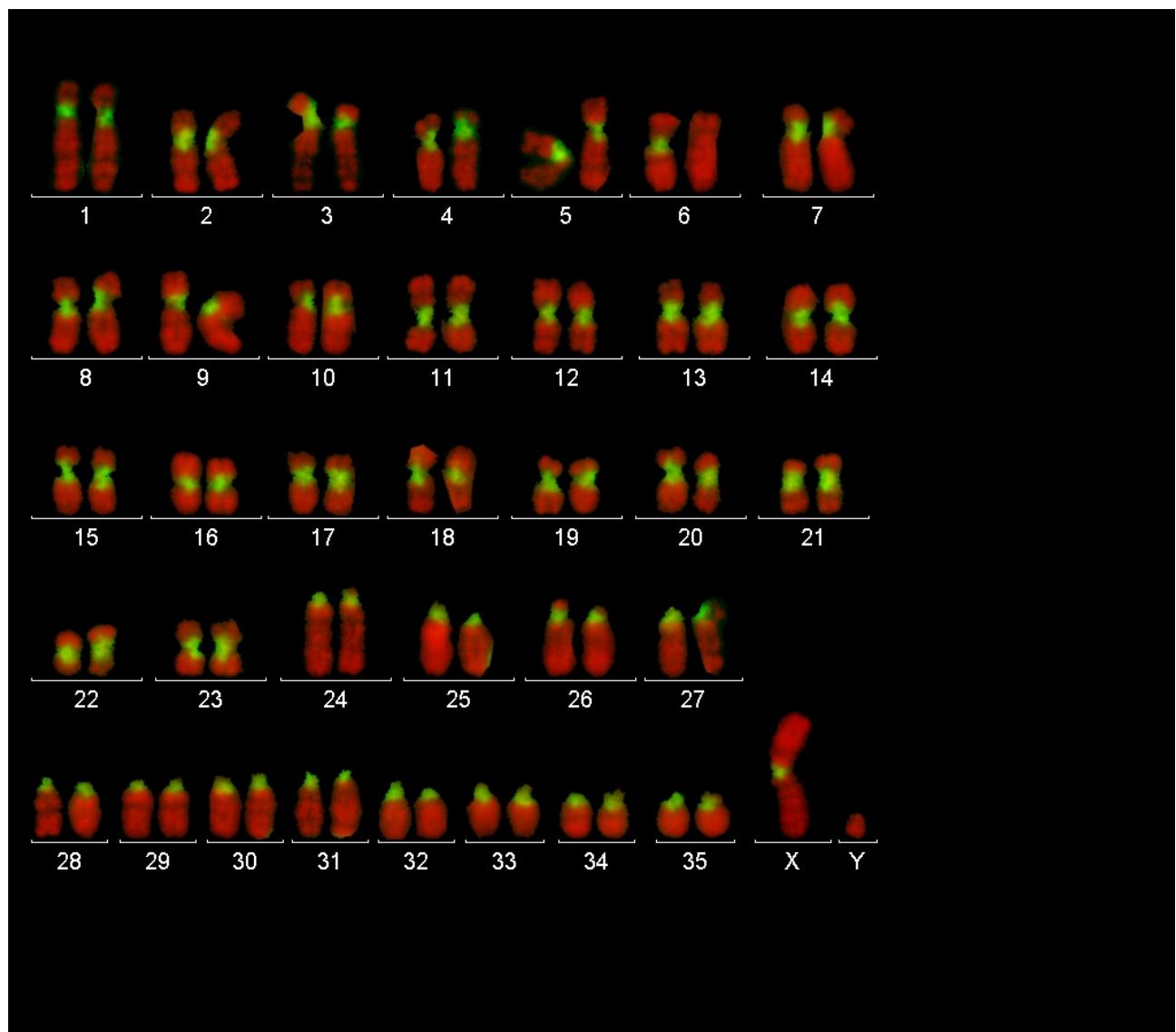
SUPPLEMENTARY FIGURE 2.3 – Interspecific comparison of alpha satellite diversity in dimer datasets. (A) Prediction of the C2 (pastel green) and C3 (dark pink) left monomers from *C. solatus* dimer dataset by using the PCA projection of *C. pogonias* left monomers (PCA projection on principal components 1 and 2 of the normalized 5-mer frequency vectors for all sequences). Each point represents a sequence. (B) Prediction of the C2 (pastel green) and C4 (light pink) right monomers from *C. solatus* dimer dataset by using the PCA projection of *C. pogonias* right monomers. (C) Phylogenetic tree (Neighbor-joining method, K2P model) for 50 randomly selected sequences within *C. solatus* C2 (pastel green) and C3 (dark pink) left monomers (labeled with arrows), within *C. solatus* C2 (pastel green) and C4 (light pink) right monomers (labeled with arrows), within *C. pogonias* C2 (pastel green) and C3 (dark pink) left monomers, and within *C. pogonias* C2 (pastel green) and C4 (light pink) right monomers. Only dimers without the Xmnl site (noX dataset) were considered in this analysis. MUSCLE was used for alignment.



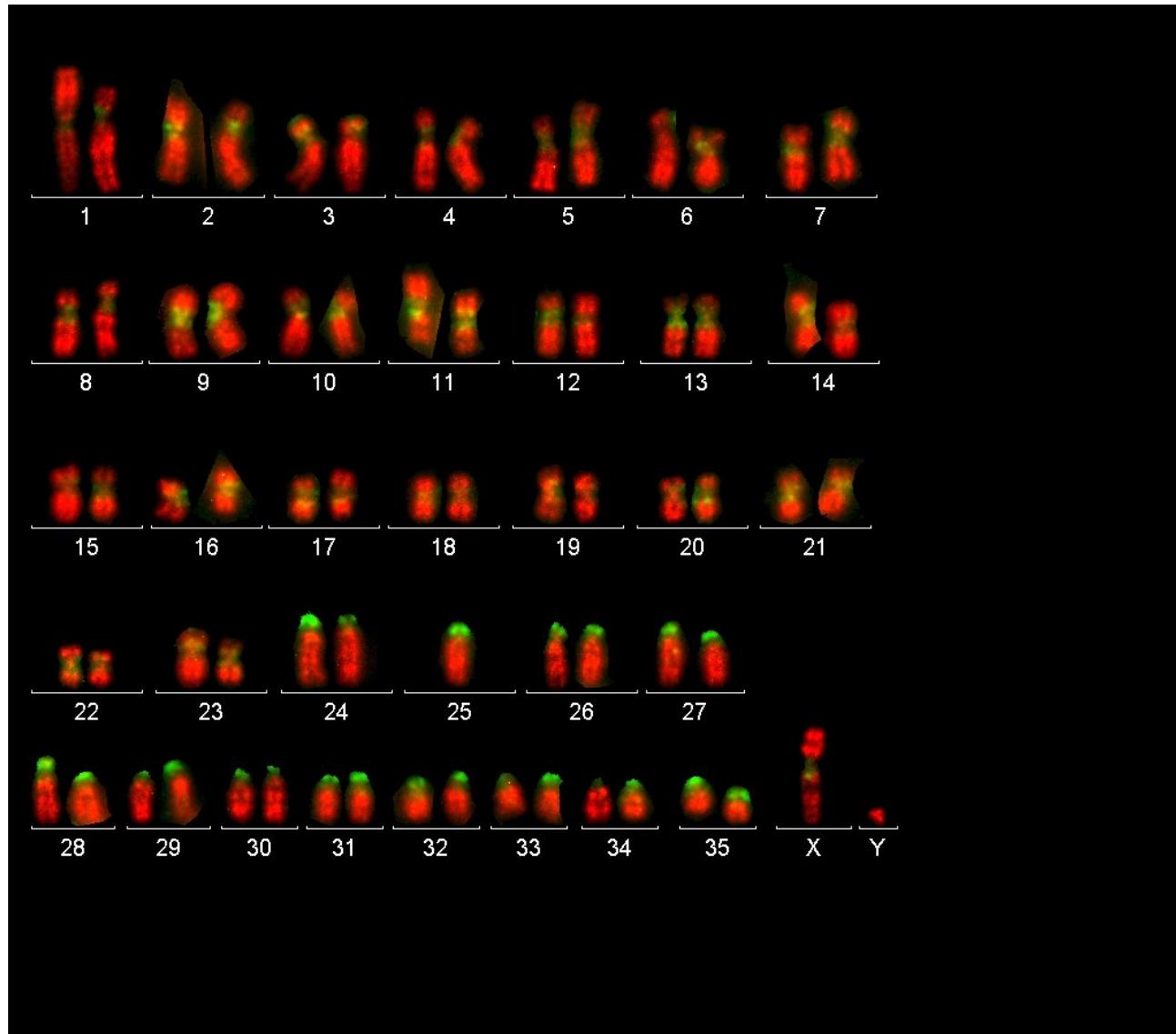
SUPPLEMENTARY FIGURE 2.4 – LNA-modified probes used to target the C1 to C6 families and the C6' sequences on *C. pogonias* chromosomes. (A) Probe binding sites on consensus. (B) Probe sequences and labels, associated to the percentage of targeted sequences within families. LNA are written in lower case and classic nucleotides are written in upper case. Percentage calculations have been made using all C1, C2, C5 and C6 sequences from the XmnI monomer dataset and all C3 and C4 sequences from the (noX) dimer dataset, when no mismatch or one mismatch was present. The sign "-" means 0. Probe C6'a (5'CaTTtTcCcTtCaAgAaTcC3', 3'Biotin) was calculated to target 75 % of HindIII C6' sequences.

Probe	Pattern (5'-3')	Label	Mismatch	C1 (%)	C2	C3	C4	C5	C6
C1a	TcCCtTtGcCaAtTcCAC	3'Cy3	0	63	1	-	-	64	-
			1	86	4	-	10	87	-
C1b	AcTgCtCtGtGtTcTGtTa	3'Digoxygenin	0	68	1	-	-	-	77
			1	91	29	-	-	1	95
C2a	TcACtTtGcAaAtTcCAC	5'AlexaFluor488	0	-	22	-	14	-	-
			1	5	57	-	52	1	-
C2b	AcTgCtTtGtGtTcTGtTa	5'Cy5	0	1	26	-	-	-	-
			1	69	58	1	-	-	78
C3a	CaGtTcTcAGATtCcAcA	3'Digoxygenin	0	-	-	28	-	-	-
			1	-	-	60	-	-	-
C3b	GcCcTaTaGtCtTcAaAg	3'Biotin	0	-	-	25	-	-	-
			1	-	-	59	-	-	-
C4a	GtGaCtTcCACtTcAcTgA	3'Biotin	0	-	-	-	30	-	-
			1	-	-	-	66	-	-
C4a'	TgAtGtGtGaCtTcCACt	3'Digoxygenin	0	-	-	-	14	-	-
			1	-	-	-	36	-	-
C4b	aGcTGtATTtTcGTgGaGc	3'Biotin	0	-	-	-	7	-	-
			1	-	-	-	37	-	-
C5a	TgAaTtCaGaGaAcAcAg	3'Biotin	0	-	-	-	-	77	-
			1	2	-	-	-	95	1
C6a	CaTTtTcCcTtCaAgAaTcC	3'Digoxygenin	0	-	-	-	-	-	72
			1	-	-	-	-	-	94
Cx	tctcagaaagctt	3'Biotin	0	77	40	65	41	76	74
			1	94	74	84	73	93	96

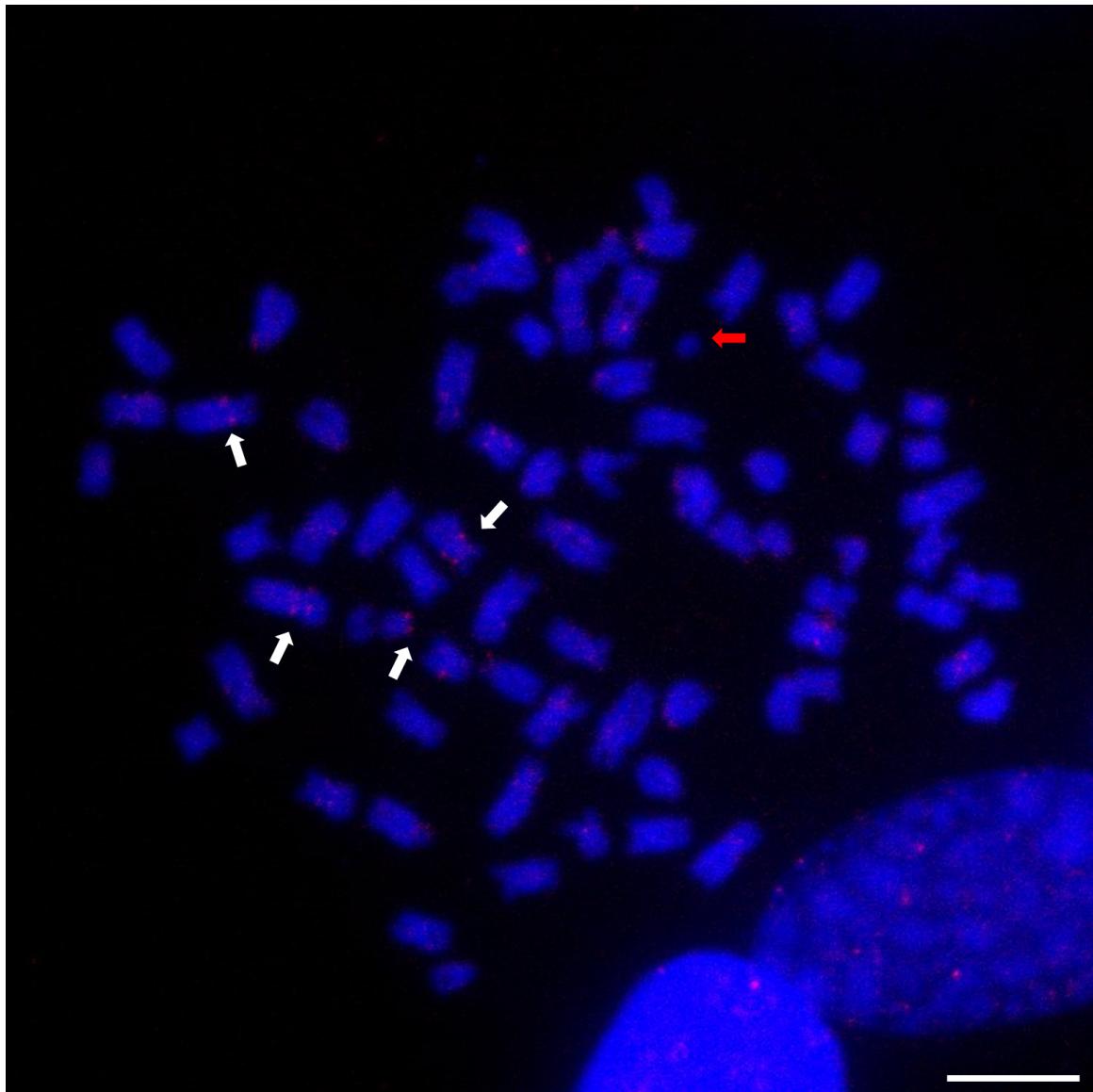
(b)



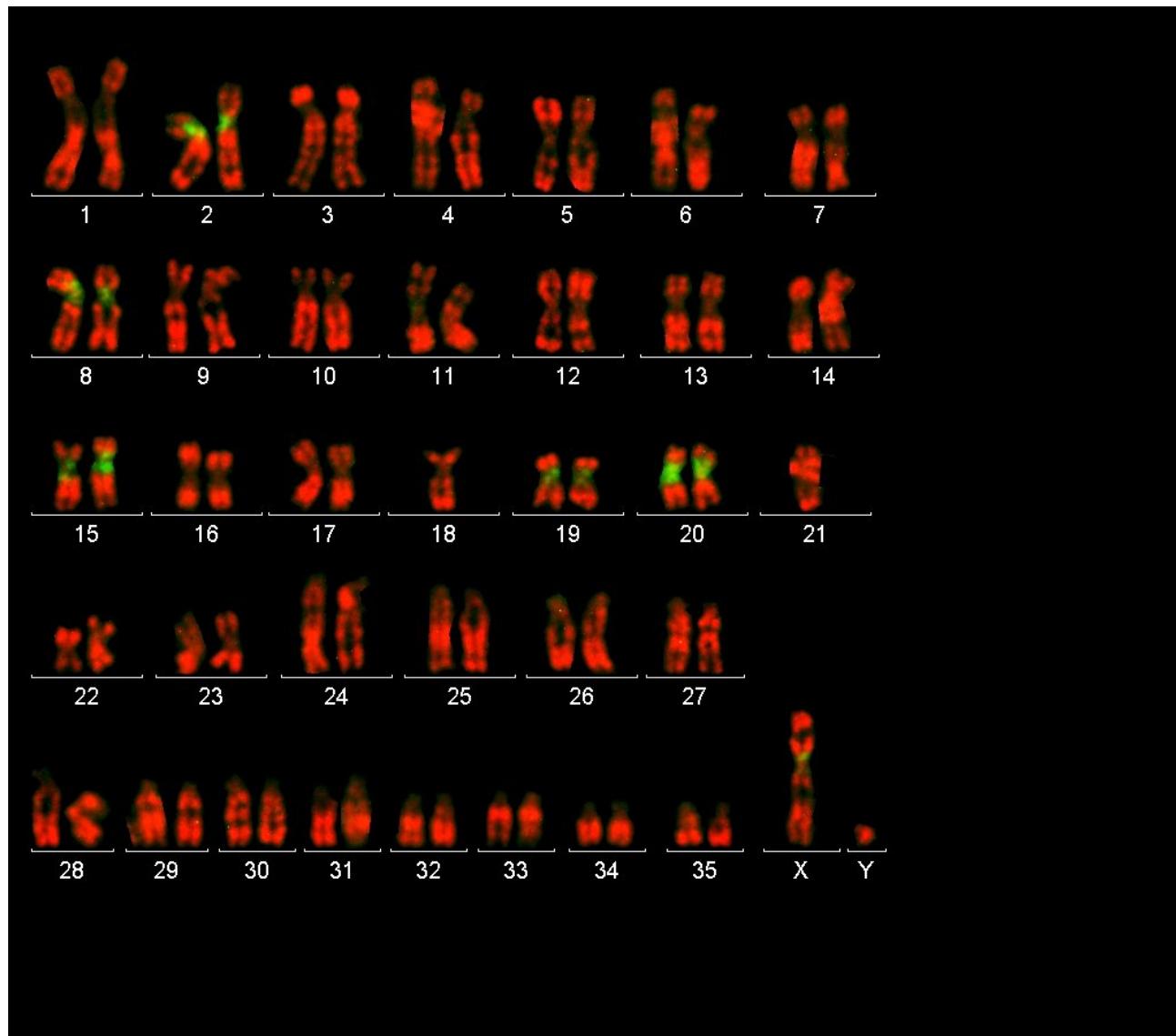
SUPPLEMENTARY FIGURE 2.5 – Hybridization pattern of probe Cx on *C. pogonias* chromosomes. Metaphase chromosomes are counterstained with propidium iodide (2 µg/mL) and shown here in red, while the green colors stand for probe Cx. RBG chromosome banding techniques were employed as in [Moulin et al. \(2008\)](#). Observation was performed with an epifluorescent microscope (Microphot-FXA, Nikon) and images were captured using a cooled CCD camera (ProgRes MFcool, Jenoplik). The metaphase was karyotyped using the Isis 5.3 software (Metasystems, Altusseheim, Germany) according to [Dutrillaux et al. \(1979\)](#); [Muleris et al. \(1984\)](#). The Y chromosome and one chromosome 6 of *C. pogonias* (homologous to HSA20-21 according to [Muleris et al. \(1984\)](#); [Moulin et al. \(2008\)](#)) are not labeled.



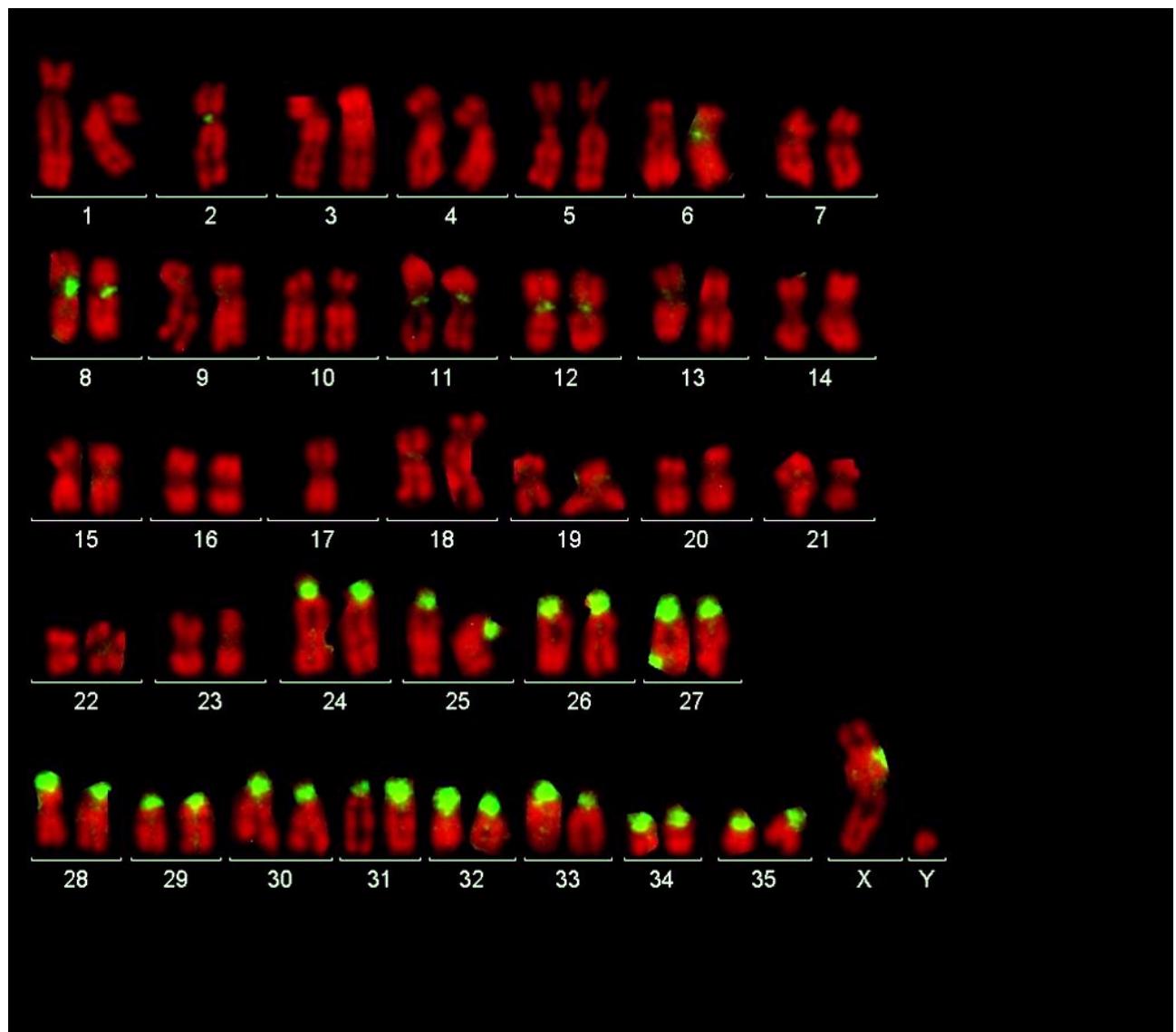
SUPPLEMENTARY FIGURE 2.6 – Distribution pattern of the C2 alpha satellite family on *C. pogonias* chromosomes. Hybridization of probe C2b (green) on metaphase chromosomes (red). See the legend of Supplementary figure 2.5 for more information about methods and references.



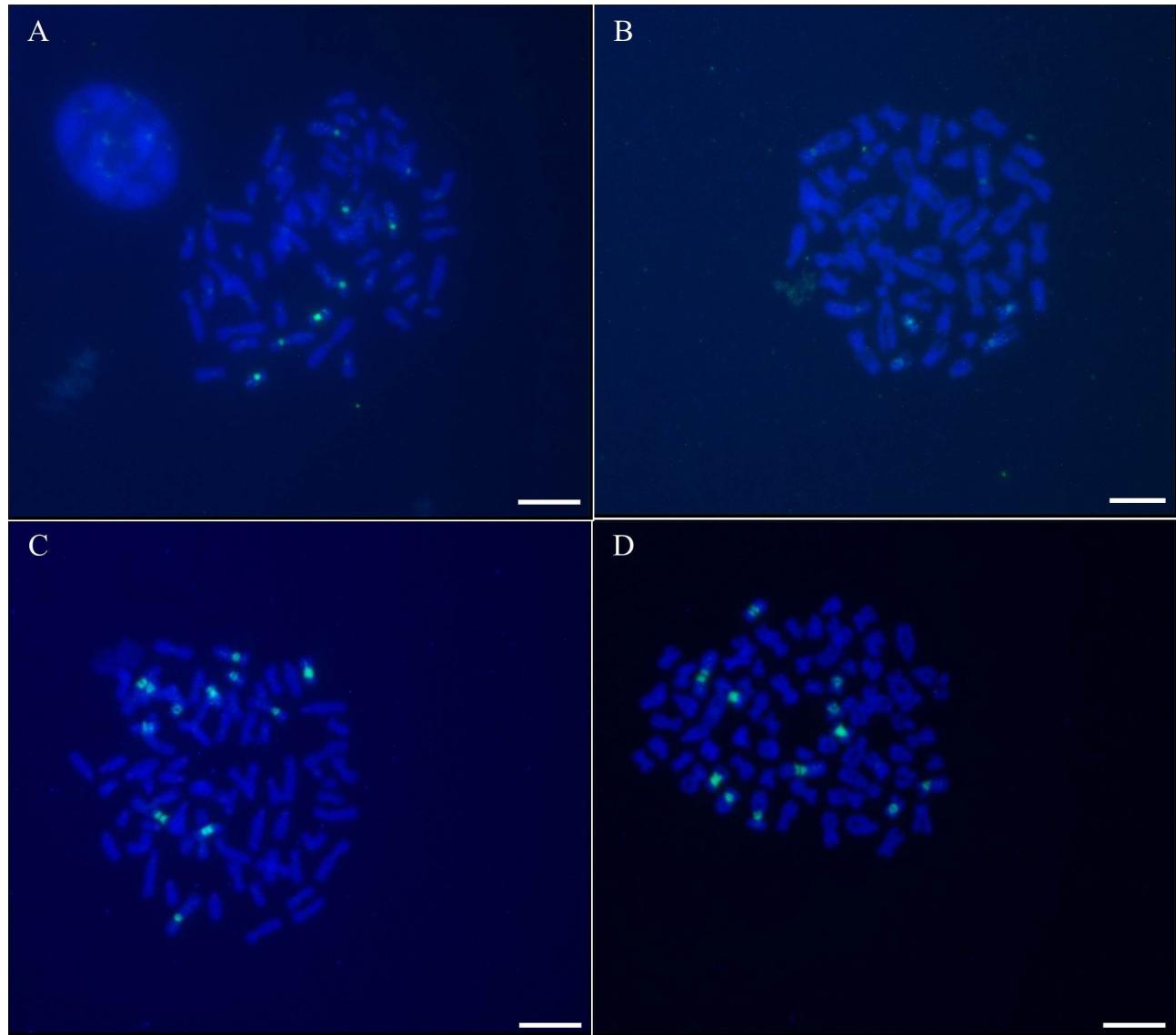
SUPPLEMENTARY FIGURE 2.7 – FISH analysis of the C4 alpha satellite family on *C. pogonias* chromosomes. Probe C4a' (red) is hybridized to *C. pogonias* metaphase chromosomes (blue). The white arrows point to some weak pericentromeric C4a' signals. The red arrow points to the Y chromosome which is not labeled by probe C4a', contrary to what was observed in *C. solatus*. Scale bar = 10 µm.



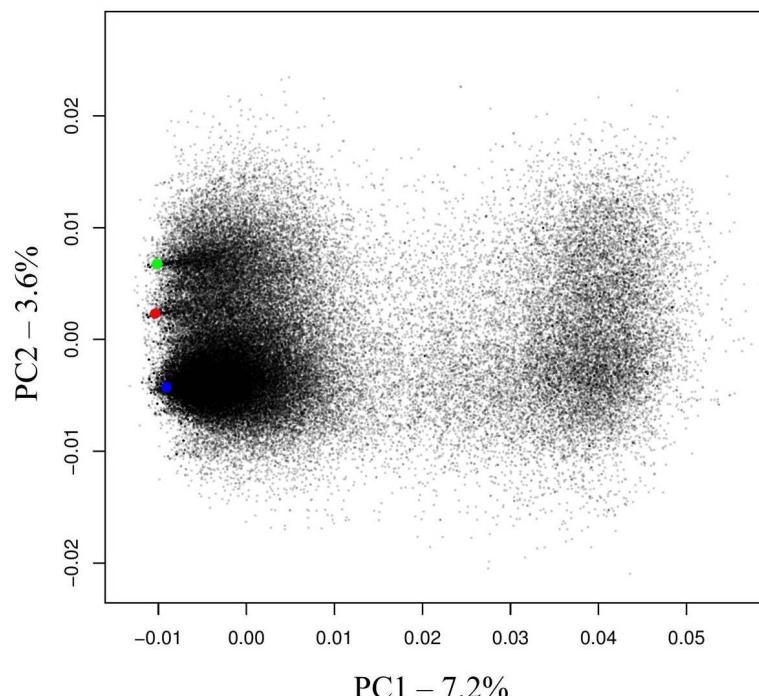
SUPPLEMENTARY FIGURE 2.8 – Distribution pattern of the C5 alpha satellite family on *C. pogonias* chromosomes. Hybridization of probe C5a (green) on metaphase chromosomes (red). See the legend of Supplementary figure 2.5 for more information about methods and references.



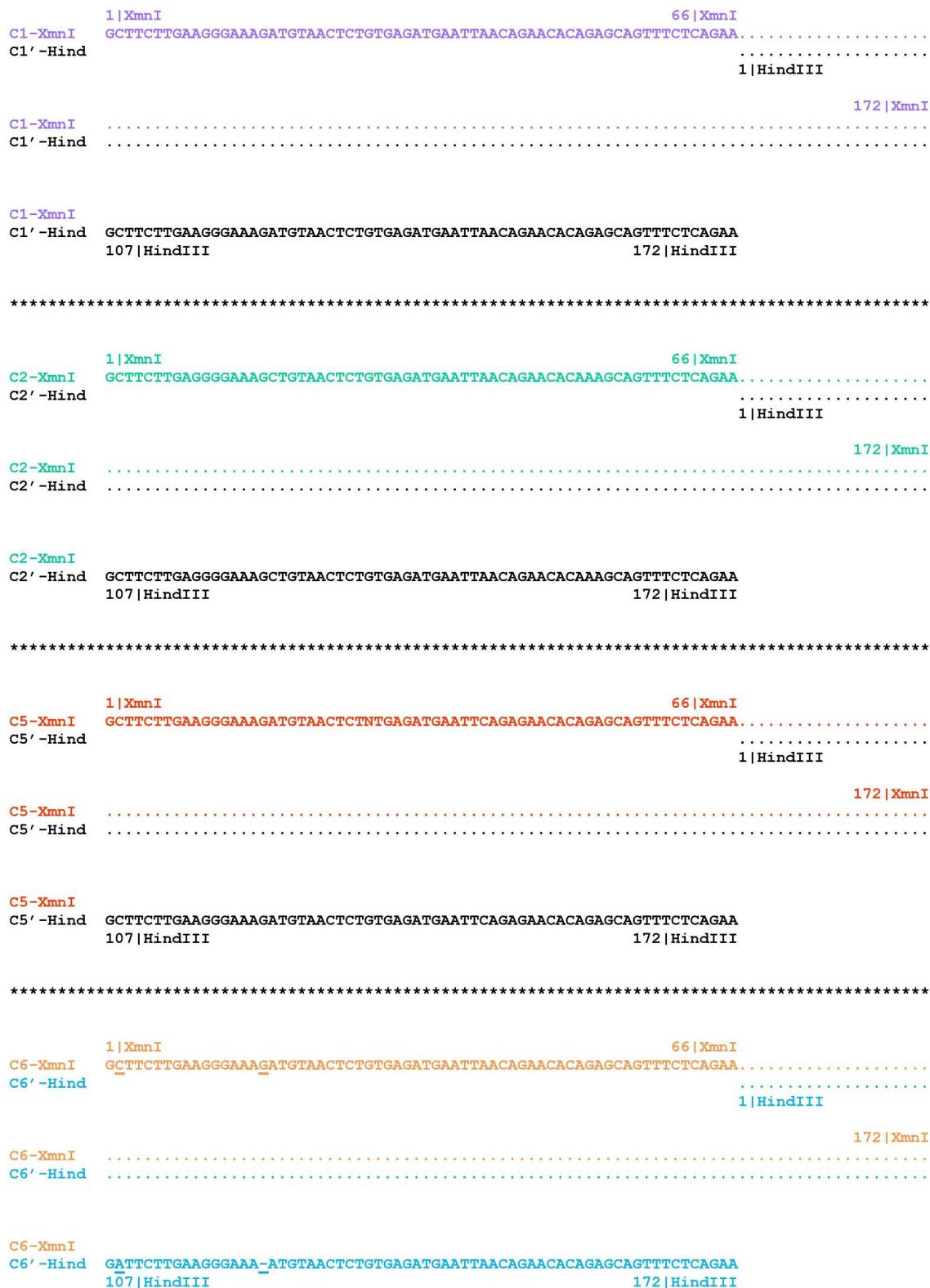
SUPPLEMENTARY FIGURE 2.9 – Distribution pattern of the C6 alpha satellite family on *C. pogonias* chromosomes. Hybridization of probe C6a (green) on metaphase chromosomes (red). The analysis of several metaphases allowed determining that a single chromosome 2 is labeled by probe C6a in the studied specimen. C6a signals on the long arm of left chromosome 27 and on chromosome X are due to chromosome superposition in the original metaphase. See the legend of Supplementary figure 2.5 for more information about methods and references.



SUPPLEMENTARY FIGURE 2.10 – **FISH analysis of the C5 alpha satellite family on *C. solatus* and *C. pogonias* chromosomes.** Metaphase chromosomes are colored in blue. (A) Hybridization of probe C5a (green) on *C. solatus* chromosomes, 63 °C washing. (B) Hybridization of probe C5a (green) on *C. solatus* chromosomes, 68 °C washing. (C) Hybridization of probe C5a (green) on *C. pogonias*, 63 °C washing. (D) Hybridization of probe C5a (green) on *C. pogonias*, 68 °C washing. Original pictures. A *C. pogonias* female sample (ID : 1979-013) was used for this control.



SUPPLEMENTARY FIGURE 2.11 – Characterization of comet-like clusters in *C. solatus* monomer dataset. PCA projection on principal components 1 and 2 of the normalized 5-mer frequency vectors for *C. pogonias* Xmni monomer dataset (sequences not shown). *Cercopithecus solatus* Xmni monomers are predicted by using this PCA projection (black points). *Cercopithecus solatus* repeated C1 sequences corresponding to the C1 consensus, or displaying a T39G variation, or displaying a A40C variation, are spotted in blue, red and green, respectively (Cacheux et al., 2016b).



SUPPLEMENTARY FIGURE 2.12 – Consensus sequence comparison between XmniI and HindIII monomer datasets. The consensus of the C1, C2, C5 and C6 families, identified in *C. pogonias* XmniI monomer dataset, are aligned with the consensus of their overlapping C1', C2', C5' or C6' group of HindIII sequences, respectively (see Figure 2.6). The overlapping 106 bp fragments are identical for each couple of consensus; each point is a nucleotide. The first 66 bp fragment of the C1, C2 and C5 consensus is identical to the last 66 bp fragment of the C1', C2' and C5' consensus, respectively. By contrast, those fragments are different between the C6 and C6' consensus (single nucleotide variations are underlined). MUSCLE was used for alignment.

4.2.2 Tables supplémentaires

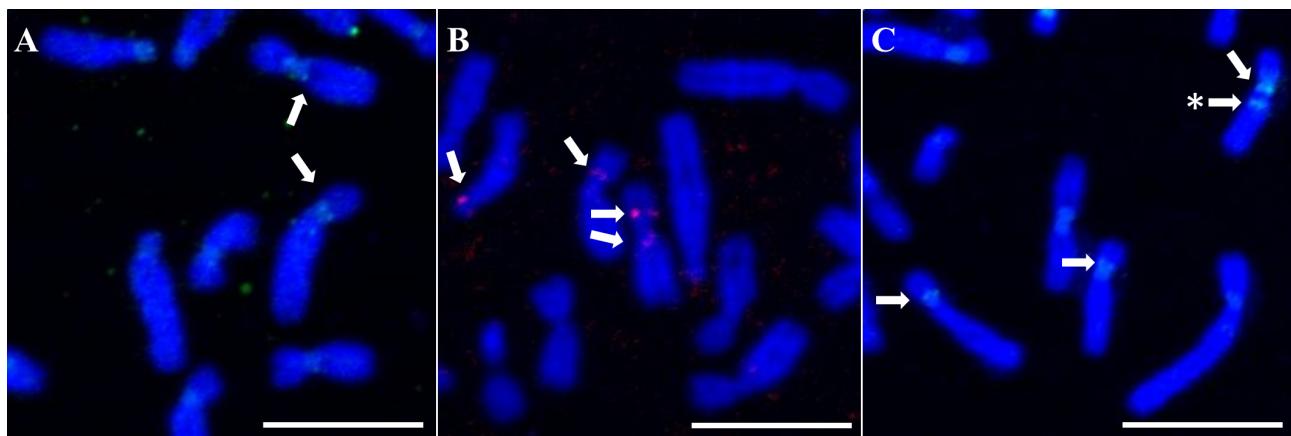
SUPPLEMENTARY TABLE 2.1 – Alpha satellite family associations in *C. pogonias* dimer dataset

A		noX dataset						B		X dataset						
		Right								Right						
Left		C1	C2	C3	C4	C5	C6	Total	Left	C1	C2	C3	C4	C5	C6	Total
C1	-	-	-	-	-	-	-	0	C1	4	2	-	-	-	-	6
C2	-	2,130	-	0	-	-	-	2,130	C2	0	57	-	-	-	-	57
C3	-	3	-	162	-	-	-	165	C3	-	-	-	-	-	-	0
C4	-	-	-	-	-	-	-	0	C4	0	1	-	-	-	-	1
C5	-	-	-	-	-	-	-	0	C5	-	-	-	-	-	-	0
C6	-	-	-	-	-	-	-	0	C6	0	1	-	-	-	-	1
Total	0	2,133	0	162	0	0	0	2,295	Total	4	61	0	0	0	0	65

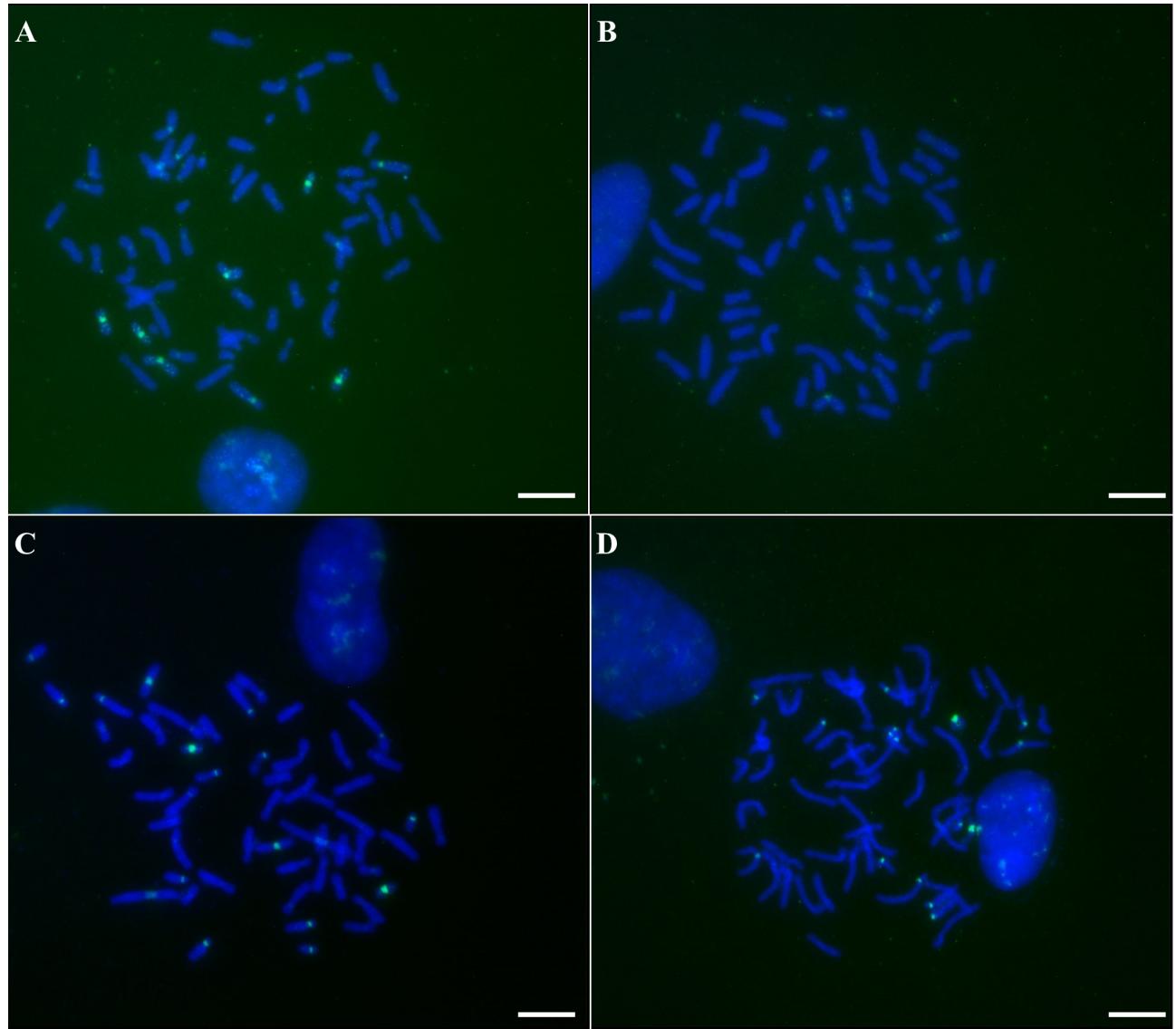
NOTE - The numbers of dimers are displayed at the intersection between left and right monomers. (A) For dimers without the Xmni site (noX), left and right monomers are not randomly associated : C2-C2 and C3-C4 over-representations are significant (Pearson's Chi-squared test ; P-value < 10^{-3}). (B) For dimers with the Xmni site (X), left and right monomers are not randomly associated : C1-C1 over-representation is significant (P-value < 10^{-3}). The sign "-" means non-relevant associations, i.e. when the considered families are absent in the left or right monomers.

4.3 Matériel supplémentaire : troisième article

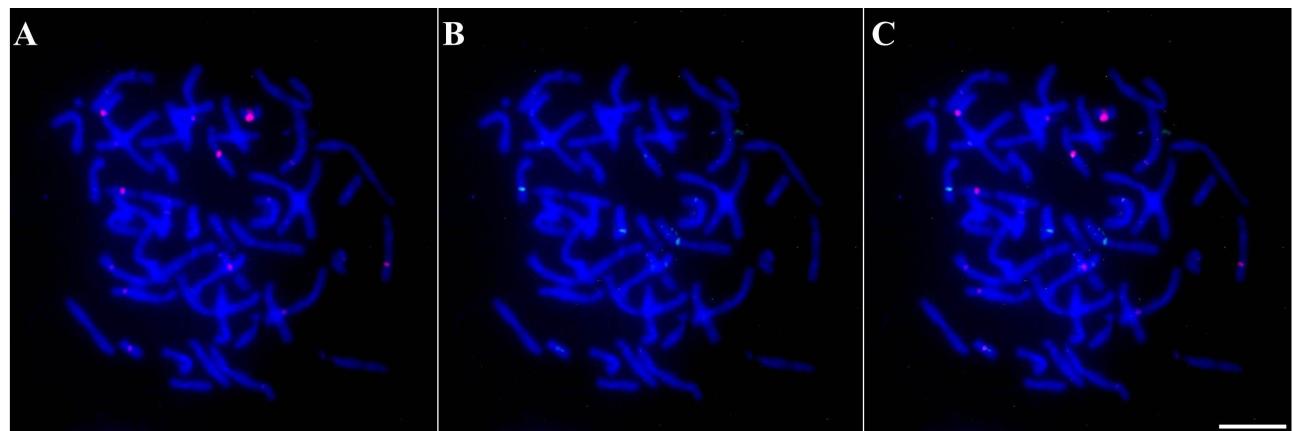
4.3.1 Figures supplémentaires



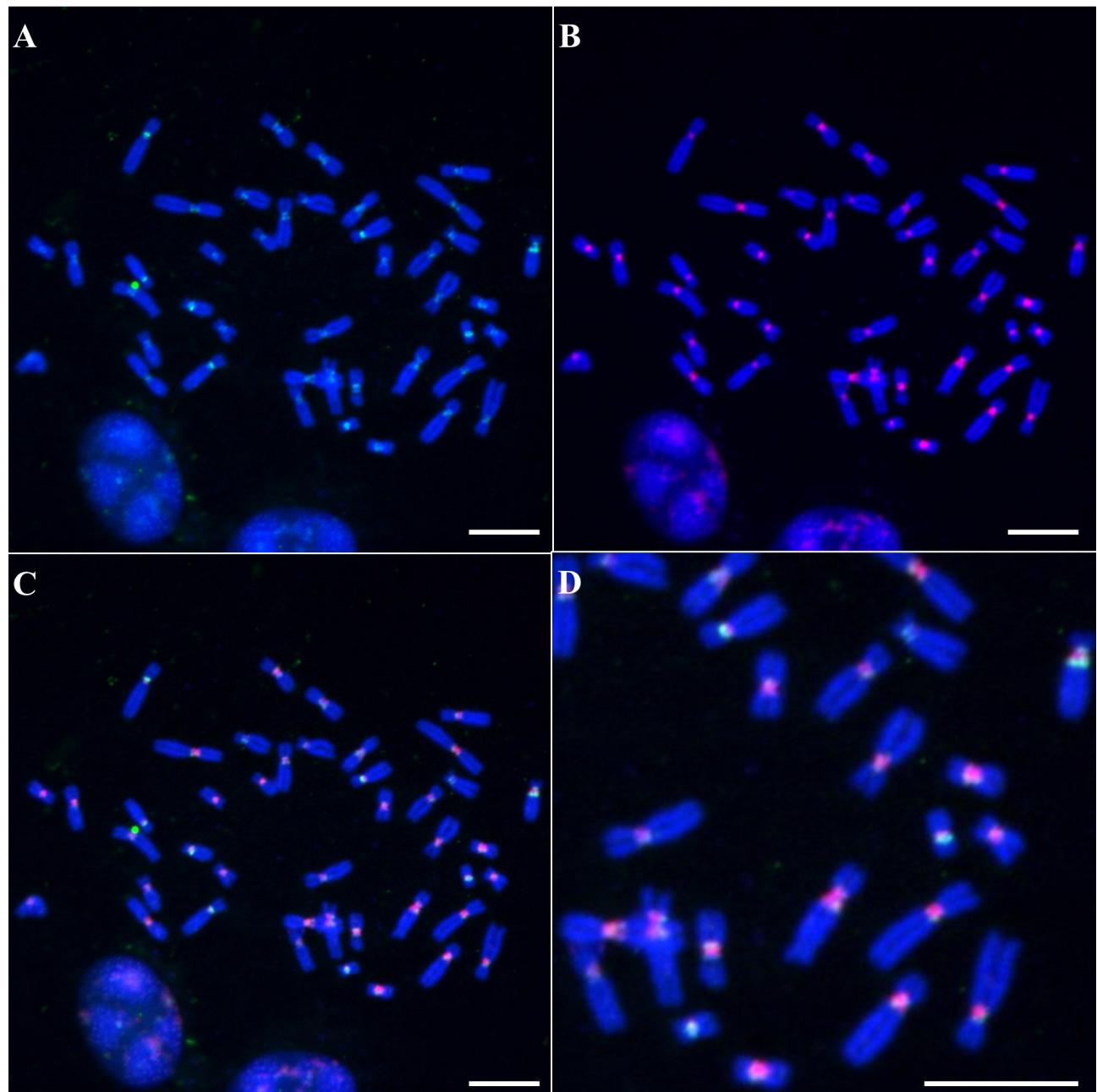
SUPPLEMENTARY FIGURE 3.1 – FISH analysis of the C3 and C4 alpha satellite families in species from the terrestrial clade. Metaphase chromosomes are colored in blue. (A) Hybridization of probe C4a (green) on *C. lhoesti* chromosomes. (B) Hybridization of probe C3b (red) on *C. lhoesti* chromosomes. (C) Hybridization of probe C4a (green) on *E. patas* chromosomes. Arrows point to sparingly labeled pericentromeres. The starry arrow points to a labeled interstitial locus. Scale bar = 10 μ m.



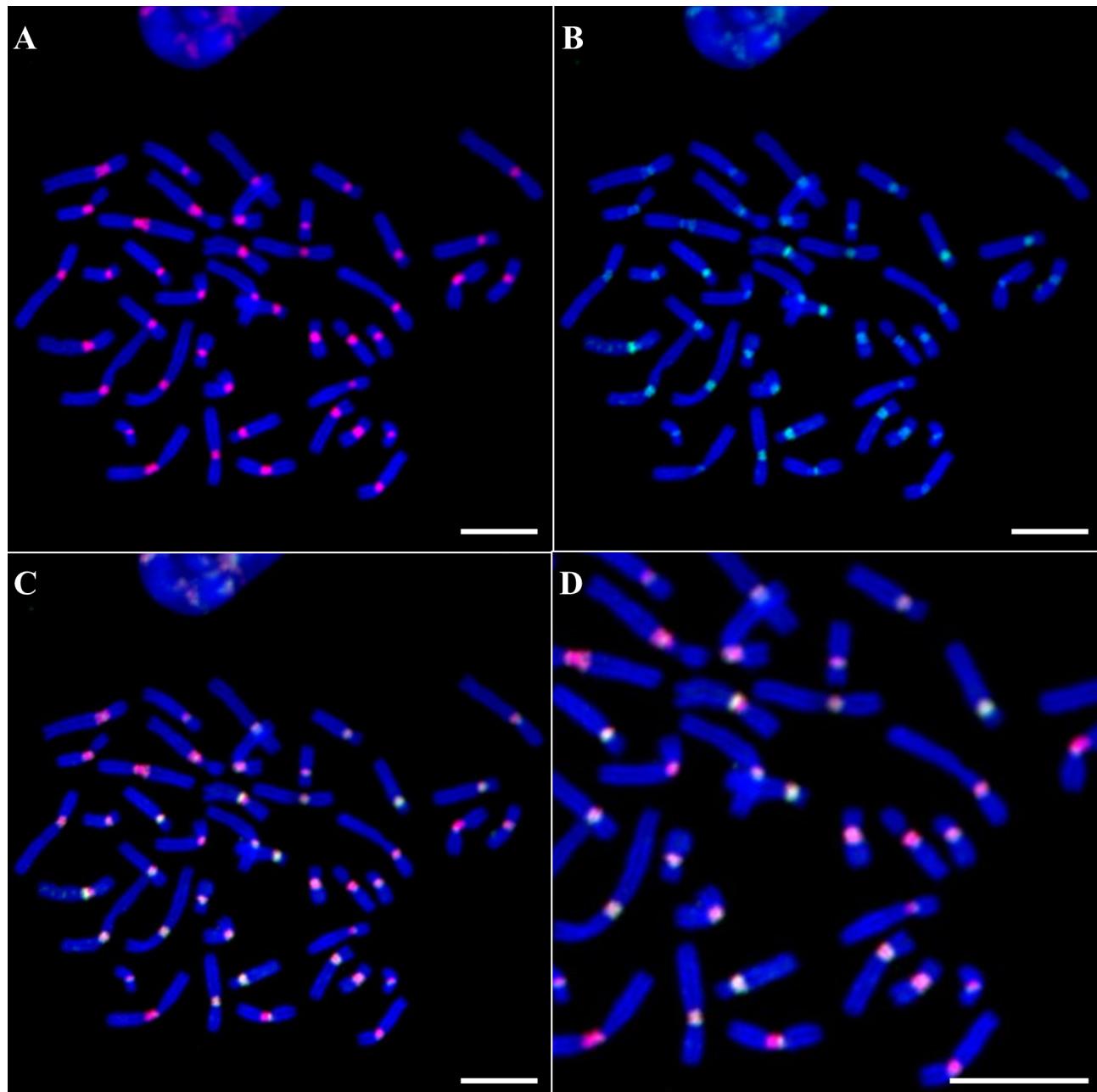
SUPPLEMENTARY FIGURE 3.2 – FISH analysis of the C5 alpha satellite family in species from the terrestrial clade. Metaphase chromosomes are colored in blue. (A) Hybridization of probe C5a (green) on *C. lhoesti* chromosomes, 63 °C washing. (B) Hybridization of probe C5a (green) on *C. lhoesti* chromosomes, 68 °C washing. (C) Hybridization of probe C5a (green) on *E. patas* chromosomes, 63 °C washing. (D) Hybridization of probe C5a (green) on *E. patas* chromosomes, 68 °C washing. Original pictures. Scale bar = 10 µm.



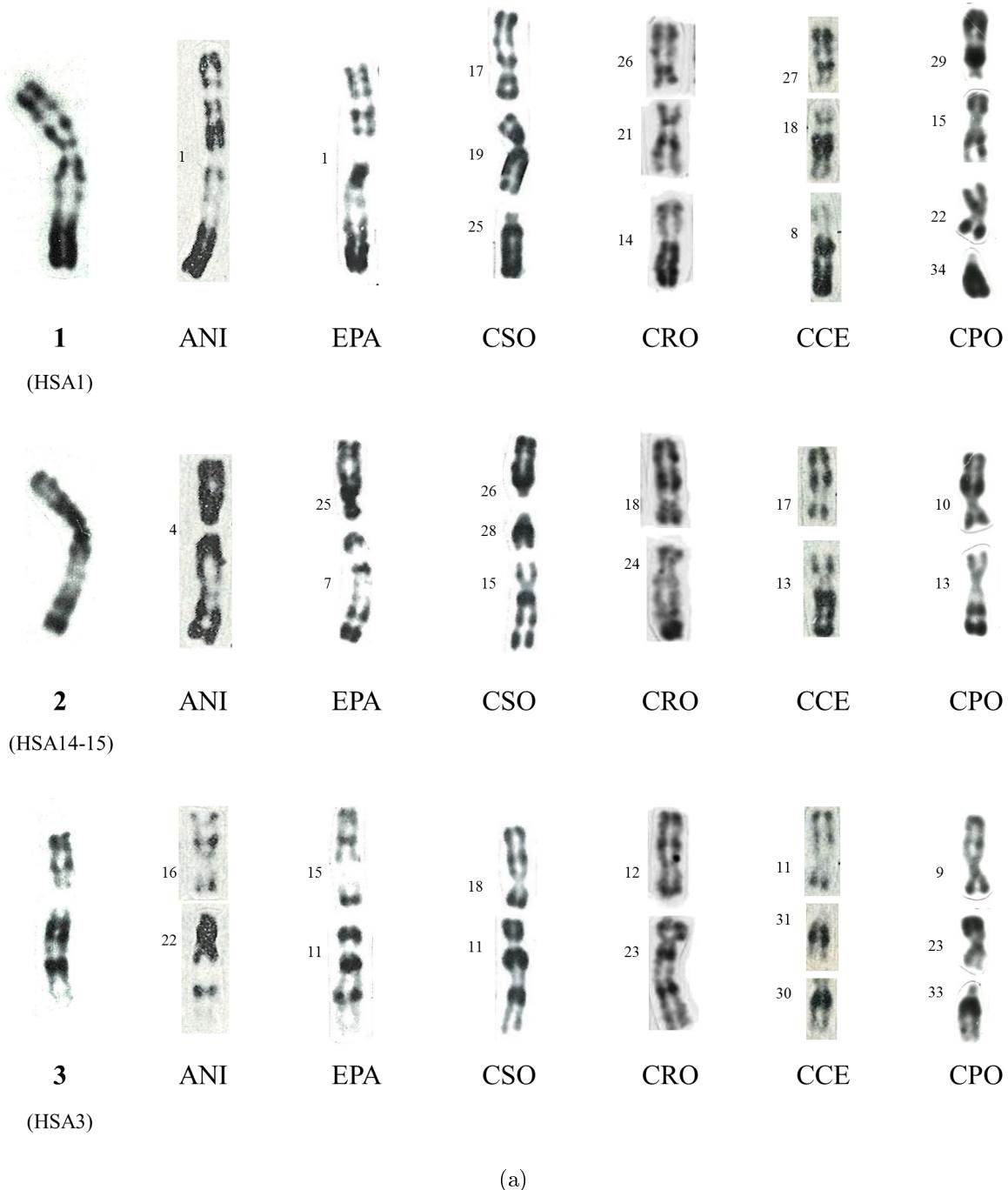
SUPPLEMENTARY FIGURE 3.3 – FISH analysis of the C1 alpha satellite family on *A. nigro-viridis* chromosomes. Probes C1a, C1b and C2b are hybridized simultaneously to metaphase chromosomes, which are colored in blue. (A) Hybridization of probe C1a (red). (B) Hybridization of probe C1b (green). (C) Superposition of the two previous images. Scale bar = 10 µm.



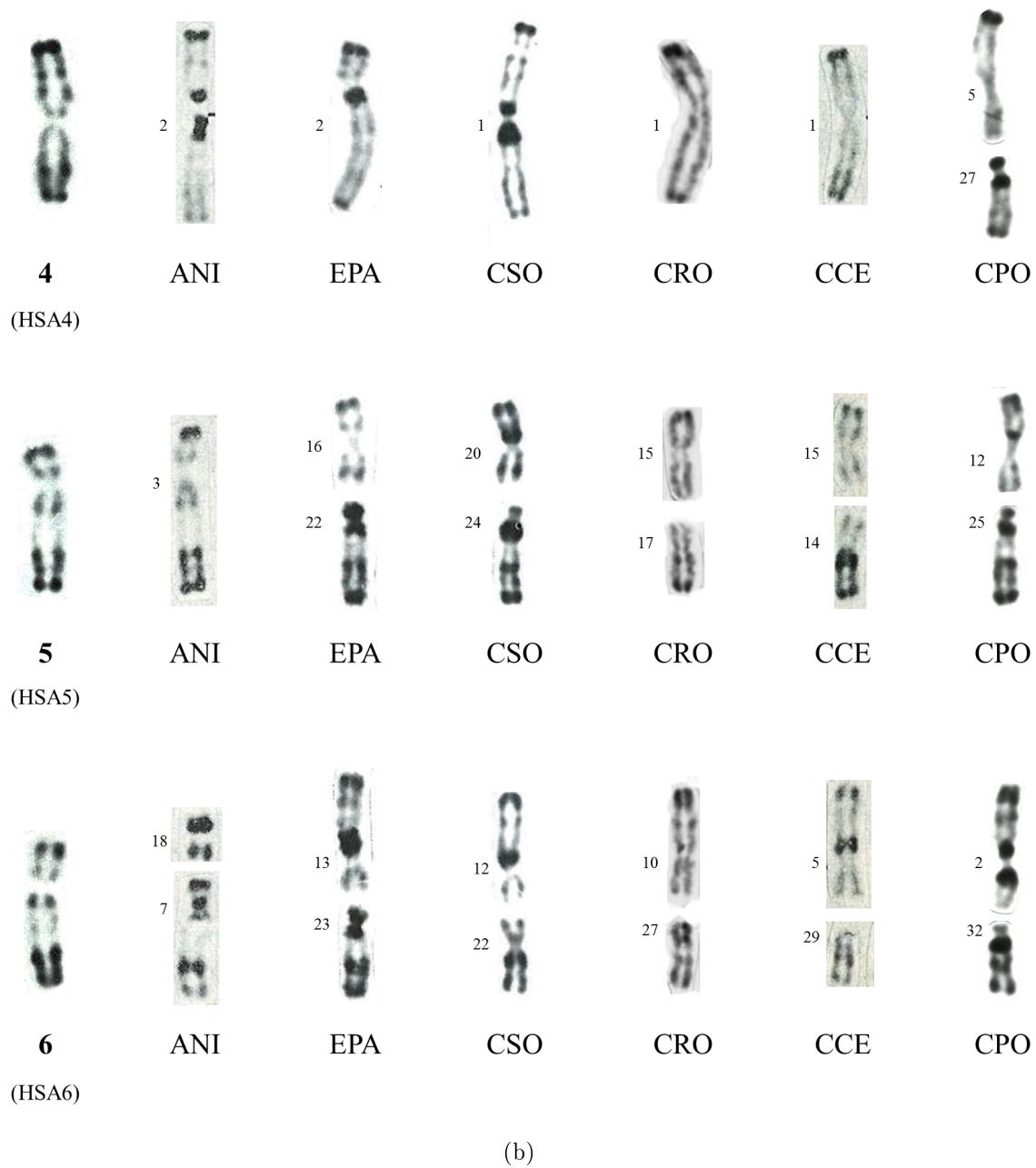
SUPPLEMENTARY FIGURE 3.4 – **FISH analysis of the C2 alpha satellite family on *M. sylvanus* chromosomes.** Probes C2a, C2b and C1b are hybridized simultaneously to metaphase chromosomes, which are colored in blue. (A) Hybridization of probe C2a (green). (B) Hybridization of probe C2b (red). (C) Superposition of the two previous images. (D) Focus on image (C). Scale bar = 10 µm.

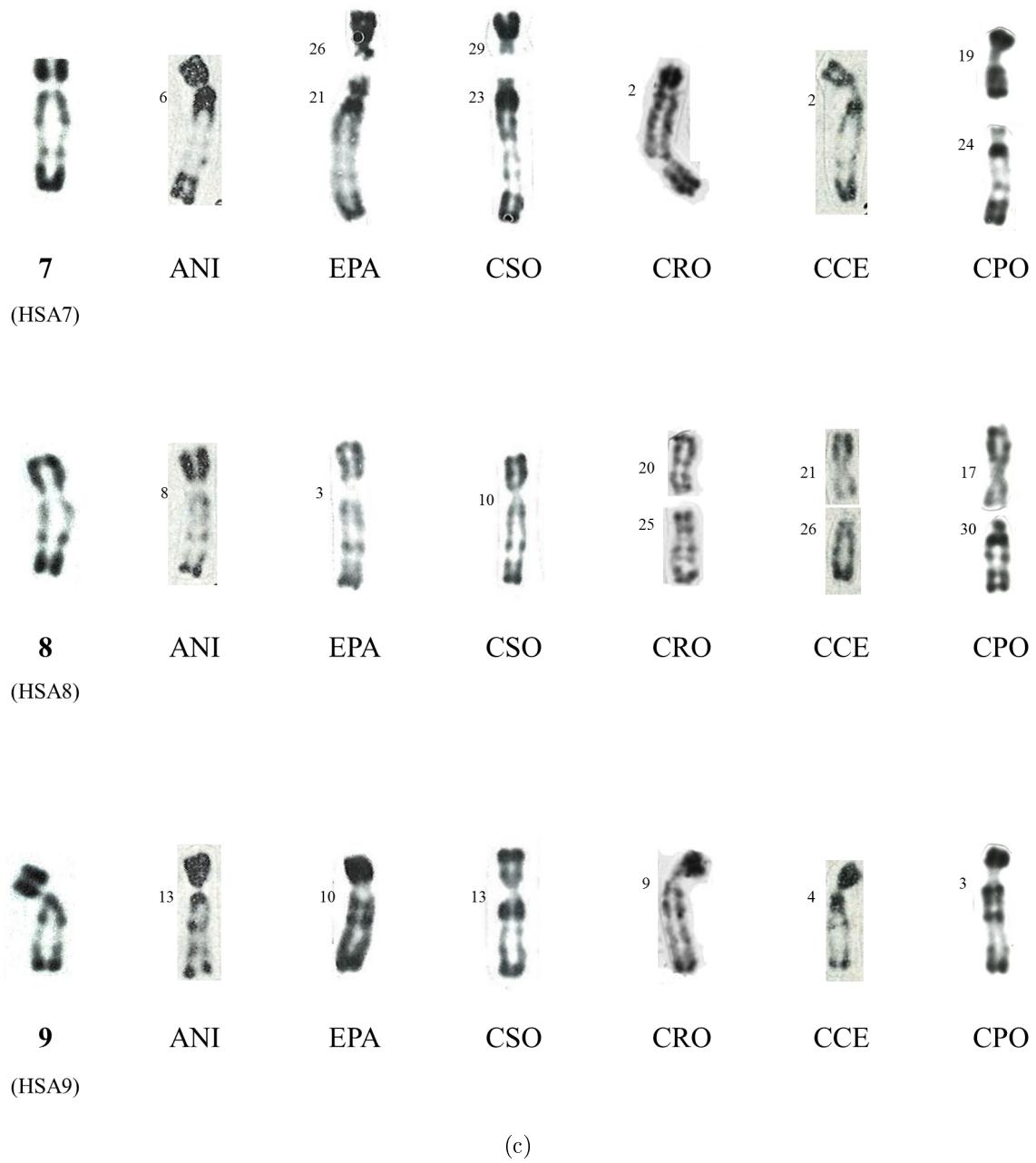


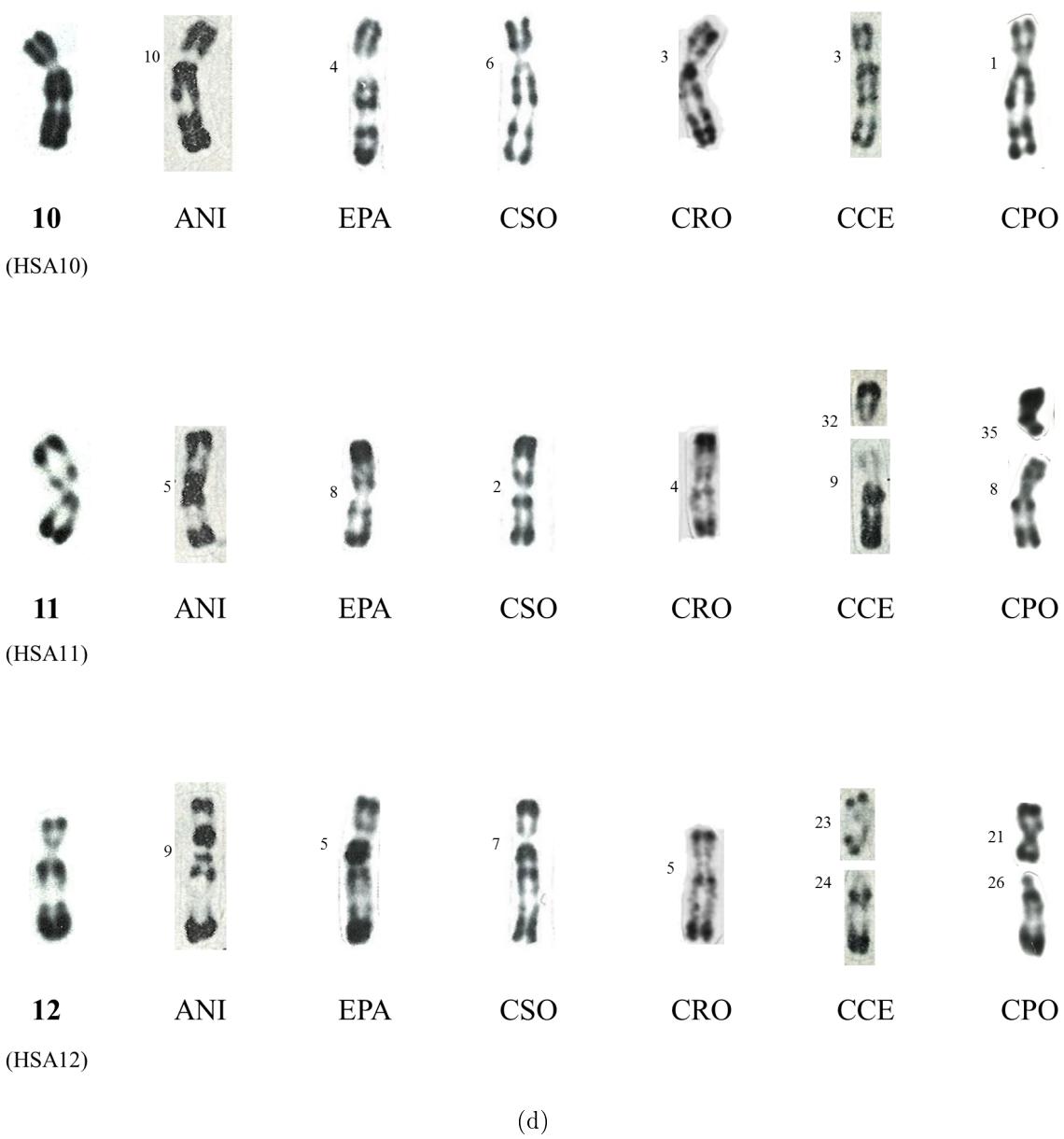
SUPPLEMENTARY FIGURE 3.5 – FISH analysis of the C3 and C4 alpha satellite families on *M. sylvanus* chromosomes. Probes C3b and C4a are hybridized simultaneously to metaphase chromosomes, which are colored in blue. (A) Hybridization of probe C3b (red). (B) Hybridization of probe C4a (green). (C) Superposition of the two previous images. (D) Focus on image (C). Scale bar = 10 µm.

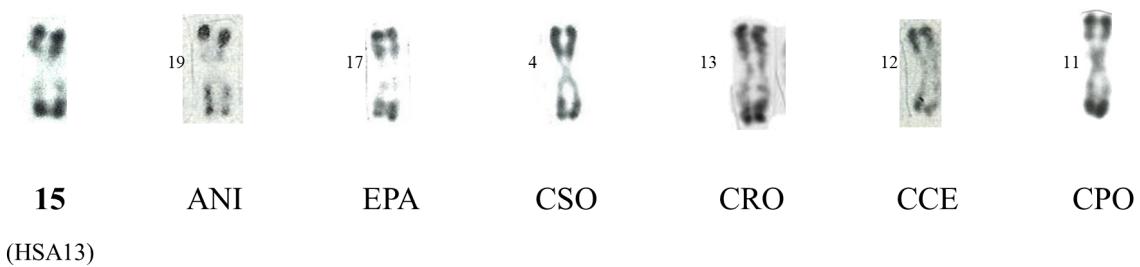
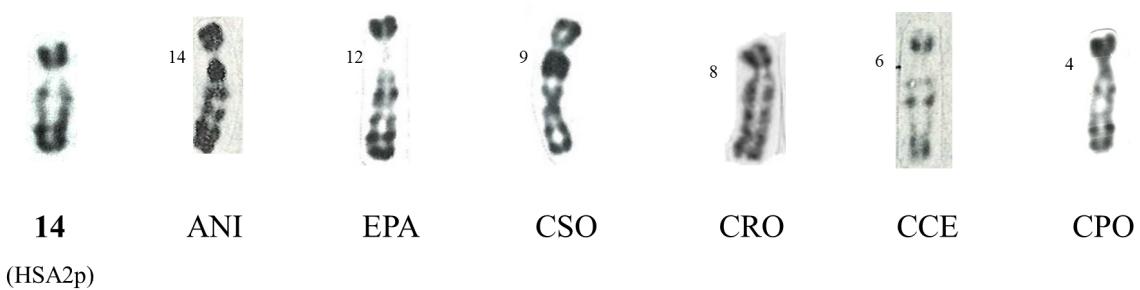
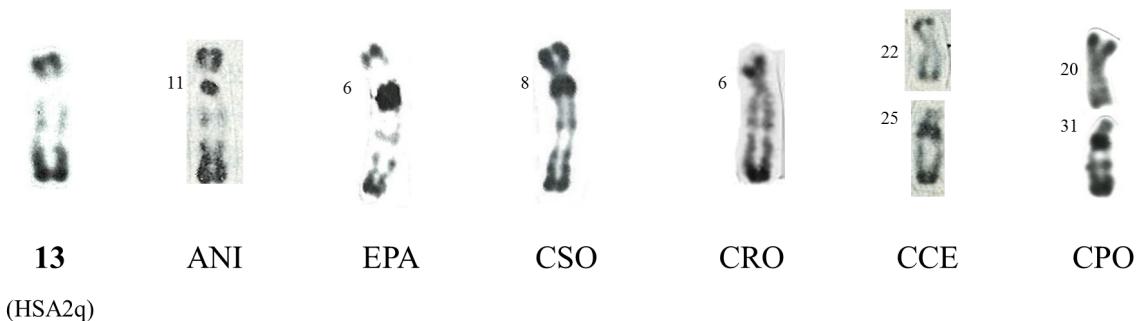


SUPPLEMENTARY FIGURE 3.6 – Alignment of homologous chromosomes between Cercopithecini species. Presumed ancestral chromosomes of Cercopithecidae, i.e. Old World monkeys (left), are aligned with their homologs in *A. nigroviridis* (ANI), *E. patas* (EPA), *C. solatus* (CSO), *C. roloway* (CRO), *C. cephus* (CCE) and *C. pogonias* (CPO) (Dutrillaux et al., 1981; Moulin et al., 2008). Chromosome numbers in reference karyotypes are displayed. Homologies with human chromosomes (HSA) are mentioned. Ancestral chromosomes 20 and 21 are associated on this figure because they are fused in all Cercopithecini species.

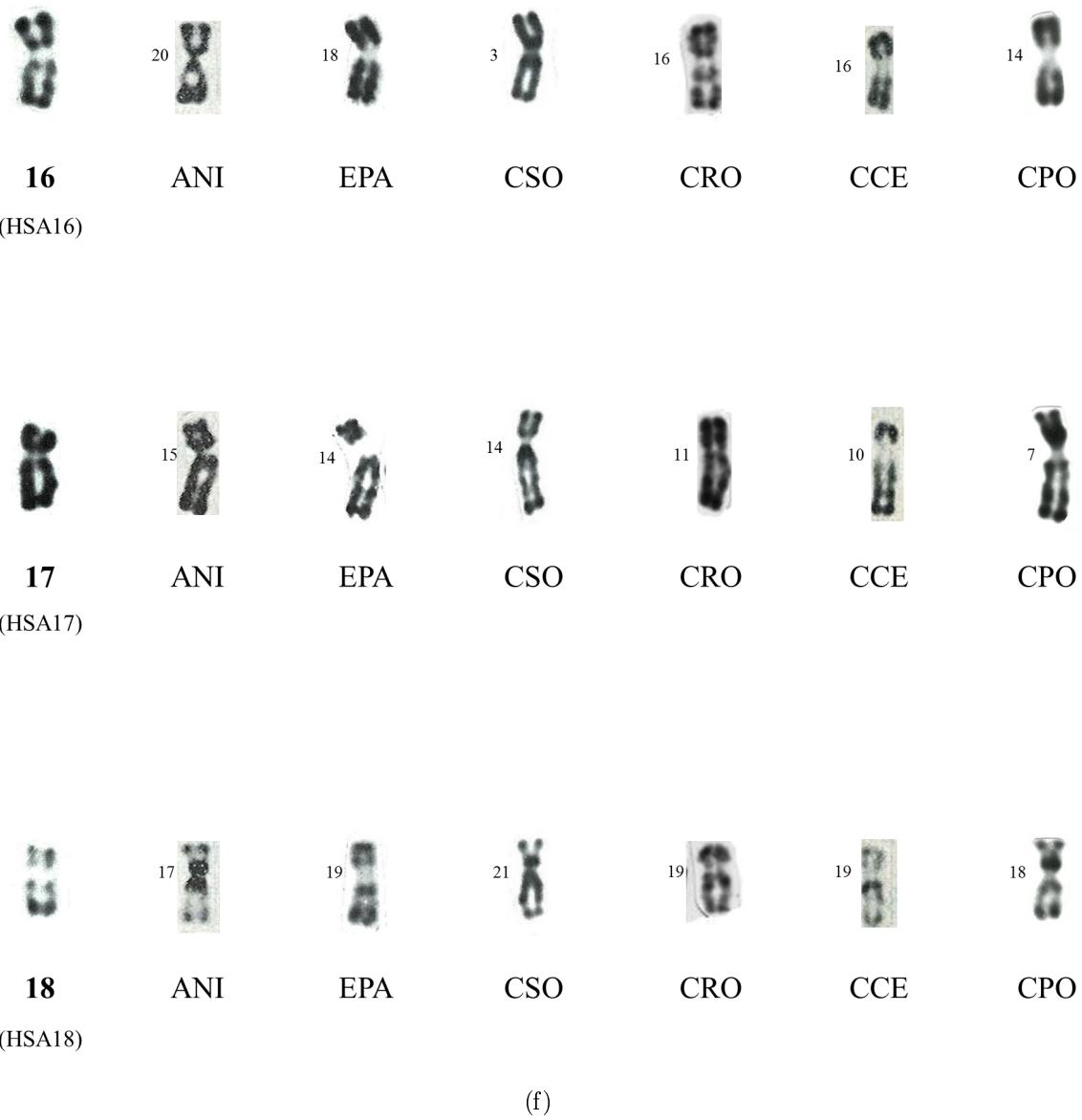


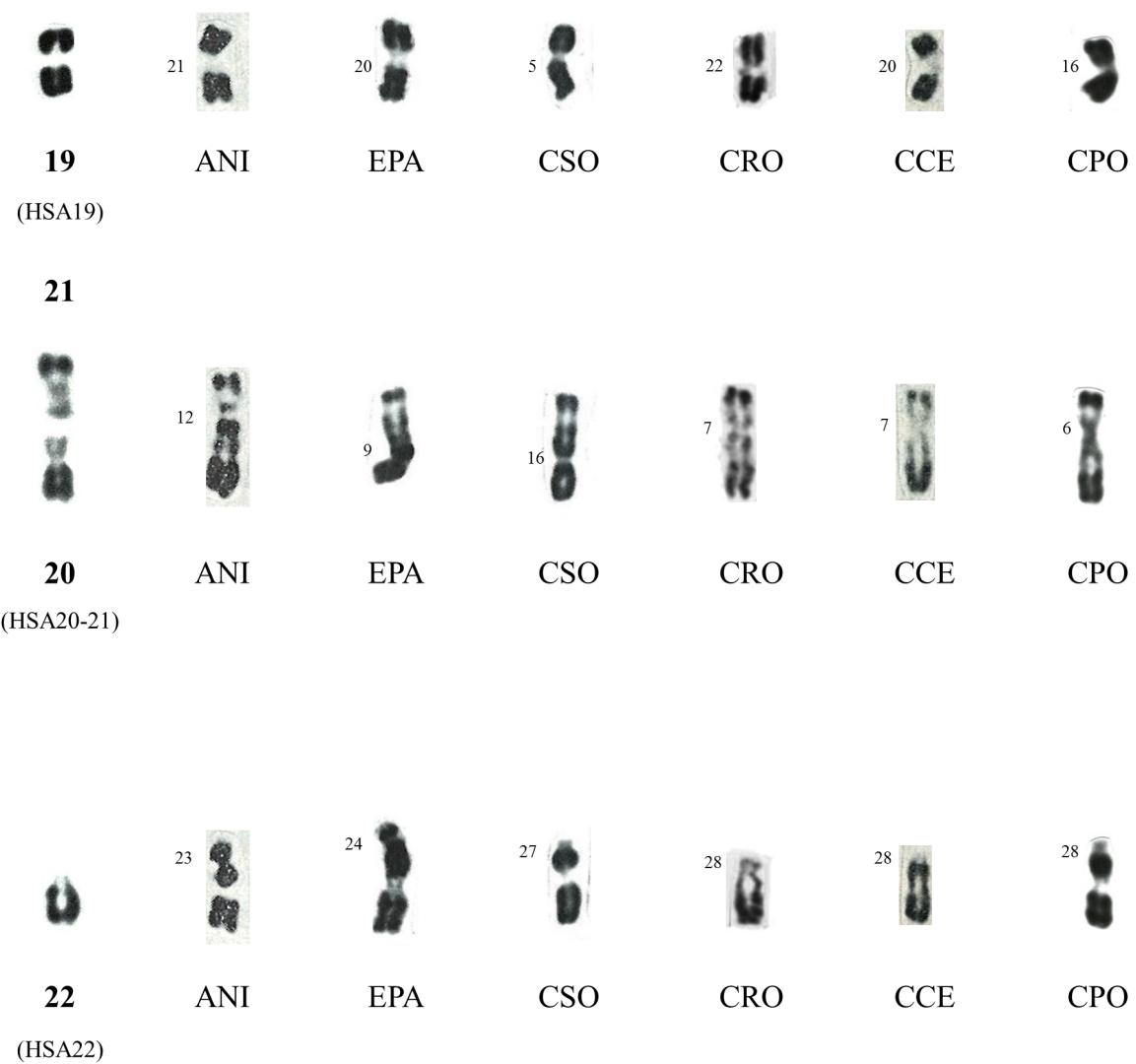


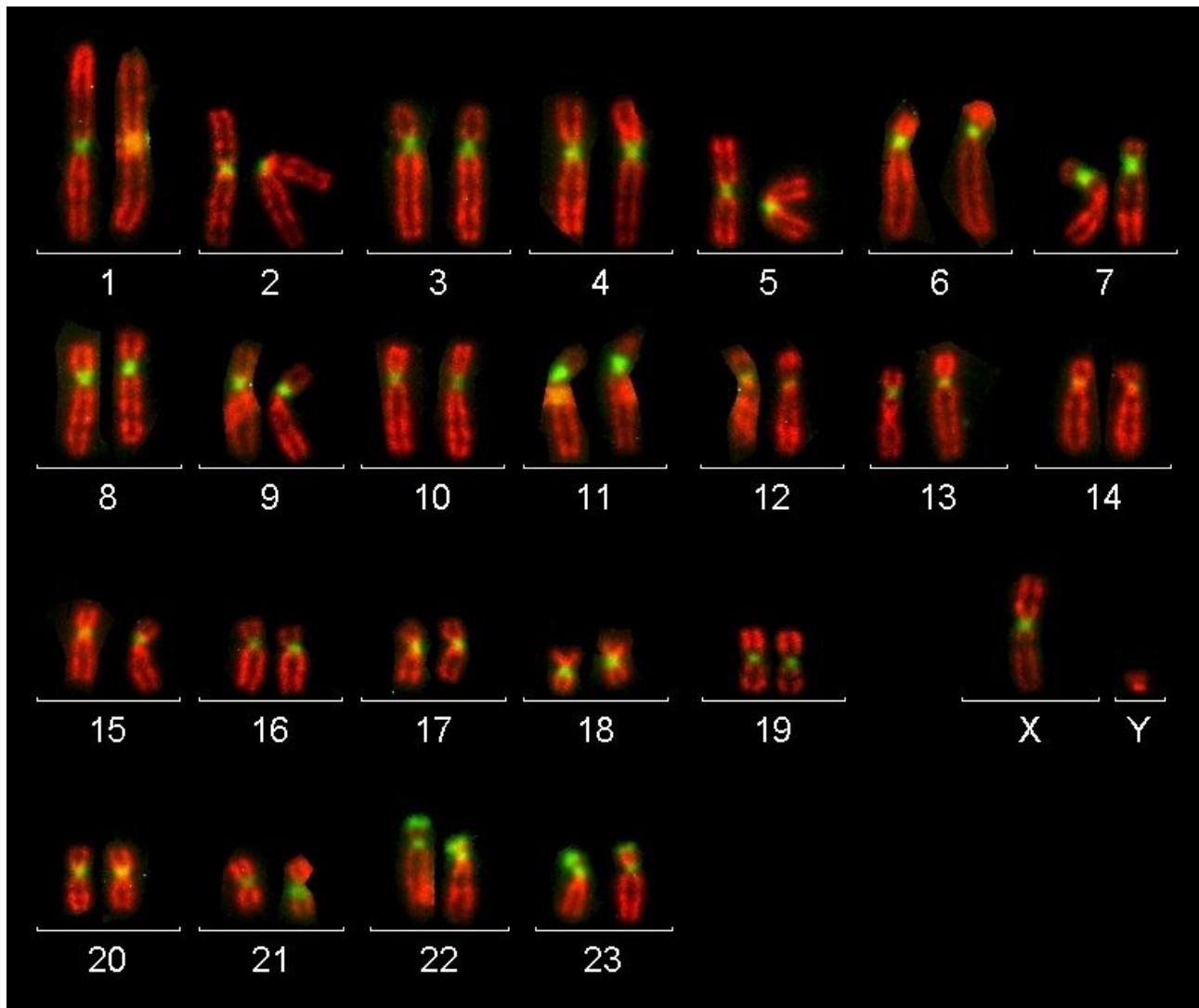




(e)

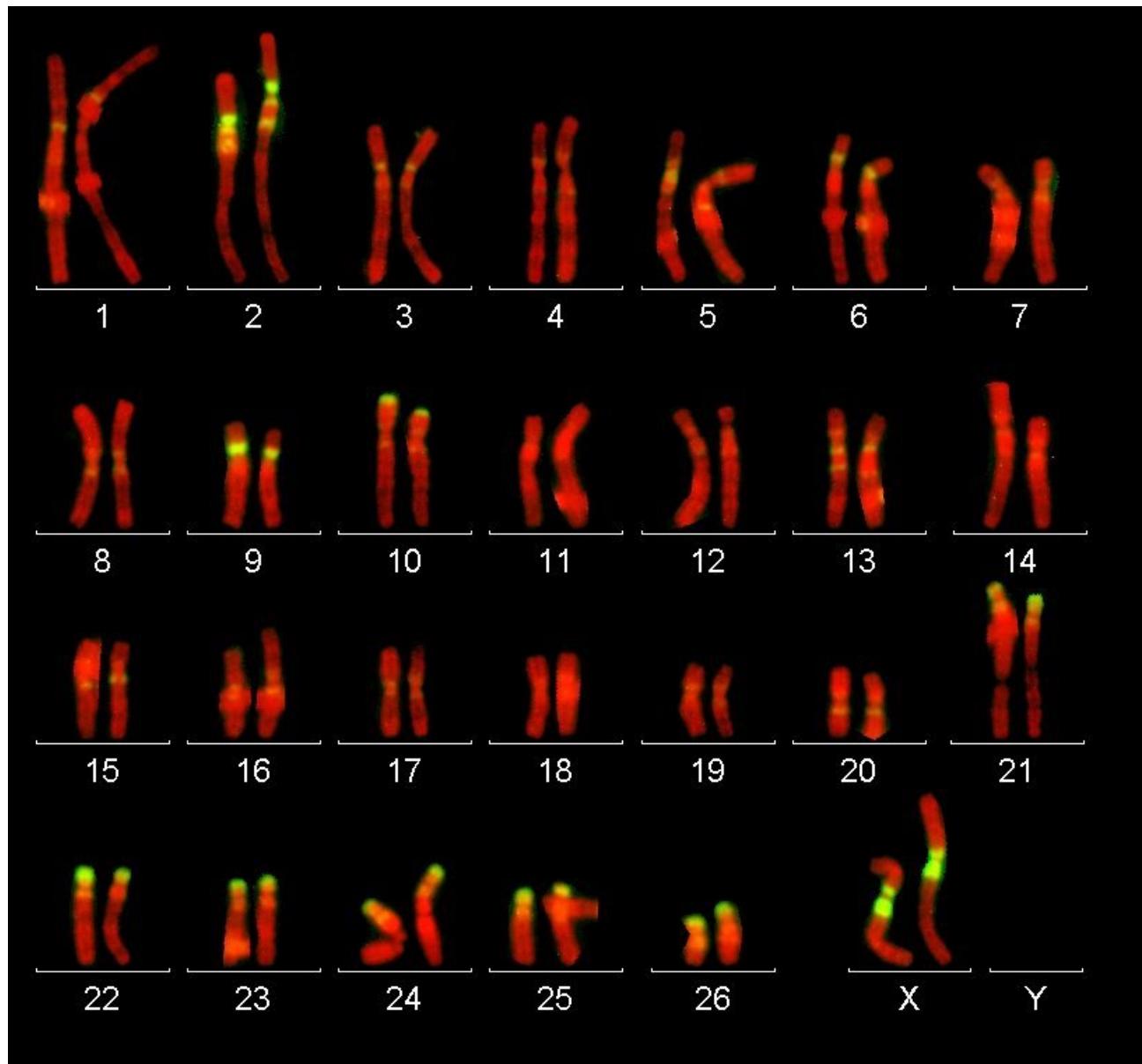




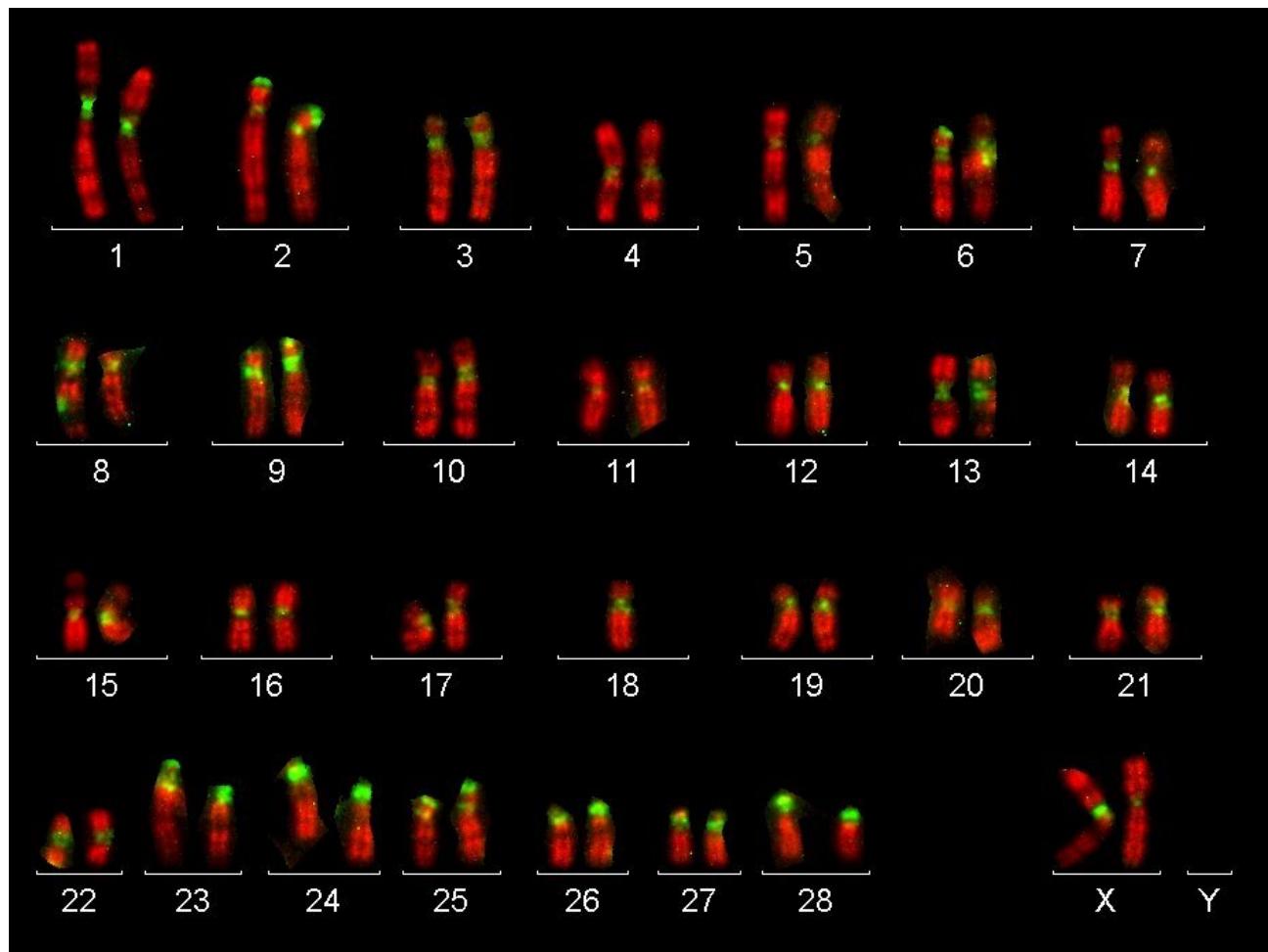


(a)

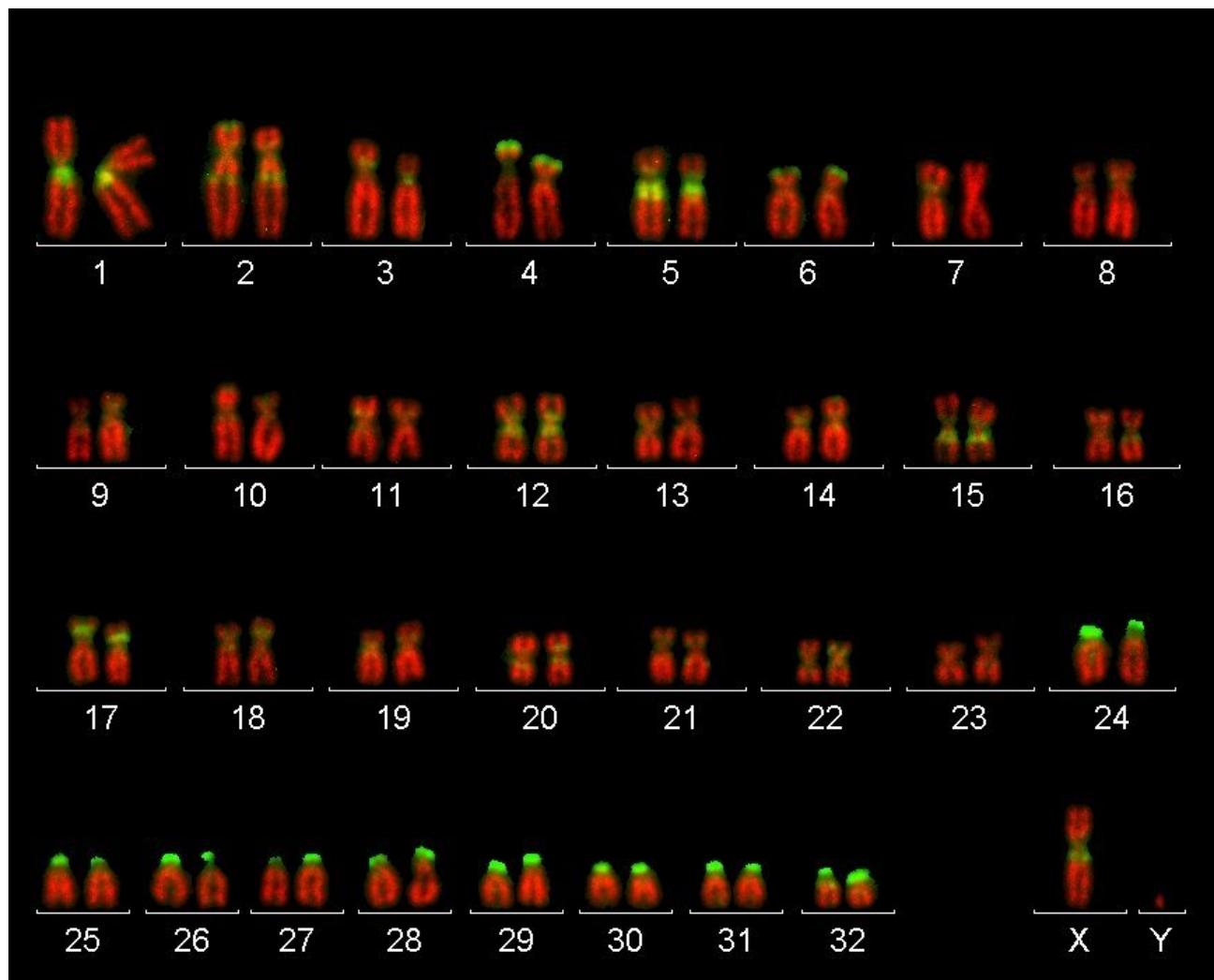
SUPPLEMENTARY FIGURE 3.7 – Distribution pattern of the C2 alpha satellite family on *Cercopithecini* chromosomes. Probes C2b (green) and C1b (not shown) are hybridized simultaneously to metaphase chromosomes (red) of (A) *A. nigroviridis*, (B) *E. patas*, (C) *C. roloway* and (D) *C. cebus*. Metaphases were karyotyped according to [Dutrillaux et al. \(1978, 1979, 1988a\)](#); [Moulin et al. \(2008\)](#).

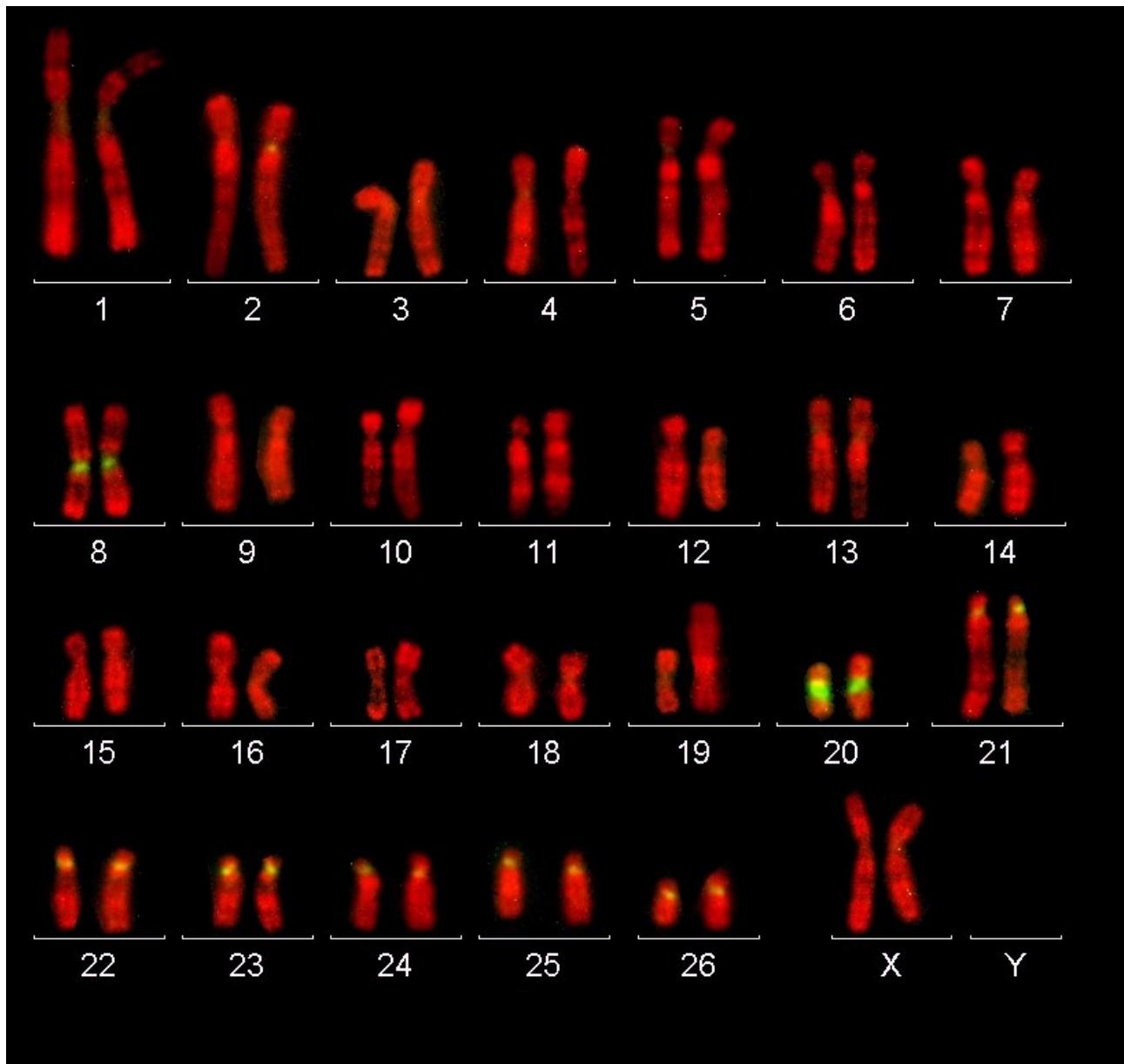


(b)

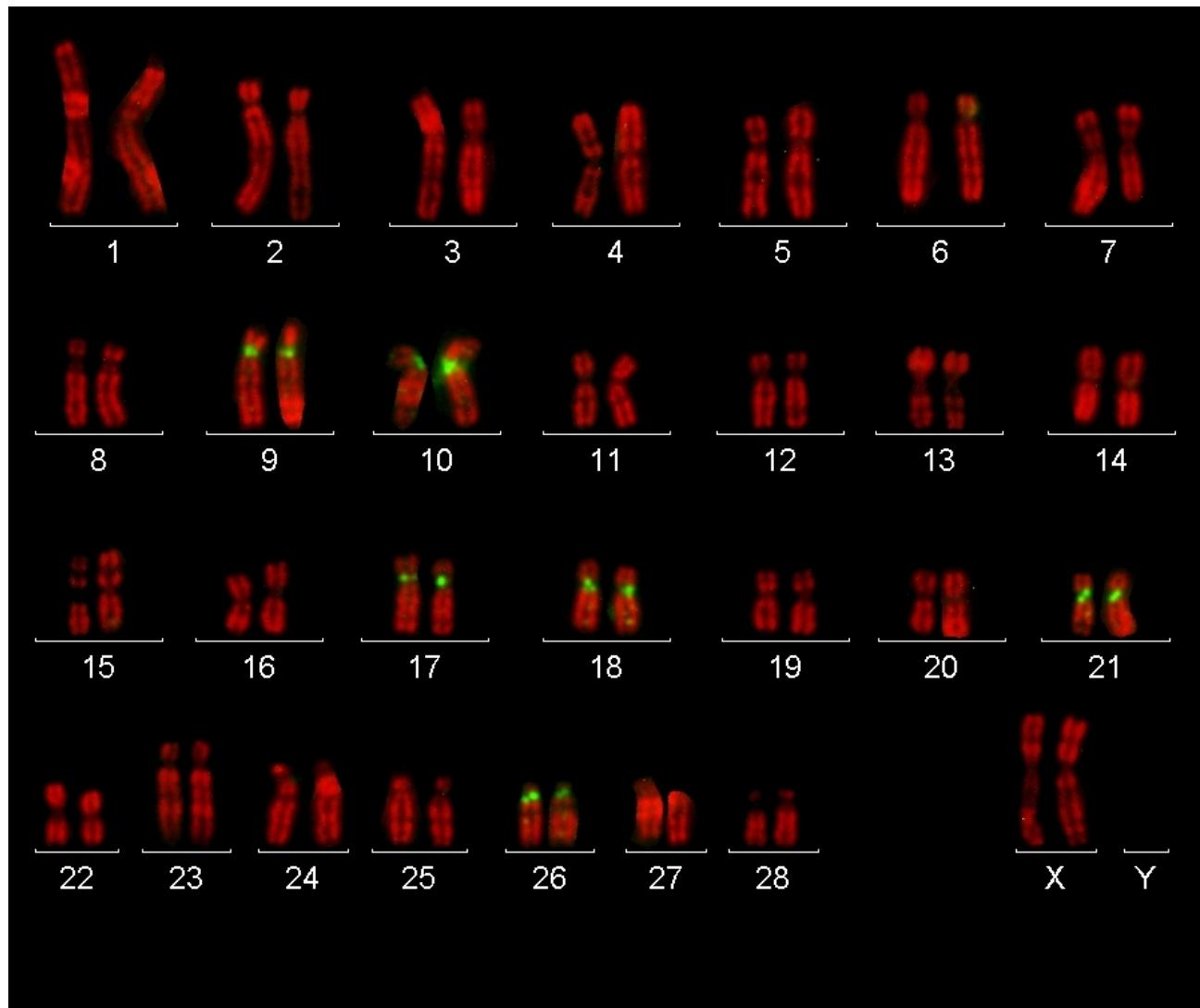


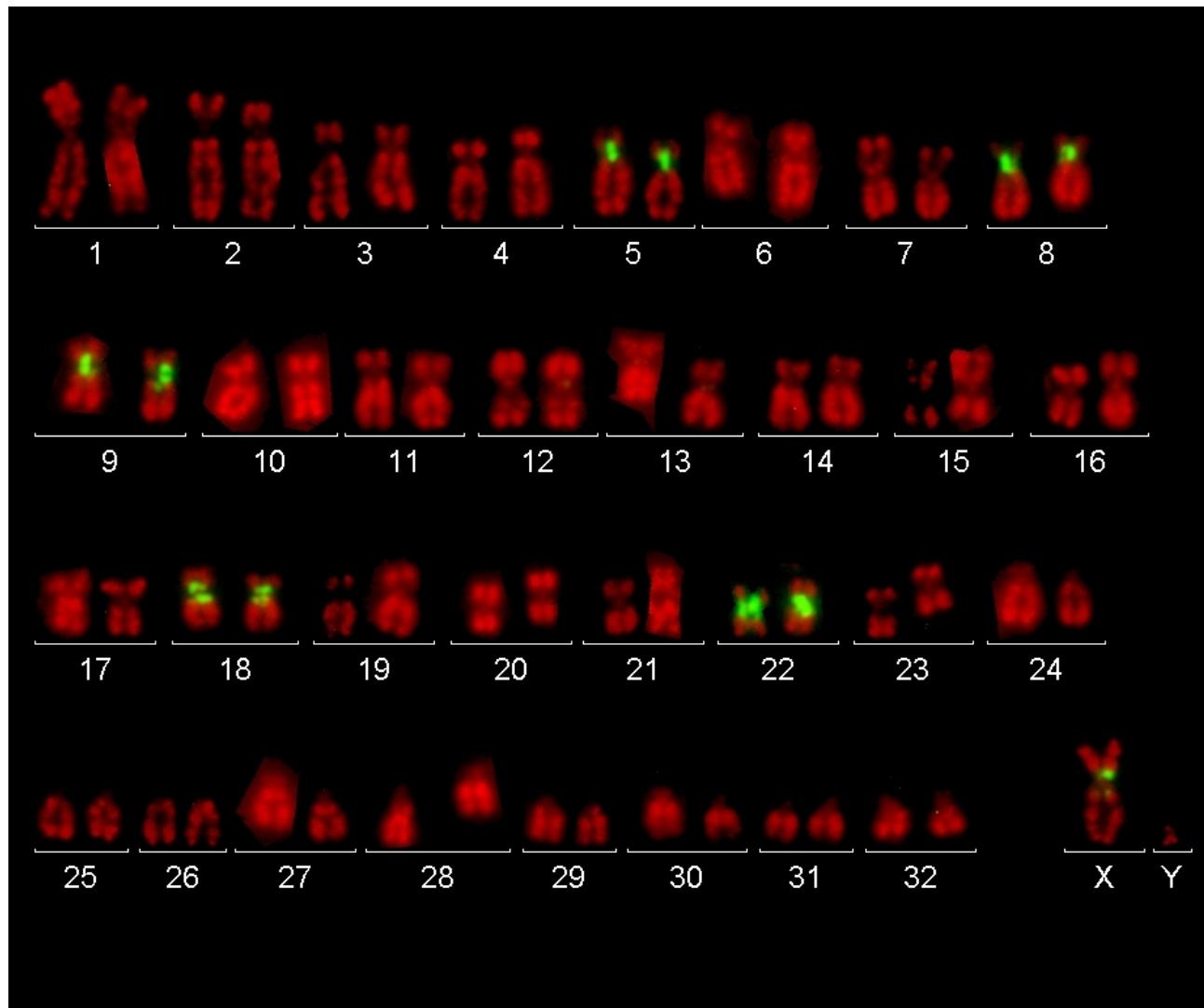
(c)



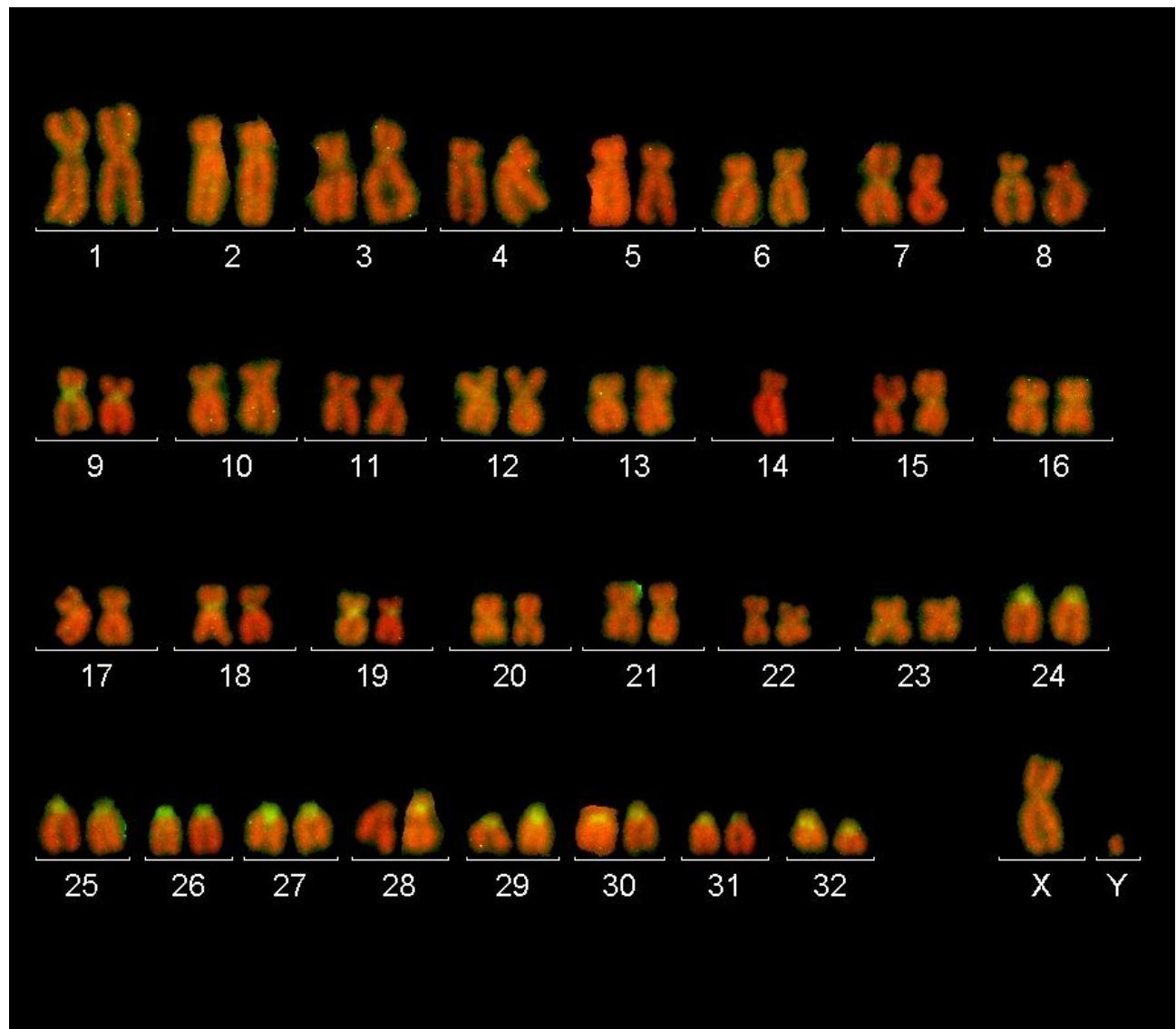


SUPPLEMENTARY FIGURE 3.8 – Distribution pattern of the C5 alpha satellite family on **Cercopithecini chromosomes**. Hybridization of probe C5a (green) on metaphase chromosomes (red) of (A) *E. patas* (label on pairs 2, 8, 20 and 21 to 26), (B) *C. roloway* (label on pairs 9, 10, 17, 18, 21 and 26) and (C) *C. cephus* (label on pairs 5, 8, 9, 18, 22 and X chromosome).





(c)



SUPPLEMENTARY FIGURE 3.9 – Distribution pattern of the C6 alpha satellite family on *C. cephush* chromosomes. Hybridization of probe C6a (green) on *C. cephush* metaphase chromosomes (red) (label on pairs 9, 18, 19 and 24 to 32).

4.3.2 Tables supplémentaires

SUPPLEMENTARY TABLE 3.1 – Patterns and labels of oligonucleotide probes used by FISH

Probe	Pattern (5'-3')	Label
C1a	TcCCtTtGcCaAtTcCAC	3'Cy3
C1b	AcTgCtCtGtGtTcTGtTa	3'Digoxygenin
C2a	TcACtTtGcAaAtTcCAC	5'AlexaFluor488
C2b	AcTgCtTtGtGtTcTGtTa	3'Cy5
C3a	CaGtTcTcAGaTtCcAcA	3'Digoxygenin
C3b	GcCcTaTaGtCtTcAaAg	3'Biotin
C4a	GtGaCtTcCACtTcAcTgA	3'Biotin
C4a'	TgAtGtGtGaCtTcCACt	3'Digoxygenin
C4b	aGcTGtATtTcGTgGaGc	3'Biotin
C5a	TgAaTtCaGaGaAcAcAg	3'Biotin
C6a	CaTTtTcCcTtCaAgAaTcC	3'Digoxygenin

NOTE - According to Cacheux et al. (2016a,b). LNA : lower case, classic nucleotides : upper case.

SUPPLEMENTARY TABLE 3.2 – Oligonucleotide probes used by FISH according to species

	C1a	C1b	C2a	C2b	C3a	C3b	C4a ⁺	C4b	C5a	C5a 68 °C	C6a
<i>C. mitis</i>	+		+		+		+		+	+	+
<i>C. nictitans</i>	+		+		+		+		+	+	+
<i>C. ascanius</i>	+		+		+		+		+	+	+
<i>C. erythrotis</i>	+		+		+		+		+	+	+
AC	<i>C. cephush</i>	+		+	+		+		+	+	+
	<i>C. pogonias</i>	+	+	+	+		+	+	+	+	+
	<i>C. mona</i>	+		+		+		+	+	+	+
	<i>C. diana</i>	+		+		+		+	+	+	+
	<i>C. roloway</i>	+		+	+		+	+	+	+	+
TC	<i>C. lhoesti</i>	+		+		+		+	+	+	+
	<i>C. solatus</i>	+	+	+	+	+	+	+	+	+	+
	<i>E. patas</i>	+		+	+		+		+	+	+
BC	<i>A. nigroviridis</i>	+	+	+	+	+	+	+	+	+	+
Out	<i>M. sylvanus</i>	+	+	+	+	+		+	+	+	+
	<i>C. angolensis</i>	+		+		+		+	+	+	+

NOTE - AC : arboreal clade. TC : terrestrial clade. BC : basal clade. Out : outgroups. C5a 68 °C : probes were washed at 68 °C after the hybridization step, instead of the usual 63 °C.

SUPPLEMENTARY TABLE 3.3 – PCR primers used for phylogenetic studies

Gene	PCR primers		Source
ABCA1	forward	CCTCCATCTTTAGCTCTACCTAC	Perelman et al. (2011)
	reverse	ACAAGAGCCTGGAGATTGGATAAC	
BRCA2	forward	CAGTTATTATTACCCCAGAGGCTGA	this study
	reverse	CAACTACACTACTCTGTAAACGTGCAG	
CFTR	forward	CTCTGTGAACACAGGATAGAAC	Perelman et al. (2011)
	reverse	TTACCTCCAGGAGGCTAAAAGCC	
DENND5A	forward	CCAGAGTTATCATGGCCAATC	Perelman et al. (2011)
	reverse	GTACCAAGCAAGAAGCTGGG	
ERC2	forward	AGCTCATCCTCCTGGTTAG	Perelman et al. (2011)
	reverse	CTCCTTGAGGATCTCAGCAAC	
LRPPRC-169	forward	CTTGGGATTGATGAGAATAATTTAG	this study
	reverse	GCATTCTGACAGATGACAAAGTT	
TTR	forward	CGGTGAGTGTCTGGGACA	this study
	reverse	GCCCTGGGTGAGGGAAAGA	
ZFX	forward	GAGAATCTATGTCAGCATAAAGCAG	this study
	reverse	GGAGTATGAGTGCTAAACCAA	

SUPPLEMENTARY TABLE 3.4 – Genbank sequences included in phylogenetic studies

Species	ABCA1	BRCA2	CFTR	DENND5A	ERC2
<i>Allenopithecus nigroviridis</i>	HM765265	HM763634	HM763500	HM759174	HM762247
<i>Chlorocebus aethiops</i>	HM765274	HM763647	HM763541	HM759189	HM762292
<i>Chlorocebus sabaeus</i>	HM765309	HM763649	HM763542	HM759190	HM762310
<i>Cercopithecus albogularis</i>	HM765273	HM763638	HM763528	HM759178	HM762227
<i>Cercopithecus ascanius</i>	HM765280	HM763639	HM763529	HM759179	HM762228
<i>Cercopithecus campbelli</i>	this study				
<i>Cercopithecus cebus</i>	this study				
<i>Cercopithecus diana</i>	HM765288	HM763641	HM763531	HM759181	HM762313
<i>Cercopithecus erythrotis</i>	x	this study	this study	this study	this study
<i>Cercopithecus hamlyni</i>	HM765293	HM763642	HM763532	HM759182	HM762233
<i>Cercopithecus lhoesti</i>	HM765294	HM763643	HM763533	HM759183	HM762225
<i>Cercopithecus mitis</i>	x	HM763644	this study	this study	this study
<i>Cercopithecus mona</i>	HM765416	this study	this study	HM759185	HM762291
<i>Cercopithecus nictitans</i>	this study				
<i>Cercopithecus petaurista</i>	HM765306	HM763645	HM763535	HM759186	HM762217
<i>Cercopithecus pogonias</i>	this study				
<i>Cercopithecus preussi</i>	this study				
<i>Cercopithecus roloway</i>	this study				
<i>Cercopithecus solatus</i>	this study				
<i>Cercopithecus wolfi</i>	HM765311	HM763646	HM763536	HM759187	HM762257
<i>Erythrocebus patas</i>	HM765317	this study	HM763547	HM759194	HM762311
<i>Miopithecus ogouensis</i>	HM765351	HM763670	x	HM759213	HM762215
<i>Macaca sylvanus</i>	HM765355	this study	HM763582	HM759208	HM762214
<i>Cercocebus torquatus</i>	HM765310	HM763637	HM763527	HM759177	HM762220
<i>Mandrillus sphinx</i>	HM765354	HM763669	HM763586	HM759212	HM762252
<i>Colobus guereza</i>	HM765292	HM763651	HM763543	HM759192	HM762317

Species	LRPPRC-169	SRY	TTR	ZFX
<i>Allenopithecus nigroviridis</i>	HM761071	AF284331	HM757630	HM757022
<i>Chlorocebus aethiops</i>	HM761081	JF293182	HM757645	HM757034
<i>Chlorocebus sabaeus</i>	HM761082	KC843986	HM757647	HM757036
<i>Cercopithecus albogularis</i>	HM761075	EF517797	HM757634	HM757025
<i>Cercopithecus ascanius</i>	this study	EF517798	HM757635	HM757026
<i>Cercopithecus campbelli</i>	this study	AY665639	x	this study
<i>Cercopithecus cephus</i>	this study	AY450883	this study	this study
<i>Cercopithecus diana</i>	this study	AY665633	HM757637	HM757028
<i>Cercopithecus erythrotis</i>	this study	x	this study	this study
<i>Cercopithecus hamlyni</i>	HM761077	AY450884	HM757638	HM757029
<i>Cercopithecus lhoesti</i>	this study	AY048067	HM757639	HM757030
<i>Cercopithecus mitis</i>	this study	AY048069	this study	this study
<i>Cercopithecus mona</i>	x	AF284332	HM757641	this study
<i>Cercopithecus nictitans</i>	this study	AY450887	this study	this study
<i>Cercopithecus petaurista</i>	HM761080	AY897617	HM757643	HM757032
<i>Cercopithecus pogonias</i>	this study	EF517799	this study	this study
<i>Cercopithecus preussi</i>	this study	AY665642	this study	this study
<i>Cercopithecus roloway</i>	this study	x	this study	this study
<i>Cercopithecus solatus</i>	this study	AY665643	this study	this study
<i>Cercopithecus wolfi</i>	x	AY450889	HM757644	HM757033
<i>Erythrocebus patas</i>	HM761084	AY048073	HM757651	HM757040
<i>Miopithecus ogouensis</i>	HM761102	x	HM757669	HM757059
<i>Macaca sylvanus</i>	HM761098	AF284326	HM757664	HM757054
<i>Cercocebus torquatus</i>	HM761074	HM757968	HM757633	HM757024
<i>Mandrillus sphinx</i>	HM761101	AF284330	HM757668	HM757058
<i>Colobus guereza</i>	HM761083	JF293183	HM757649	HM757038

(b)

SUPPLEMENTARY TABLE 3.5 – RBCell collection specimens included in cytogenetic studies

Species	Specimen Id number
<i>Allenopithecus nigroviridis</i>	1979-009
<i>Cercopithecus ascanius</i>	1979-017
<i>Cercopithecus cebus</i>	2012-006
<i>Cercopithecus diana</i>	1978-003
<i>Cercopithecus erythrotis</i>	1995-002
<i>Cercopithecus lhoesti</i>	1998-009
<i>Cercopithecus mitis</i>	1986-003
<i>Cercopithecus mona</i>	1979-014
<i>Cercopithecus nictitans</i>	2013-017
<i>Cercopithecus pogonias</i>	2001-027
<i>Cercopithecus roloway</i>	2003-064
<i>Cercopithecus solatus</i>	2012-028
<i>Erythrocebus patas</i>	1978-006
<i>Macaca sylvanus</i>	1971-019
<i>Colobus angolensis</i>	1978-005

Bibliographie

- Abdellah, Z., Ahmadi, A., Ahmed, S., Aimable, M., Ainscough, R., Almeida, J., et al. (2004). International human genome sequencing consortium. *Nature*, 409 :860–921. [1.3.3](#)
- Abràmoff, M. D., Magalhães, P. J., and Ram, S. J. (2004). Image processing with imagej. *Biophotonics international*, 11(7) :36–42. [2.1.6](#), [2.2.6](#), [2.3.6](#)
- Agrawal, A., Eastman, Q. M., and Schatz, D. G. (1998). Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature*, 394(6695) :744–751. [1.1.3](#)
- AIexandrov, I., Medvedev, L., Mashkova, T., Kisseelev, L., Romanova, L., and Yurov, Y. (1993). Definition of a new alpha satellite suprachromosomal family characterized by monomeric organization. *Nucleic acids research*, 21(9) :2209–2215. [2.1.2](#), [2.1.4](#)
- Aldrich-Blake, F. (1968). A fertile hybrid between two *Cercopithecus* spp. in the budongo forest, uganda. *Folia Primatologica*, 9(1) :15–21. [1.4.1](#)
- Aldrup-MacDonald, M. E. and Sullivan, B. A. (2014). The past, present, and future of human centromere genomics. *Genes*, 5(1) :33–50. [1.2.2](#)
- Alexandrov, I., Iurov, I., Mitkevich, S., and Gindilis, V. (1986). Chromosome organization of human alphoid DNA. *Doklady Akademii nauk SSSR*, 288(1) :242. [1.3.2](#)
- Alexandrov, I., Kazakov, A., Tumeneva, I., Shepelev, V., and Yurov, Y. (2001). Alpha-satellite DNA of primates : old and new families. *Chromosoma*, 110(4) :253–266. [1.3.1](#), [3.1](#), [1.3.1](#), [1.3.2](#), [2.1.2](#), [2.1.3](#), [1.7](#), [2.1.4](#), [2.1.6](#), [2.2.2](#), [3.2.1](#)
- Alexandrov, I., Mashkova, T., Akopian, T., Medvedev, L., Kisseelev, L., Mitkevich, S., and Yurov, Y. (1991). Chromosome-specific alpha satellites : two distinct families on human chromosome 18. *Genomics*, 11(1) :15–23. [1.3.2](#)
- Alexandrov, I., Mashkova, T., Romanova, L., Yurov, Y., and Kisseelev, L. (1993). Segment substitutions in alpha satellite DNA : unusual structure of human chromosome 3-specific alpha satellite repeat unit. *Journal of molecular biology*, 231(2) :516–520. [3.1](#), [1.3.1](#), [2.2.2](#)

- Alexandrov, I., Mitkevich, S., and Yurov, Y. (1988). The phylogeny of human chromosome specific alpha satellites. *Chromosoma*, 96(6) :443–453. [1.3.1](#), [1.3.2](#), [1.3.2](#), [2.1.2](#), [2.3.2](#)
- Alkan, C., Cardone, M. F., Catacchio, C. R., Antonacci, F., O'Brien, S. J., Ryder, O. A., Purgato, S., Zoli, M., Della Valle, G., Eichler, E. E., et al. (2011). Genome-wide characterization of centromeric satellites from multiple mammalian genomes. *Genome research*, 21(1) :137–145. [1.2.2](#)
- Alkan, C., Eichler, E. E., Bailey, J. A., Cenk Sahinalp, S., and Tüzün, E. (2004). The role of unequal crossover in alpha-satellite DNA evolution : a computational analysis. *Journal of Computational Biology*, 11(5) :933–944. [1.3.3](#)
- Alkan, C., Ventura, M., Archidiacono, N., Rocchi, M., Sahinalp, S. C., and Eichler, E. E. (2007). Organization and evolution of primate centromeric DNA from whole-genome shotgun sequence data. *PLoS Comput Biol*, 3(9) :e181. [1.3.1](#), [1.3.2](#), [1.3.3](#), [2.1.2](#), [2.2.2](#), [2.3.2](#), [3.2.1](#)
- Allen, E. S. (2002). *Long Term Hybridization and the Maintenance of Species Identity in Orioles (Icterus)*. Indiana University. [1.4.2](#)
- Allen, W. L., Stevens, M., and Higham, J. P. (2014). Character displacement of Cercopithecini primate visual signals. *Nature communications*, 5. [1.4.1](#)
- Altemose, N., Miga, K. H., Maggioni, M., and Willard, H. F. (2014). Genomic characterization of large heterochromatic gaps in the human genome assembly. *PLOS Comput Biol*, 10(5) :e1003628. [1.2.2](#)
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3) :403–410. [2.1.6](#), [2.2.6](#)
- Alves, G., Seuánez, H. N., and Fanning, T. (1994). Alpha satellite DNA in neotropical primates (Platyrrhini). *Chromosoma*, 103(4) :262–267. [3.1](#), [1.3.1](#), [2.1.2](#), [2.2.2](#), [2.2.4](#)
- Alves, G., Seuánez, H. N., and Fanning, T. (1998). A clade of New World primates with distinctive alphoid satellite DNAs. *Molecular phylogenetics and evolution*, 9(2) :220–224. [1.3.1](#), [3.1](#), [1.3.1](#), [2.2.2](#)
- Amano, M., Suzuki, A., Hori, T., Backer, C., Okawa, K., Cheeseman, I. M., and Fukagawa, T. (2009). The CENP-S complex is essential for the stable assembly of outer kinetochore structure. *The Journal of cell biology*, 186(2) :173–182. [1.2.1](#)
- Amor, D. J. and Choo, K. A. (2002). Neocentromeres : role in human disease, evolution, and centromere study. *The American Journal of Human Genetics*, 71(4) :695–714. [1.2.2](#)
- Archidiacono, N., Antonacci, R., Marzella, R., Finelli, P., Lonoce, A., and Rocchi, M. (1995). Comparative mapping of human alphoid sequences in great apes using fluorescence in situ hybridization. *Genomics*, 25(2) :477–484. [1.3.2](#), [2.1.2](#), [2.3.2](#)

- Armour, J., Crosier, M., and Jeffreys, A. (1996). Distribution of tandem repeat polymorphism within minisatellite MS621 (D5S110). *Annals of human genetics*, 60(1) :11–20. [1.1.2](#)
- Arnold, C., Matthews, L. J., and Nunn, C. L. (2010). The 10kTrees website : a new online resource for primate phylogeny. *Evolutionary Anthropology : Issues, News, and Reviews*, 19(3) :114–118. [3.1](#)
- Arnold, N., Stanyon, R., Jauch, A., O'brien, P., and Wienberg, J. (1996). Identification of complex chromosome rearrangements in the gibbon by fluorescent in situ hybridization (FISH) of a human chromosome 2q specific microlibrary, yeast artificial chromosomes, and reciprocal chromosome painting. *Cytogenetic and Genome Research*, 74(1-2) :80–85. [1.4.3](#)
- Assaad, F. F., Tucker, K. L., and Signer, E. R. (1993). Epigenetic repeat-induced gene silencing (RIGS) in Arabidopsis. *Plant molecular biology*, 22(6) :1067–1085. [1.1.3](#)
- Assum, G., Fink, T., Steinbeißer, T., and Fisel, K. J. (1993). Analysis of human extrachromosomal DNA elements originating from different β -satellite subfamilies. *Human genetics*, 91(5) :489–495. [1.3.3](#)
- Ayala, F. J. and Coluzzi, M. (2005). Chromosome speciation : humans, Drosophila, and mosquitoes. *Proceedings of the National Academy of Sciences*, 102(suppl 1) :6535–6542. [1.4.3](#)
- Baicharoen, S., Arsaithamkul, V., Hirai, Y., Hara, T., Koga, A., and Hirai, H. (2012). In situ hybridization analysis of gibbon chromosomes suggests that amplification of alpha satellite DNA in the telomere region is confined to two of the four genera. *Genome*, 55(11) :809–812. [1.3.2](#)
- Baldini, A., Archidiacono, N., Carbone, R., Bolino, A., Shridhar, V., Miller, O., Miller, D., Ward, D., and Rocchi, M. (1992). Isolation and comparative mapping of a human chromosome 20-specific α -satellite DNA clone. *Cytogenetic and Genome Research*, 59(1) :12–16. [1.3.2](#)
- Baldini, A., Miller, D. A., Miller, O. J., Ryder, O. A., and Mitchell, A. R. (1991). A chimpanzee-derived chromosome-specific alpha satellite DNA sequence conserved between chimpanzee and human. *Chromosoma*, 100(3) :156–161. [1.3.2](#), [2.2.2](#)
- Bejerano, G., Lowe, C. B., Ahituv, N., King, B., Siepel, A., Salama, S. R., Rubin, E. M., Kent, W. J., and Haussler, D. (2006). A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature*, 441(7089) :87–90. [1.1.3](#), [2.2.2](#)
- Bender, M. A. and Mettler, L. E. (1958). Chromosome studies of primates. *Science*, 128(3317) :186–190. [1.4.3](#)
- Benfante, R., Landsberger, N., Maiorano, D., and Badaracco, G. (1990). A binding protein (p82 protein) recognizes specifically the curved heterochromatic DNA in *Artemia franciscana*. *Gene*, 94(2) :217–222. [1.2.2](#)

- Bergmann, J. H., Rodríguez, M. G., Martins, N. M., Kimura, H., Kelly, D. A., Masumoto, H., Larionov, V., Jansen, L. E., and Earnshaw, W. C. (2011). Epigenetic engineering shows H3K4me2 is required for HJURP targeting and CENP-A assembly on a synthetic human kinetochore. *The EMBO journal*, 30(2) :328–340. [1.2.1](#)
- Bernard, P., Maure, J.-F., Partridge, J. F., Genier, S., Javerzat, J.-P., and Allshire, R. C. (2001). Requirement of heterochromatin for cohesion at centromeres. *Science*, 294(5551) :2539–2542. [1.2](#)
- Best, R., Diamond, D., Crawford, E., Grass, F., Janish, C., Lear, T., Soenksen, D., Szalay, A., and Moore, C. (1998). Baboon/human homologies examined by spectral karyotyping (SKY) : a visual comparison. *Cytogenetic and Genome Research*, 82(1-2) :83–87. [1.4.3](#)
- Bigoni, F., Koehler, U., Stanyon, R., Ishida, T., and Wienberg, J. (1997). Fluorescence in situ hybridization establishes homology between human and silvered leaf monkey chromosomes, reveals reciprocal translocations between chromosomes homologous to human Y/5, 1/9, and 6/16, and delineates an X 1 X 2 Y 1 Y 2/X 1 X 1 X 2 X 2 sex-chromosome system. *American journal of physical anthropology*, 102(3) :315–327. [1.4.3](#)
- Bigot, Y., Hamelin, M., and Periquet, G. (1990). Heterochromatin condensation and evolution of unique satellite-DNA families in two parasitic wasp species : Diadromus pulchellus and Eupelmus vuilleti (Hymenoptera). *Molecular biology and evolution*, 7(4) :351–364. [1.2.2](#)
- Bingham, P. M., Kidwell, M. G., and Rubin, G. M. (1982). The molecular basis of PM hybrid dysgenesis : the role of the P element, a P-strain-specific transposon family. *Cell*, 29(3) :995–1004. [1.1.1](#)
- Biscotti, M. A., Olmo, E., and Heslop-Harrison, J. P. (2015). Repetitive DNA in eukaryotic genomes. *Chromosome Research*, 23(3) :415–420. [1.1.2](#)
- Bloom, K. S. (2014). Centromeric heterochromatin : the primordial segregation machine. *Annual review of genetics*, 48 :457. [1.2.1](#)
- Blower, M. D., Sullivan, B. A., and Karpen, G. H. (2002). Conserved organization of centromeric chromatin in flies and humans. *Developmental cell*, 2(3) :319–330. [1.2.1](#)
- Bodor, D. L., Mata, J. F., Sergeev, M., David, A. F., Salimian, K. J., Panchenko, T., Cleveland, D. W., Black, B. E., Shah, J. V., and Jansen, L. E. (2014). The quantitative architecture of centromeric chromatin. *Elife*, 3 :e02137. [1.2.1](#)
- Borgaonkar, D. S. (1966). A list of chromosome numbers in primates. *Journal of Heredity*, 57(2) :60–64. [1.4.3](#)
- Bowers, J. E., Dangl, G. S., and Meredith, C. P. (1999). Development and characterization of additional microsatellite DNA markers for grape. *American Journal of Enology and Viticulture*, 50(3) :243–246. [1.1.2](#)

- Bragg, L. M., Stone, G., Butler, M. K., Hugenholtz, P., and Tyson, G. W. (2013). Shining a light on dark sequencing : characterising errors in Ion Torrent PGM data. *PLoS Comput Biol*, 9(4) :e1003031. [2.1.3](#)
- Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G. J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell*, 128(6) :1089–1103. [1.1.3](#)
- Broccoli, D., Trevor, K. T., Miller, O. J., and Miller, D. A. (1991). Isolation of a variant family of mouse minor satellite DNA that hybridizes preferentially to chromosome 4. *Genomics*, 10(1) :68–74. [1.2.2](#)
- Brown, J. D. and O'Neill, R. J. (2010). Chromosomes, conflict, and epigenetics : chromosomal speciation revisited. *Annual review of genomics and human genetics*, 11 :291–316. [1.4.3](#), [4.21](#)
- Brown, T. A. (2002). Genome replication. [3.9](#)
- Brugiere, D., Gautier, J.-P., Moungazi, A., and Gautier-Hion, A. (2002). Primate diet and biomass in relation to vegetation composition and fruiting phenology in a rain forest in Gabon. *International Journal of Primatology*, 23(5) :999–1024. [1.4.1](#)
- Brumfield, R. T., Jernigan, R. W., McDonald, D. B., and Braun, M. J. (2001). Evolutionary implications of divergent clines in an avian (*Manacus* : Aves) hybrid zone. *Evolution*, 55(10) :2070–2087. [1.4.2](#)
- Buckland, R. A. (1983). Comparative structure and evolution of goat and sheep satellite I DNAs. *Nucleic acids research*, 11(5) :1349–1360. [1.2.2](#)
- Bühler, M., Verdel, A., and Moazed, D. (2006). Tethering RITS to a nascent transcript initiates RNAi-and heterochromatin-dependent gene silencing. *Cell*, 125(5) :873–886. [1.1.3](#)
- Bulazel, K., Metcalfe, C., Ferreri, G. C., Yu, J., Eldridge, M. D., and O'Neill, R. J. (2006). Cytogenetic and molecular evaluation of centromere-associated DNA sequences from a marsupial (*Macropodidae* : *Macropus rufogriseus*) X chromosome. *Genetics*, 172(2) :1129–1137. [1.2.2](#), [2.4](#)
- Burkin, D. J., Broad, T. E., and Jones, C. (1996). The chromosomal distribution and organization of sheep satellite I and II centromeric DNA using characterized sheep-hamster somatic cell hybrids. *Chromosome Research*, 4(1) :49–55. [1.2.2](#)
- Butynski, T. M. (1990). Comparative ecology of blue monkeys (*Cercopithecus mitis*) in high- and low-density subpopulations. *Ecological Monographs*, 60(1) :1–26. [1.4.1](#)
- Byrd, K. and Corces, V. G. (2003). Visualization of chromatin domains created by the gypsy insulator of *Drosophila*. *The Journal of cell biology*, 162(4) :565–574. [1.1.3](#)

- Cacheux, L., Ponger, L., Gerbault-Seureau, M., Loll, F., Gey, D., Richard, F. A., and Escudé, C. (2016a). Copious alpha satellite diversification events in the *Cercopithecus pogonias* lineage. [2.3.2](#), [2.3.3](#), [2.3.3](#), [2.3.3](#), [2.3.4](#), [2.3.4](#), [2.3.4](#), [2.3.6](#), [3.1](#)
- Cacheux, L., Ponger, L., Gerbault-Seureau, M., Richard, F. A., and Escudé, C. (2016b). Diversity and distribution of alpha satellite DNA in the genome of an Old World monkey : *Cercopithecus solatus*. *BMC genomics*, 17(1) :916. [2.2.2](#), [2.2.3](#), [2.2.4](#), [2.2.6](#), [2.3.2](#), [2.3.3](#), [2.3.3](#), [2.3.4](#), [2.3.4](#), [2.3.4](#), [2.3.6](#), [2.11](#), [3.1](#)
- Canavez, F., Alves, G., Fanning, T. G., and Seuánez, H. N. (1996). Comparative karyology and evolution of the Amazonian *Callithrix* (Platyrrhini, Primates). *Chromosoma*, 104(5) :348–357. [1.4.3](#)
- Capozzi, O., Carbone, L., Stanyon, R. R., Marra, A., Yang, F., Whelan, C. W., De Jong, P. J., Rocchi, M., and Archidiacono, N. (2012). A comprehensive molecular cytogenetic analysis of chromosome rearrangements in gibbons. *Genome research*, 22(12) :2520–2528. [1.4.3](#)
- Capy, P., Gasperi, G., Biémont, C., and Bazin, C. (2000). Stress and transposable elements : co-evolution or useful parasites ? *Heredity*, 85(2) :101–106. [2.3.4](#)
- Carbone, L., Harris, R. A., Vessere, G. M., Mootnick, A. R., Humphray, S., Rogers, J., Kim, S. K., Wall, J. D., Martin, D., Jurka, J., et al. (2009). Evolutionary breakpoints in the gibbon suggest association between cytosine methylation and karyotype evolution. *PLoS Genet*, 5(6) :e1000538. [1.4.3](#)
- Cardini, A. and Elton, S. (2008). Does the skull carry a phylogenetic signal ? Evolution and modularity in the guenons. *Biological Journal of the Linnean Society*, 93(4) :813–834. [2.3.4](#)
- Cardone, M., Ballarati, L., Ventura, M., Rocchi, M., Marozzi, A., Ginelli, E., and Meneveri, R. (2004). Evolution of beta satellite DNA sequences : evidence for duplication-mediated repeat amplification and spreading. *Molecular biology and evolution*, 21(9) :1792–1799. [1.2.2](#)
- Carmell, M. A., Girard, A., van de Kant, H. J., Bourc'his, D., Bestor, T. H., de Rooij, D. G., and Hannon, G. J. (2007). MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Developmental cell*, 12(4) :503–514. [1.1.3](#)
- Caspersson, T., Farber, S., Foley, G., Kudynowski, J., Modest, E., Simonsson, E., Wagh, U., and Zech, L. (1968). Chemical differentiation along metaphase chromosomes. *Experimental cell research*, 49(1) :219–222. [1.4.3](#)
- Catacchio, C., Ragone, R., Chiatante, G., and Ventura, M. (2015). Organization and evolution of Gorilla centromeric DNA from old strategies to new approaches. *Scientific reports*, 5. [1.3.1](#), [3.1](#), [1.3.1](#), [2.1.2](#), [2.2.2](#), [2.2.4](#), [2.3.2](#)
- Cellamare, A., Catacchio, C., Alkan, C., Giannuzzi, G., Antonacci, F., Cardone, M., Della Valle, G., Malig, M., Rocchi, M., Eichler, E., et al. (2009). New insights into centromere organization

- and evolution from the white-cheeked gibbon and marmoset. *Molecular biology and evolution*, 26(8) :1889–1900. [1.3.1](#), [2.2.4](#)
- Cerveau, N., Leclercq, S., Bouchon, D., and Cordaux, R. (2011). Evolutionary dynamics and genomic impact of prokaryote transposable elements. In *Evolutionary biology—concepts, biodiversity, macroevolution and genome evolution*, pages 291–312. Springer. [1.1.1](#)
- Chan, F. L., Marshall, O. J., Saffery, R., Kim, B. W., Earle, E., Choo, K. A., and Wong, L. H. (2012). Active transcription and essential role of RNA polymerase II at the centromere during mitosis. *Proceedings of the National Academy of Sciences*, 109(6) :1979–1984. [1.2.2](#), [2.2.2](#)
- Charlesworth, B., Sniegowski, P., and Stephan, W. (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. [1.1.2](#)
- Chaves, R., Adega, F., Heslop-Harrison, J., Guedes-Pinto, H., and Wienberg, J. (2003). Complex satellite DNA reshuffling in the polymorphic t (1 ; 29) Robertsonian translocation and evolutionarily derived chromosomes in cattle. *Chromosome Research*, 11(7) :641–648. [1.2.2](#)
- Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C., and Patrinos, G. P. (2007). Gene conversion : mechanisms, evolution and human disease. *Nature Reviews Genetics*, 8(10) :762–775. [3.8](#)
- Cheney, D. L. (1987). Interactions and relationships between groups. *Primate societies*. [1.4.1](#)
- Chiarelli, B. (1968). Chromosome polymorphism in the species of the genus Cercopithecus. *Cytologia*, 33(1) :1–16. [1.4.3](#)
- Chism, J. (1978). Relationships between patas infants and group members other than the mother. *Recent advances in primatology*, 1 :173–176. [1.4.1](#)
- Chism, J. (1999). Intergroup encounters in wild patas monkeys (*Erythrocebus patas*) in Kenya. *American Journal of Primatology*, (46) :43. [1.4.1](#)
- Chism, J. and Rogers, W. (2004). Grooming and social cohesion in patas monkeys and other guenons. In *The guenons : Diversity and adaptation in African monkeys*, pages 233–244. Springer. [1.4.1](#)
- Chism, J., Rowell, T., and Olson, D. (1984). Life history patterns of female patas monkeys. *Female primates : studies by women primatologists*. New York : Alan R. Liss. p, pages 175–190. [1.4.1](#)
- Chism, J. B. and Rowell, T. (1986). Mating and residence patterns of male patas monkeys. *Ethology*, 72(1) :31–39. [1.4.1](#)
- Choo, K., Earle, E., and McQuillan, C. (1990). A homologous subfamily of satellite III DNA on human chromosomes 14 and 22. *Nucleic acids research*, 18(19) :5641–5648. [2.1.4](#), [2.3.4](#)

- Chu, E. and Bender, M. (1961). Chromosome cytology and evolution in primates. *Science*, 133(3462) :1399–1405. [1.4.3](#)
- Clarke, L., Baum, M., Marschall, L., Ngan, V., and Steiner, N. (1993). Structure and function of Schizosaccharomyces pombe centromeres. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 58, pages 687–695. Cold Spring Harbor Laboratory Press. [1.2.2](#)
- Clemente, I., Ponsà, M., Garcia, M., and Egoscue, J. (1990). Chromosome evolution in the Cercopithecidae and its relationship to human fragile sites and neoplasia. *International journal of primatology*, 11(4) :377–398. [1.4.3](#)
- Cohen, A., Huh, T., and Helleiner, C. (1973). Transcription of satellite DNA in mouse L-cells. *Canadian journal of biochemistry*, 51(5) :529–532. [1.2.2](#)
- Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., Walters, L., et al. (1998). New goals for the US human genome project : 1998-2003. *Science*, 282(5389) :682–689. [1.3.3](#)
- Coluzzi, M. (1981). Spatial distribution of chromosomal inversions and speciation in Anopheline mosquitoes. *Progress in clinical and biological research*, 96 :143–153. [1.4.3](#)
- Comings, D. (1973). Biochemical mechanisms of chromosome banding and color banding with acridine orange. In *Nobel Symposium*, volume 23, pages 293–299. [1.4.3](#)
- Consigliere, S., Stanyon, R., Koehler, U., Agoramoothy, G., and Wienberg, J. (1996). Chromosome painting defines genomic rearrangements between red howler monkey subspecies. *Chromosome Research*, 4(4) :264–270. [1.4.3](#)
- Consigliere, S., Stanyon, R., Koehler, U., Arnold, N., and Wienberg, J. (1998). In situ hybridization (FISH) maps chromosomal homologies between Alouatta belzebul (Platyrrhini, Cebidae) and other primates and reveals extensive interchromosomal rearrangements between howler monkey genomes. *American journal of primatology*, 46(2) :119–133. [1.4.3](#)
- Cords, M. (1986). Interspecific and intraspecific variation in diet of two forest guenons, Cercopithecus ascanius and C. mitis. *The Journal of Animal Ecology*, pages 811–827. [1.4.1](#)
- Cords, M. (1988). Mating systems of forest guenons : a preliminary review. *A primate radiation : Evolutionary biology of the African guenons*, pages 323–339. [1.4.1](#)
- Cords, M. (1990). Mixed-species association of East African guenons : General patterns or specific examples ? *American Journal of Primatology*, 21(2) :101–114. [1.4.1](#)
- Cords, M. (2000). Grooming partners of immature blue monkeys (Cercopithecus mitis) in the Kakamega. *International Journal of Primatology*, 21(2) :239–254. [1.4.1](#)
- Cords, M. (2004). When are there influxes in blue monkey groups ? In *The Guenons : Diversity and adaptation in African monkeys*, pages 189–201. Springer. [1.4.1](#)

- Cords, M. et al. (1987). *Mixed-species association of Cercopithecus monkeys in the Kakamega Forest, Kenya*. Number DEPOSITO FV. [1.4.1](#)
- Cords, M. and Fuller, J. L. (2010). Infanticide in *Cercopithecus mitis stuhlmanni* in the Kakamega Forest, Kenya : variation in the occurrence of an adaptive behavior. *International Journal of Primatology*, 31(3) :409–431. [1.4.1](#)
- Cournac, A., Koszul, R., and Mozziconacci, J. (2016). The 3D folding of metazoan genomes correlates with the association of similar repetitive elements. *Nucleic acids research*, 44(1) :245–255. [1.1.3](#), [2.2.2](#)
- Cremer, T., Lichter, P., Borden, J., Ward, D., and Manuelidis, L. (1988). Detection of chromosome aberrations in metaphase and interphase tumor cells by in situ hybridization using chromosome-specific library probes. *Human genetics*, 80(3) :235–246. [1.4.3](#)
- D'Aiuto, L., Antonacci, R., Marzella, R., Archidiacono, N., and Rocchi, M. (1993). Cloning and comparative mapping of a human chromosome 4-specific alpha satellite DNA sequence. *Genomics*, 18(2) :230–235. [1.3.2](#)
- D'aiuto, L., Barsanti, P., Mauro, S., Cserpan, I., Lanave, C., and Ciccarese, S. (1997). Physical relationship between satellite I and II DNA in centromeric regions of sheep chromosomes. *Chromosome Research*, 5(6) :375–381. [1.2.2](#)
- Dawlaty, M. M., Malureanu, L., Jegannathan, K. B., Kao, E., Sustmann, C., Tahk, S., Shuai, K., Grosschedl, R., and van Deursen, J. M. (2008). Resolution of sister centromeres requires RanBP2-mediated SUMOylation of topoisomerase II α . *Cell*, 133(1) :103–115. [1.2.2](#)
- de Laat, W. and Dekker, J. (2012). 3C-based technologies to study the shape of the genome. *Methods (San Diego, Calif.)*, 58(3). [1.1.3](#)
- Deininger, P. L., Moran, J. V., Batzer, M. A., and Kazazian, H. H. (2003). Mobile elements and mammalian genome evolution. *Current opinion in genetics & development*, 13(6) :651–658. [1.3.2](#)
- Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *science*, 295(5558) :1306–1311. [1.1.3](#)
- Dewannieux, M., Esnault, C., and Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. *Nature genetics*, 35(1) :41–48. [1.1.1](#)
- Dimitriadis, E. K., Weber, C., Gill, R. K., Diekmann, S., and Dalal, Y. (2010). Tetrameric organization of vertebrate centromeric nucleosomes. *Proceedings of the National Academy of Sciences*, 107(47) :20317–20322. [1.2.1](#)
- Disotell, T. R. and Raaum, R. L. (2004). Molecular timescale and gene tree incongruence in the guenons. In *The guenons : Diversity and adaptation in African monkeys*, pages 27–36. Springer. [1.4.2](#)

- Donehower, L., Furlong, C., Gillespie, D., and Kurnit, D. (1980). DNA sequence of baboon highly repeated DNA : evidence for evolution by nonrandom unequal crossovers. *Proceedings of the National Academy of Sciences*, 77(4) :2129–2133. [1.3.1](#), [1.3.1](#)
- Dorer, D. R. and Henikoff, S. (1994). Expansions of transgene repeats cause heterochromatin formation and gene silencing in *Drosophila*. *Cell*, 77(7) :993–1002. [1.1.3](#)
- Dover, G. (1982). Molecular drive : a cohesive mode of species evolution. *Nature*, 299(5879) :111–117. [1.3.3](#)
- Dover, G. A. (1986). Molecular drive in multigene families : how biological novelties arise, spread and are assimilated. *Trends in genetics*, 2 :159–165. [2.2.4](#)
- Dumas, F., Stanyon, R., Sineo, L., Stone, G., and Bigoni, F. (2007). Phylogenomics of species from four genera of New World monkeys by flow sorting and reciprocal chromosome painting. *BMC evolutionary biology*, 7(2) :1. [1.4.3](#)
- Durfy, S. J. and Willard, H. F. (1990). Concerted evolution of primate alpha satellite DNA : evidence for an ancestral sequence shared by gorilla and human X chromosome alpha satellite. *Journal of molecular biology*, 216(3) :555–566. [1.3.2](#)
- Dutrillaux, B. and Couturier, J. (1986). Principes de l'analyse chromosomique appliquée à la phylogénie : l'exemple des Pongidae et des Hominidae. *Mammalia*, 50 :22–37. [1.4.3](#)
- Dutrillaux, B., Couturier, J., and Chauvier, G. (1979). Chromosomal evolution of 19 species of sub-species of Cercopithecinae. In *Annales de genetique*, volume 23, pages 133–143. [1.4.2](#), [1.4.3](#), [1.4.3](#), [2.2.2](#), [2.3.2](#), [3.6](#), [2.3.3](#), [2.3.4](#), [2.3.4](#), [2.5](#), [3.7](#)
- Dutrillaux, B., Couturier, J., Muleris, M., Lombard, M., and Chauvier, G. (1981). Chromosomal phylogeny of forty-two species or subspecies of cercopithecoids (Primates Catarrhini). In *Annales de genetique*, volume 25, pages 96–109. [1.4.2](#), [1.4.3](#), [1.4.3](#), [2.3.2](#), [2.3.3](#), [3.6](#), [2.3.4](#), [3.6](#)
- Dutrillaux, B. and Covic, M. (1974). Etude de facteurs influençant la dénaturation thermique ménagée des chromosomes. *Experimental cell research*, 85(1) :143–153. [1.4.3](#)
- Dutrillaux, B., Dutrillaux, A.-M., Lombard, M., Gautier, J.-P., Cooper, R., Moysan, F., and Lernould, J. (1988a). The karyotype of *Cercopithecus solatus* Harrison 1988, a new species belonging to *C. lhoesti*, and its phylogenetic relationships with other guenons. *Journal of Zoology*, 215(4) :611–617. [1.4.2](#), [4.15](#), [1.4.3](#), [2.2.2](#), [3.6](#), [1.4](#), [3.7](#)
- Dutrillaux, B. and Lejeune, J. (1971). Sur une nouvelle technique d'analyse du caryotype humain. *CR Acad. Sci.(Paris)*, 272 :2638–2640. [1.4.3](#)
- Dutrillaux, B., Muleris, M., and Couturier, J. (1988b). Chromosomal evolution of Cercopithecinae. *A Primate Radiation : Evolutionary Biology of the African Guenons*, pages 150–159. [4.8](#), [1.4.2](#)

- Dutrillaux, B., Rethoré, M., Aurias, A., and Goustand, M. (1975). Analyse du caryotype de deux espèces de Gibbons (*Hylobates lar* et *H. concolor*) par différentes techniques de marquage. *Cytogenet. Cell Genet.*, 15 :81–91. [1.4.3](#)
- Dutrillaux, B., Rethoré, M.-O., Prieur, M., and Lejeune, J. (1973). Analyse de la structure fine des chromosomes du Gorille (*Gorilla gorilla*). *Humangenetik*, 20(4) :343–354. [1.4.3](#)
- Dutrillaux, B., Viegas-Pequignot, E., Couturier, J., and Chauvier, G. (1978). Identity of eu-chromatic bands from man to Cercopithecidae. *Human genetics*, 45(3) :283–296. [1.4.2](#), [4.14](#), [3.6](#), [3.7](#)
- Dutrillaux, B., R. M. and Lejeune, J. (1975). Comparaison du caryotype de l'orang-outang (*Pongo pygmaeus*) à celui de l'homme, du chimpanzé et du gorille. *Annales de genetique*, 18(3) :153–161. [1.4.3](#)
- Edgar, R. C. (2004). MUSCLE : multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5) :1792–1797. [2.1.6](#), [2.2.6](#), [2.3.6](#), [1.2](#), [2.1](#)
- Eichler, E. E., Clark, R. A., and She, X. (2004). An assessment of the sequence gaps : unfinished business in a finished human genome. *Nature Reviews Genetics*, 5(5) :345–354. [1.3.3](#), [2.2.4](#), [2.3.2](#)
- Eickbush, T. H. and Jamburuthugoda, V. K. (2008). The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus research*, 134(1) :221–234. [1.1.1](#)
- Elder Jr, J. F. and Turner, B. J. (1995). Concerted evolution of repetitive DNA sequences in eukaryotes. *Quarterly Review of Biology*, pages 297–320. [1.3.3](#)
- Eldridge, M. D. and Close, R. L. (1993). Radiation of chromosome shuffles. *Current opinion in genetics & development*, 3(6) :915–922. [1.2.2](#)
- Enstam, K. L. and Isbell, L. A. (2002). Comparison of responses to alarm calls by patas (*Erythrocebus patas*) and vervet (*Cercopithecus aethiops*) monkeys in relation to habitat structure. *American Journal of Physical Anthropology*, 119(1) :3–14. [1.4.1](#)
- Enstam, K. L. and Isbell, L. A. (2007). The guenons (genus *Cercopithecus*) and their allies : behavioral ecology of polypatric associations. *Primates in perspective*, pages 252–274. [1.4.1](#), [1.4.1](#)
- Enstam, K. L., Isbell, L. A., and De Maar, T. W. (2002). Male demography, female mating behavior, and infanticide in wild patas monkeys (*Erythrocebus patas*). *International Journal of Primatology*, 23(1) :85–104. [1.4.1](#)
- Feliciello, I., Akrap, I., and Ugarković, D. (2015). Satellite DNA modulates gene expression in the beetle *Tribolium castaneum* after heat stress. *PLoS Genet*, 11(8) :e1005466. [1.1.3](#), [2.1.2](#)

- Feliciello, I., Picariello, O., and Chinali, G. (2006). Intra-specific variability and unusual organization of the repetitive units in a satellite DNA from *Rana dalmatina* : molecular evidence of a new mechanism of DNA repair acting on satellite DNA. *Gene*, 383 :81–92. [1.3.3](#)
- Ferguson-Smith, M., Yang, F., Rens, W., and O'Brien, P. (2004). The impact of chromosome sorting and painting on the comparative analysis of primate genomes. *Cytogenetic and genome research*, 108(1-3) :112–121. [1.4.3](#)
- Ferreri, G., Marzelli, M., Rens, W., and O'Neill, R. (2004). A centromere-specific retroviral element associated with breaks of synteny in macropodine marsupials. *Cytogenetic and genome research*, 107(1-2) :115–118. [1.2.2](#)
- Ferreri, G. C., Brown, J. D., Obergfell, C., Jue, N., Finn, C. E., O'Neill, M. J., and O'Neill, R. J. (2011). Recent amplification of the kangaroo endogenous retrovirus, KERV, limited to the centromere. *Journal of virology*, 85(10) :4761–4771. [1.2.2](#)
- Ferri, F., Bouzinba-Segard, H., Velasco, G., Hube, F., and Francastel, C. (2009). Non-coding murine centromeric transcripts associate with and potentiate Aurora B kinase. *Nucleic acids research*, page gkp529. [1.2.2](#), [2.2.2](#)
- Ferrigno, O., Virolle, T., Djabari, Z., Ortonne, J.-P., White, R. J., and Aberdam, D. (2001). Transposable B2 SINE elements can provide mobile RNA polymerase II promoters. *Nature genetics*, 28(1) :77–81. [1.1.3](#)
- Feschotte, C. and Mouches, C. (2000). Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a pogo-like DNA transposon. *Molecular Biology and Evolution*, 17(5) :730–737. [1.1.1](#)
- Feschotte, C. and Pritham, E. J. (2007). DNA transposons and the evolution of eukaryotic genomes. *Annual review of genetics*, 41 :331. [1.1.1](#)
- Finaz, C. and De Grouchy, J. (1971). Le caryotype humain après traitement par l'α-chymotrypsine. *Ann. Genet*, 14 :309–311. [1.4.3](#)
- Finaz, C., Turleau, C., J, D., VANCONG, N., R, R., and Frezal, J. (1973). Comparison of man and chimpanzee syntenic groups by cell hybridization-preliminary report. *BIOMEDICINE EXPRESS*, 19(12) :526–531. [1.4.3](#)
- Finelli, P., Stanyon, R., Plesker, R., Ferguson-Smith, M., O'brien, P., and Wienberg, J. (1999). Reciprocal chromosome painting shows that the great difference in diploid number between human and African green monkey is mostly due to non-robertsonian fissions. *Mammalian Genome*, 10(7) :713–718. [1.4.3](#)
- Finnegan, D. J. (1989). Eukaryotic transposable elements and genome evolution. *Trends in genetics*, 5 :103–107. [1.1.1](#)
- Finnegan, D. J. (2012). Retrotransposons. *Current Biology*, 22(11) :R432–R437. [1.1.1](#)

- Fittler, F. (1977). Analysis of the α -satellite dna from African green monkey cells by restriction nucleases. *European Journal of Biochemistry*, 74(2) :343–352. [2.1.2](#)
- Flemming, W. (1882). *Zellsubstanz, kern und zelltheilung*. Vogel. [1.2](#)
- Fleury, M.-C. and Gautier-Hion, A. (1997). Better to live with allogenerics than to live alone ? the case of single male Cercopithecus pogonias in troops of Colobus satanas. *International journal of primatology*, 18(6) :967–974. [1.4.1](#)
- Foltz, D. R., Jansen, L. E., Black, B. E., Bailey, A. O., Yates, J. R., and Cleveland, D. W. (2006). The human CENP-A centromeric nucleosome-associated complex. *Nature cell biology*, 8(5) :458–469. [1.2.1](#)
- Francesca, D. and Luca, S. (2010). Chromosomal dynamics in Cercopithecini studied by Williams-Beuren probe mapping. *Caryologia*, 63(4) :435–442. [1.4.3](#), [1.4.3](#)
- Fukagawa, T. and Earnshaw, W. C. (2014). The centromere : chromatin foundation for the kinetochore machinery. *Developmental cell*, 30(5) :496–508. [1.2.1](#), [2.9](#), [2.2.2](#)
- Funk, D. J. and Omland, K. E. (2003). Species-level paraphyly and polyphyly : frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology, Evolution, and Systematics*, pages 397–423. [1.4.2](#)
- Furuichi, T. (2006). Red-tailed monkeys (*Cercopithecus ascanius*) hunt green pigeons (*Treron calva*) in the Kalinzu Forest in Uganda. *Primates*, 47(2) :174–176. [1.4.1](#)
- Gaff, C., Du Sart, D., KalltsIs, P., Nagy, A., Choo, K., et al. (1994). A novel nuclear protein binds centromeric alpha satellite DNA. *Human molecular genetics*, 3(5) :711–716. [1.2.1](#)
- Garrick, D., Fiering, S., Martin, D. I., and Whitelaw, E. (1998). Repeat-induced gene silencing in mammals. *Nature genetics*, 18(1) :56–59. [1.1.3](#)
- Gautier, J.-P. (1975). *Étude comparée des systèmes d'intercommunication sonore chez quelques cercopithécines forestiers africains : mise en évidence de corrélations phylogénétiques et socio-écologiques*. PhD thesis. [1.4.1](#)
- Gautier, J.-P. (1985). Quelques caractéristiques écologiques du singe des marais : *Allenopithecus nigroviridis* Lang 1923. [1.4.1](#), [1.4.1](#)
- Gautier, J.-P. (1988). Interspecific affinities among guenons as deduced from vocalizations. *A primate radiation : Evolutionary biology of the African guenons*, pages 194–226. [2.3.4](#)
- Gautier-Hion, A. (1971). L’écologie du talapoin du Gabon. [1.4.1](#), [1.4.1](#)
- Gautier-Hion, A. (1973). Social and ecological features of talapoin monkey—comparisons with sympatric cercopithecines. *Comparative Ecology and Behavior of Primates*, pages 147–170. [1.4.1](#), [1.4.1](#)

- Gautier-Hion, A. (1988). Polyspecific associations among forest guenons : ecological, behavioural and evolutionary aspects. *A primate radiation : Evolutionary biology of the African guenons*, pages 452–476. [1.4.1](#)
- Gautier-Hion, A. and Gautier, J.-P. (1974). Les associations polyspécifiques de cercopithèques du plateau de M'passa (Gabon)(part 1 of 4). *Folia primatologica*, 22(2-3) :134–144. [1.4.1](#)
- Gautier-Hion, A. and Gautier, J.-P. (1978). Le singe de Brazza : une stratégie originale. *Zeitschrift für Tierpsychologie*, 46(1) :84–104. [1.4.1](#), [1.4.1](#)
- Gautier-Hion, A., Gautier, J.-P., and Quris, R. (1981). Forest structure and fruit availability as complementary factors influencing habitat use by a troop of monkeys (*Cercopithecus cebus*). [1.4.1](#)
- Gautier-Hion, A., Quris, R., and Gautier, J.-P. (1983). Monospecific vs polyspecific life : a comparative study of foraging and antipredatory tactics in a community of *Cercopithecus* monkeys. *Behavioral Ecology and Sociobiology*, 12(4) :325–335. [1.4.1](#)
- Gerasimova, T. I., Byrd, K., and Corces, V. G. (2000). A chromatin insulator determines the nuclear localization of DNA. *Molecular cell*, 6(5) :1025–1035. [1.1.3](#)
- Giannuzzi, G., Catacchio, C. R., and Ventura, M. (2012). *Centromere Evolution : Digging into Mammalian Primary Constriction*. INTECH Open Access Publisher. [1.2.2](#), [2.3](#)
- Glauser, D. A. (2014). The multiplicity of alternative splicing decisions in *Caenorhabditis elegans* is linked to specific intronic regulatory motifs and minisatellites. *BMC genomics*, 15(1) :1. [1.1.3](#), [2.2.2](#)
- Goldberg, I. G., Sawhney, H., Pluta, A. F., Warburton, P. E., and Earnshaw, W. C. (1996). Surprising deficiency of CENP-B binding sites in African green monkey alpha-satellite DNA : implications for CENP-B function at centromeres. *Molecular and cellular biology*, 16(9) :5156–5168. [1.3.1](#), [1.3.1](#), [2.1.4](#), [3.2.1](#)
- Gonzalez, C., Jimenez, J. C., Ripoll, P., and Sunkel, C. E. (1991). The spindle is required for the process of sister chromatid separation in *Drosophila* neuroblasts. *Experimental cell research*, 192(1) :10–15. [1.2.1](#)
- Gopalakrishnan, S., Sullivan, B. A., Trazzi, S., Della Valle, G., and Robertson, K. D. (2009). DNMT3B interacts with constitutive centromere protein CENP-C to modulate DNA methylation and the histone code at centromeric regions. *Human molecular genetics*, 18(17) :3178–3193. [1.2.1](#)
- Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView version 4 : a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular biology and evolution*, 27(2) :221–224. [2.1.6](#), [2.2.6](#)

- Gray, J., Carrano, A., Steinmetz, L., Van Dilla, M., Moore, D., Mayall, B., and Mendelsohn, M. (1975). Chromosome measurement and sorting by flow systems. *Proceedings of the National Academy of Sciences*, 72(4) :1231–1234. [1.4.3](#)
- Grouchy, J. d., Turleau, C., and Finaz, C. (1978). Chromosomal phylogeny of the primates. *Annual review of genetics*, 12(1) :289–328. [4.13](#)
- Grouchy, J. d., Turleau, C., Roubin, M., and Klein, M. (1972). Evolutions caryotypiques de l'homme et du chimpanzé. étude comparative des topographies de bandes après dénaturation ménagée. *Ann. Genet.*, 15 :79–84. [1.4.3](#)
- Groves, C. P. (2001). Primate taxonomy. [1.4.1](#), [2.3.3](#)
- Guenatri, M., Bailly, D., Maison, C., and Almouzni, G. (2004). Mouse centric and pericentric satellite repeats form distinct functional heterochromatin. *The Journal of cell biology*, 166(4) :493–505. [2.5](#), [2.1.4](#)
- Guschanski, K., Krause, J., Sawyer, S., Valente, L. M., Bailey, S., Finstermeier, K., Sabin, R., Gilissen, E., Sonet, G., Nagy, Z. T., et al. (2013). Next-generation museomics disentangles one of the largest primate radiations. *Systematic biology*, 62(4) :539–554. [1.4.2](#), [4.11](#), [4.12](#), [2.1.2](#), [2.2.2](#), [2.3.2](#), [2.3.4](#), [2.3.6](#), [3.2.2](#), [3.2.3](#)
- Guy, J., Hearn, T., Crosier, M., Mudge, J., Viggiano, L., Koczan, D., Thiesen, H.-J., Bailey, J. A., Horvath, J. E., Eichler, E. E., et al. (2003). Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10p. *Genome research*, 13(2) :159–172. [1.2.2](#)
- Haaf, T. and Willard, H. F. (1997). Chromosome-specific α -satellite DNA from the centromere of chimpanzee chromosome 4. *Chromosoma*, 106(4) :226–232. [1.3.2](#), [2.2.2](#), [2.3.2](#)
- Haaf, T. and Willard, H. F. (1998). Orangutan α -satellite monomers are closely related to the human consensus sequence. *Mammalian genome*, 9(6) :440–447. [1.3.2](#), [2.2.2](#)
- Hacch, F. and Mazrimas, J. (1974). Fractionation and characterization of satellite DNAs of the kangaroo rat (*Dipodomys ordii*). *Nucleic acids research*, 1(4) :559–576. [1.1.2](#)
- Haldane, J. B. (1922). Sex ratio and unisexual sterility in hybrid animals. *Journal of genetics*, 12(2) :101–109. [1.4.2](#)
- Hall, K. (1966). Behaviour and ecology of the wild patas monkey, *Erythrocebus patas*, in Uganda. *Journal of Zoology*, 148(1) :15–87. [1.4.1](#)
- Haltenorth, T. and Diller, H. (1988). *The Collins Field Guide to the Mammals of Africa : Including Madagascar*. S. Greene Press. [1.4.1](#)
- Han, J. S., Szak, S. T., and Boeke, J. D. (2004). Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature*, 429(6989) :268–274. [1.1.3](#)

- Harding, R. S. and Olson, D. K. (1986). Patterns of mating among male patas monkeys (*Erythrocebus patas*) in Kenya. *American Journal of Primatology*, 11(4) :343–358. [1.4.1](#)
- Harel, J., Hanania, N., Tapiero, H., and Harel, L. (1968). RNA replication by nuclear satellite DNA in different mouse cells. *Biochemical and biophysical research communications*, 33(4) :696–701. [1.2.2](#)
- Harrison, M. J. (1988). A new species of guenon (genus *Cercopithecus*) from Gabon. *Journal of Zoology*, 215(3) :561–575. [1.4.1](#), [2.1.2](#)
- Hasson, D., Panchenko, T., Salimian, K. J., Salman, M. U., Sekulic, N., Alonso, A., Warburton, P. E., and Black, B. E. (2013). The octamer is the major form of centromeric nucleosomes at human centromeres. *Nature structural & molecular biology*, 20(6) :687–695. [1.2.1](#)
- Hauser, M. D., Cheney, D. L., and Seyfarth, R. M. (1986). Group extinction and fusion in free-ranging vervet monkeys. *American Journal of Primatology*, 11(1) :63–77. [1.4.1](#)
- Hayashi, T., Ohtsuka, H., Kuwabara, K., Mafune, Y., Miyashita, N., Moriwaki, K., Takahashi, Y., and Kominami, R. (1993). A variant family of mouse minor satellite located on the centromeric region of chromosome 2. *Genomics*, 17(2) :490–492. [1.2.2](#)
- Hayden, K. E. (2012). Human centromere genomics : now it's personal. *Chromosome research*, 20(5) :621–633. [2.7](#), [1.2.2](#), [1.3.2](#), [3.5](#), [1.3.2](#), [2.1.2](#), [2.2.2](#)
- Hayden, K. E., Strome, E. D., Merrett, S. L., Lee, H.-R., Rudd, M. K., and Willard, H. F. (2013). Sequences associated with centromere competency in the human genome. *Molecular and cellular biology*, 33(4) :763–772. [1.2.2](#), [1.3.1](#)
- Hayman, D. (1990). Marsupial cytogenetics. *Australian Journal of Zoology*, 37(3) :331–349. [1.2.2](#)
- Henikoff, J. G., Thakur, J., Kasinathan, S., and Henikoff, S. (2015). A unique chromatin complex occupies young α -satellite arrays of human centromeres. *Science advances*, 1(1) :e1400234. [2.1.4](#)
- Henikoff, S. (2002). Near the edge of a chromosome's 'black hole'. *TRENDS in Genetics*, 18(4) :165–167. [2.1.4](#)
- Henikoff, S., Ahmad, K., and Malik, H. S. (2001). The centromere paradox : stable inheritance with rapidly evolving dna. *Science*, 293(5532) :1098–1102. [1.2.2](#)
- Hill, C. M. (1994). The role of female diana monkeys, *cercopithecus diana*, in territorial defence. *Animal behaviour*, 47(2) :425–431. [1.4.1](#)
- Hodgson, J. W., Argiropoulos, B., and Brock, H. W. (2001). Site-specific recognition of a 70-base-pair element containing d (GA) n repeats mediates bithoraxoid Polycomb group Response Element-dependent silencing. *Molecular and cellular biology*, 21(14) :4528–4543. [1.1.3](#)

- Holoch, D. and Moazed, D. (2015). RNA-mediated epigenetic regulation of gene expression. *Nature Reviews Genetics*, 16(2) :71–84. [1.2](#)
- Hori, T., Amano, M., Suzuki, A., Backer, C. B., Welburn, J. P., Dong, Y., McEwen, B. F., Shang, W.-H., Suzuki, E., Okawa, K., et al. (2008). CCAN makes multiple contacts with centromeric DNA to provide distinct pathways to the outer kinetochore. *Cell*, 135(6) :1039–1052. [1.2.1](#)
- Hori, T., Shang, W.-H., Toyoda, A., Misu, S., Monma, N., Ikeo, K., Molina, O., Vargiu, G., Fujiyama, A., Kimura, H., et al. (2014). Histone H4 lys 20 monomethylation of the CENP-A nucleosome is essential for kinetochore assembly. *Developmental cell*, 29(6) :740–749. [1.2.1](#)
- Houwing, S., Kamminga, L. M., Berezikov, E., Cronembold, D., Girard, A., Van Den Elst, H., Filippov, D. V., Blaser, H., Raz, E., Moens, C. B., et al. (2007). A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell*, 129(1) :69–82. [1.1.3](#)
- Hua-Van, A., Le Rouzic, A., Boutin, T. S., Filée, J., and Capy, P. (2011). The struggle for life of the genome's selfish architects. *Biology direct*, 6(1) :1. [1.1.1](#)
- Hudson, D. F., Fowler, K. J., Earle, E., Saffery, R., Kalitsis, P., Trowell, H., Hill, J., Wreford, N. G., de Kretser, D. M., Cancilla, M. R., et al. (1998). Centromere protein B null mice are mitotically and meiotically normal but have lower body and testis weights. *The Journal of cell biology*, 141(2) :309–319. [1.2.1](#)
- Huelsenbeck, J. P., Ronquist, F., et al. (2001). MRBAYES : Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8) :754–755. [2.3.6](#)
- Isbell, L. A. (1998). Diet for a small primate : Insectivory and gummivory in the (large) patas monkey(*Erythrocebus patas pyrrhonotus*). *American Journal of Primatology*, 45(4) :381–398. [1.4.1](#)
- Isbell, L. A., Cheney, D. L., and Seyfarth, R. M. (1991). Group fusions and minimum group sizes in vervet monkeys (*Cercopithecus aethiops*). *American Journal of Primatology*, 25(1) :57–65. [1.4.1](#)
- Isbell, L. A. and Pruetz, J. D. (1998). Differences between vervets (*Cercopithecus aethiops*) and patas monkeys (*Erythrocebus patas*) in agonistic interactions between adult females. *International Journal of Primatology*, 19(5) :837–855. [1.4.1](#)
- Jabs, E. W., Wolf, S. F., and Migeon, B. R. (1984). Characterization of a cloned DNA sequence that is present at centromeres of all human autosomes and the X chromosome and shows polymorphic variation. *Proceedings of the National Academy of Sciences*, 81(15) :4884–4888. [1.3.2](#)
- Jarne, P. and Lagoda, P. J. (1996). Microsatellites, from molecules to populations and back. *Trends in ecology & evolution*, 11(10) :424–429. [1.1.2](#)

- Jauch, A., Wienberg, J., Stanyon, R., Arnold, N., Tofanelli, S., Ishida, T., and Cremer, T. (1992). Reconstruction of genomic rearrangements in great apes and gibbons by chromosome painting. *Proceedings of the National Academy of Sciences*, 89(18) :8611–8615. [1.4.3](#)
- Jenuwein, T. (2001). Re-SET-ting heterochromatin by histone methyltransferases. *Trends in cell biology*, 11(6) :266–273. [1.2.1](#)
- Jjingo, D., Conley, A. B., Wang, J., Mariño-Ramírez, L., Lunyak, V. V., and Jordan, I. K. (2014). Mammalian-wide interspersed repeat (MIR)-derived enhancers and the regulation of human gene expression. *Mobile DNA*, 5(1) :1. [1.1.3](#)
- Joglekar, A. P., Bouck, D., Finley, K., Liu, X., Wan, Y., Berman, J., He, X., Salmon, E. D., and Bloom, K. S. (2008). Molecular architecture of the kinetochore-microtubule attachment site is conserved between point and regional centromeres. *The Journal of cell biology*, 181(4) :587–594. [1.2.1](#)
- Johnston, K., Joglekar, A., Hori, T., Suzuki, A., Fukagawa, T., and Salmon, E. D. (2010). Vertebrate kinetochore protein architecture : protein copy number. *The Journal of cell biology*, 189(6) :937–943. [1.2.1](#)
- Jonstrup, A. T., Thomsen, T., Wang, Y., Knudsen, B. R., Koch, J., and Andersen, A. H. (2008). Hairpin structures formed by alpha satellite DNA of human centromeres are cleaved by human topoisomerase II α . *Nucleic acids research*, 36(19) :6165–6174. [1.2.2](#)
- Jordan, I. K., Rogozin, I. B., Glazko, G. V., and Koonin, E. V. (2003). Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends in Genetics*, 19(2) :68–72. [1.1.3](#)
- Jørgensen, A., Jones, C., Bostock, C., and Bak, A. L. (1987). Different subfamilies of alphoid repetitive DNA are present on the human and chimpanzee homologous chromosomes 21 and 22. *The EMBO journal*, 6(6) :1691. [1.3.2](#), [2.1.2](#)
- Jørgensen, A., Laursen, H. B., Jones, C., and Bak, A. L. (1992). Evolutionarily different alphoid repeat DNA on homologous chromosomes in human and chimpanzee. *Proceedings of the National Academy of Sciences*, 89(8) :3310–3314. [1.3.2](#)
- Kapitonov, V. V. and Jurka, J. (2001). Rolling-circle transposons in eukaryotes. *Proceedings of the National Academy of Sciences*, 98(15) :8714–8719. [1.1.1](#)
- Kapoor, M., de Oca Luna, R. M., Liu, G., Lozano, G., Cummings, C., Mancini, M., Ouspenski, I., Brinkley, B., and May, G. S. (1998). The cenpB gene is not essential in mice. *Chromosoma*, 107(8) :570–576. [1.2.1](#)
- Kazakov, A. E., Shepelev, V. A., Tumeneva, I. G., Alexandrov, A. A., Yurov, Y. B., and Alexandrov, I. A. (2003). Interspersed repeats are found predominantly in the “old” α satellite families. *Genomics*, 82(6) :619–627. [1.2.2](#)

- Kiburz, B. M., Reynolds, D. B., Megee, P. C., Marston, A. L., Lee, B. H., Lee, T. I., Levine, S. S., Young, R. A., and Amon, A. (2005). The core centromere and Sgo1 establish a 50-kb cohesin-protected domain around centromeres during meiosis I. *Genes & development*, 19(24) :3017–3030. [1.2](#)
- Kim, J.-H., Ebersole, T., Kouprina, N., Noskov, V. N., Ohzeki, J.-I., Masumoto, H., Mravinac, B., Sullivan, B. A., Pavlicek, A., Dovat, S., et al. (2009). Human gamma-satellite DNA maintains open chromatin structure and protects a transgene from epigenetic silencing. *Genome research*, 19(4) :533–544. [1.2.2](#)
- Kingdon, J. (1980). The role of visual signals and face patterns in African forest monkeys (guenons) of the genus *Cercopithecus*. *The Transactions of the Zoological Society of London*, 35(4) :425–475. [1.4.1](#)
- Kingdon, J. (2007). Primate visual signals in noisy environments. *Folia Primatologica*, 78(5–6) :389–404. [1.4.1](#)
- Kingdon, J., Happold, D., Butynski, T., Hoffmann, M., Happold, M., and Kalina, J. (2013). *Mammals of Africa*, volume 1. A&C Black. [1.4.2](#)
- Kipling, D., Ackford, H. E., Taylor, B. A., and Cooke, H. J. (1991). Mouse minor satellite DNA genetically maps to the centromere and is physically linked to the proximal telomere. *Genomics*, 11(2) :235–241. [1.2.2](#)
- Kipling, D., Wilson, H. E., Mitchell, A. R., Taylor, B. A., and Cooke, H. J. (1994). Mouse centromere mapping using oligonucleotide probes that detect variants of the minor satellite. *Chromosoma*, 103(1) :46–55. [1.2.2](#)
- Kit, S. (1961). Equilibrium sedimentation in density gradients of DNA preparations from animal tissues. *Journal of molecular biology*, 3(6) :711IN1–716IN2. [1.1.2](#)
- Koehler, U., Bigoni, F., Wienberg, J., and Stanyon, R. (1995). Genomic reorganization in the concolor gibbon (*Hylobates concolor*) revealed by chromosome painting. *Genomics*, 30(2) :287–292. [1.4.3](#)
- Koga, A., Hirai, Y., Hara, T., and Hirai, H. (2012). Repetitive sequences originating from the centromere constitute large-scale heterochromatin in the telomere region in the siamang, a small ape. *Heredity*, 109(3) :180–187. [1.3.2](#)
- Koga, A., Hirai, Y., Terada, S., Jahan, I., Baicharoen, S., Arsaithamkul, V., and Hirai, H. (2014). Evolutionary origin of higher-order repeat structure in alpha-satellite DNA of primate centromeres. *DNA Research*, page dsu005. [1.3.1](#), [2.2.2](#), [2.3.2](#)
- Koga, A., Notohara, M., and Hirai, H. (2011). Evolution of subterminal satellite (StSat) repeats in hominids. *Genetica*, 139(2) :167–175. [1.3.2](#)

- Komissarov, A. S., Gavrilova, E. V., Demin, S. J., Ishov, A. M., and Podgornaya, O. I. (2011). Tandemly repeated DNA families in the mouse genome. *BMC genomics*, 12(1) :1. [1.2.2](#), [2.6](#)
- Kopecka, H., Macaya, G., Cortadas, J., Thiéry, J.-P., and Bernardi, G. (1978). Restriction enzyme analysis of satellite DNA components from the bovine genome. *European Journal of Biochemistry*, 84(1) :189–195. [1.2.2](#)
- Krüger, J. and Vogel, F. (1975). Population genetics of unequal crossing over. *Journal of Molecular Evolution*, 4(3) :201–247. [1.3.3](#)
- Kurnit, D. M. and Maio, J. J. (1974). Variable satellite DNA's in the African green monkey *Cercopithecus aethiops*. *Chromosoma*, 45(4) :387–400. [1.3](#)
- Kurnit, D. M., Shafit, B. R., and Maio, J. J. (1973). Multiple satellite deoxyribonucleic acids in the calf and their relation to the sex chromosomes. *Journal of molecular biology*, 81(3) :273–284. [1.2.2](#)
- Kuznetsova, I. S., Prusov, A. N., Enukashvily, N. I., and Podgornaya, O. I. (2005). New types of mouse centromeric satellite DNAs. *Chromosome Research*, 13(1) :9–25. [1.2.2](#)
- Lanfear, R., Calcott, B., Ho, S. Y., and Guindon, S. (2012). PartitionFinder : combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular biology and evolution*, 29(6) :1695–1701. [2.3.6](#)
- Lawes, M., Cords, M., and Lehn, C. (2013). *Cercopithecus mitis* profile. *Mammals of Africa*, 2 :354–362. [1.4.1](#)
- Lawes, M., Henzi, S., and Perrin, M. (1990). Diet and feeding behaviour of samango monkeys (*Cercopithecus mitis labiatus*) in Ngoye forest, South Africa. *Folia Primatologica*, 54(1-2) :57–69. [1.4.1](#)
- Lee, C., Wevrick, R., Fisher, R., Ferguson-Smith, M., and Lin, C. (1997). Human centromeric DNAs. *Human genetics*, 100(3-4) :291–304. [1.2.2](#), [2.1.2](#), [2.3.2](#)
- Lee, H.-R., Hayden, K. E., and Willard, H. F. (2011). Organization and molecular evolution of CENP-A-associated satellite DNA families in a basal primate genome. *Genome biology and evolution*, 3 :1136–1149. [1.3](#)
- Lee, T. N. and Singer, M. F. (1982). Structural organization of α -satellite DNA in a single monkey chromosome. *Journal of molecular biology*, 161(2) :323–342. [2.1.3](#)
- Lejeune, J., Dutrillaux, B., Rethoré, M. O., and Prieur, M. (1973). Comparaison de la structure fine des chromatides d'*Homo sapiens* et de *Pan troglodytes*. *Chromosoma*, 43(4) :423–444. [1.4.3](#)
- Lichter, P., Cremer, T., Borden, J., Manuelidis, L., and Ward, D. (1988). Delineation of individual human chromosomes in metaphase and interphase cells by in situ suppression hybridization using recombinant DNA libraries. *Human genetics*, 80(3) :224–234. [1.4.3](#)

- Lin, M., Comings, D., and Alfi, O. (1977). Optical studies of the interaction of 4-6-diamidino-2-phenylindole with DNA and metaphase chromosomes. *Chromosoma*, 60(1) :15–25. [1.4.3](#)
- Litt, M. and Luty, J. A. (1989). A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *American journal of human genetics*, 44(3) :397. [1.1.2](#)
- Locke, D. P., Hillier, L. W., Warren, W. C., Worley, K. C., Nazareth, L. V., Muzny, D. M., Yang, S.-P., Wang, Z., Chinwalla, A. T., Minx, P., et al. (2011). Comparative and demographic analysis of orang-outan genomes. *Nature*, 469(7331) :529–533. [1.2.2](#), [2.3.2](#)
- Lunyak, V. V., Prefontaine, G. G., Núñez, E., Cramer, T., Ju, B.-G., Ohgi, K. A., Hutt, K., Roy, R., García-Díaz, A., Zhu, X., et al. (2007). Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science*, 317(5835) :248–251. [1.1.3](#), [2.2.2](#)
- Lynch, V. J., Leclerc, R. D., May, G., and Wagner, G. P. (2011). Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nature genetics*, 43(11) :1154–1159. [1.1.3](#)
- Macas, J., Meszaros, T., and Nouzova, M. (2002). PlantSat : a specialized database for plant satellite repeats. *Bioinformatics*, 18(1) :28–35. [1.1.2](#)
- Madhani, H. D., Leadon, S. A., Smith, C. A., and Hanawalt, P. (1986). alpha DNA in African green monkey cells is organized into extremely long tandem arrays. *Journal of Biological Chemistry*, 261(5) :2314–2318. [2.1.2](#)
- Maio, J. J. (1971). DNA strand reassociation and polyribonucleotide binding in the African green monkey, *Cercopithecus aethiops*. *Journal of molecular biology*, 56(3) :579–595. [1.3](#), [2.1.2](#)
- Maio, J. J., Brown, F. L., and Musich, P. R. (1981). Toward a molecular paleontology of primate genomes. *Chromosoma*, 83(1) :103–125. [1.3](#), [2.1.2](#)
- Malik, H. S. and Henikoff, S. (2002). Conflict begets complexity : the evolution of centromeres. *Current opinion in genetics & development*, 12(6) :711–718. [1.2.2](#), [2.2](#), [1.2.2](#), [2.1.2](#)
- Maloney, K. A., Sullivan, L. L., Matheny, J. E., Strome, E. D., Merrett, S. L., Ferris, A., and Sullivan, B. A. (2012). Functional epialleles at an endogenous human centromere. *Proceedings of the National Academy of Sciences*, 109(34) :13704–13709. [1.3.2](#)
- Manuelidis, L. (1978). Chromosomal localization of complex and simple repeated human DNAs. *Chromosoma*, 66(1) :23–32. [1.2.2](#)
- Marshall, O. J., Chueh, A. C., Wong, L. H., and Choo, K. A. (2008). Neocentromeres : new insights into centromere structure, disease development, and karyotype evolution. *The American Journal of Human Genetics*, 82(2) :261–282. [1.2.2](#)

- Masumoto, H., Masukata, H., Muro, Y., Nozaki, N., and Okazaki, T. (1989). A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite. *The Journal of Cell Biology*, 109(5) :1963–1973. [1.2.1](#)
- McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, 36(6) :344–355. [1.1.1](#)
- Mestrović, N., Plohl, M., Mravinac, B., and Ugarković, D. (1998). Evolution of satellite DNAs from the genus Palorus-experimental evidence for the "library" hypothesis. *Molecular biology and evolution*, 15(8) :1062–1068. [1.1.2](#)
- Miga, K. H. (2015). Completing the human genome : the progress and challenge of satellite DNA assembly. *Chromosome Research*, 23(3) :421–426. [2.1.2](#)
- Miga, K. H., Newton, Y., Jain, M., Altemose, N., Willard, H. F., and Kent, W. J. (2014). Centromere reference models for human chromosomes X and Y satellite arrays. *Genome research*, 24(4) :697–707. [2.8](#), [1.3.1](#), [2.1.2](#), [2.3.4](#)
- Mirkin, S. M. (2007). Expandable DNA repeats and human disease. *Nature*, 447(7147) :932–940. [1.3.3](#)
- Montefalcone, G., Tempesta, S., Rocchi, M., and Archidiacono, N. (1999). Centromere repositioning. *Genome Research*, 9(12) :1184–1188. [1.2.2](#), [1.4.3](#), [2.3.2](#)
- Morescalchi, M., Schempp, W., Consigliere, S., Bigoni, F., Wienberg, J., and Stanyon, R. (1997). Mapping chromosomal homology between humans and the black-handed spider monkey by fluorescence in situ hybridization. *Chromosome Research*, 5(8) :527–536. [1.4.3](#)
- Motsch, P., Le Flohic, G., Dilger, C., Delahaye, A., Chateau-Smith, C., and Couette, S. (2015). Degree of terrestrial activity of the elusive sun-tailed monkey (*Cercopithecus solatus*) in Gabon : Comparative study of behavior and postcranial morphometric data. *American journal of primatology*, 77(10) :1060–1074. [1.4.2](#)
- Moulin, S., Gerbault-Seureau, M., Dutrillaux, B., and Richard, F. A. (2008). Phylogenomics of African guenons. *Chromosome Research*, 16(5) :783–799. [1.4.2](#), [4.9](#), [1.4.3](#), [2.1.6](#), [2.2.6](#), [2.3.2](#), [2.3.3](#), [3.6](#), [2.3.3](#), [2.3.4](#), [2.3.4](#), [1.4](#), [2.5](#), [3.6](#), [3.7](#)
- Mravinac, B. and Plohl, M. (2010). Parallelism in evolution of highly repetitive DNAs in sibling species. *Molecular biology and evolution*, 27(8) :1857–1867. [2.1.4](#)
- Muleris, M., Gautier, J., Lombard, M., and Dutrillaux, B. (1984). [cytogenetic study of *Cercopithecus wolfi*, *Cercopithecus erythrotis* and a hybrid *Cercopithecus ascanius* x *Cercopithecus pogonias grayi*]. In *Annales de genetique*, volume 28, pages 75–80. [2.2.2](#), [2.5](#)
- Müller, S., Hollatz, M., and Wienberg, J. (2003). Chromosomal phylogeny and evolution of gibbons (Hylobatidae). *Human genetics*, 113(6) :493–501. [1.4.3](#)

- Müller, S., O'Brien, P., Ferguson-Smith, M., and Wienberg, J. (1998). Cross-species colour segmenting : a novel tool in human karyotype analysis. *Cytometry*, 33(4) :445–452. [1.4.3](#)
- Müller, S., O'Brien, P. C., Ferguson-Smith, M. A., and Wienberg, J. (1997). A novel source of highly specific chromosome painting probes for human karyotype analysis derived from primate homologues. *Human genetics*, 101(2) :149–153. [1.4.3](#)
- Müller, S., Stanyon, R., O'brien, P., Ferguson-Smith, M., Plesker, R., and Wienberg, J. (1999). Defining the ancestral karyotype of all primates by multidirectional chromosome painting between tree shrews, lemurs and humans. *Chromosoma*, 108(6) :393–400. [1.4.3](#)
- Musich, P. R., Brown, F. L., and Maio, J. J. (1980). Highly repetitive component α and related alphoid DNAs in man and monkeys. *Chromosoma*, 80(3) :331–348. [1.3](#), [3.3](#), [2.1.2](#), [2.1.3](#)
- Nakagawa, N. (1989). Activity budget and diet of patas monkeys in Kala Maloue National Park, Cameroon : a preliminary report. *Primates*, 30(1) :27–34. [1.4.1](#)
- Nakagawa, N. (1999). Differential habitat utilization by patas monkeys (*Erythrocebus patas*) and tantalus monkeys (*Cercopithecus aethiops tantalus*) living sympatrically in northern Cameroon. *American Journal of Primatology*, 49(3) :243–264. [1.4.1](#)
- Navarro, A. and Barton, N. H. (2003a). Accumulating postzygotic isolation genes in parapatry : a new twist on chromosomal speciation. *Evolution*, 57(3) :447–459. [1.4.3](#)
- Navarro, A. and Barton, N. H. (2003b). Chromosomal speciation and molecular divergence-accelerated evolution in rearranged chromosomes. *Science*, 300(5617) :321–324. [1.4.3](#)
- Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3) :443–453. [2.1.6](#), [2.2.6](#)
- Nijman, I. J. and Lenstra, J. A. (2001). Mutation and recombination in cattle satellite DNA : a feedback model for the evolution of satellite DNA repeats. *Journal of Molecular Evolution*, 52(4) :361–371. [1.1.2](#)
- Nishino, T., Takeuchi, K., Gascoigne, K. E., Suzuki, A., Hori, T., Oyama, T., Morikawa, K., Cheeseman, I. M., and Fukagawa, T. (2012). CENP-TWSX forms a unique centromeric chromatin structure with a histone-like fold. *Cell*, 148(3) :487–501. [1.2.1](#)
- Nonaka, N., Kitajima, T., Yokobayashi, S., Xiao, G., Yamamoto, M., Grewal, S. I., and Watanabe, Y. (2002). Recruitment of cohesin to heterochromatic regions by Swi6/HP1 in fission yeast. *Nature Cell Biology*, 4(1) :89–93. [1.2](#)
- Norris, J., Fan, D., Aleman, C., Marks, J. R., Futreal, P. A., Wiseman, R. W., Iglehart, J. D., Deininger, P. L., and McDonnell, D. P. (1995). Identification of a new subclass of Alu DNA repeats which can function as estrogen receptor-dependent transcriptional enhancers. *Journal of Biological Chemistry*, 270(39) :22777–22782. [1.1.3](#)

- Ohta, T. (1983). On the evolution of multigene families. *Theoretical population biology*, 23(2) :216–240. [1.3.3](#)
- Ohta, T. and Dover, G. A. (1983). Population genetics of multigene families that are dispersed into two or more chromosomes. *Proceedings of the National Academy of Sciences*, 80(13) :4079–4083. [1.3.3](#)
- Ohta, T. and Dover, G. A. (1984). The cohesive population genetics of molecular drive. *Genetics*, 108(2) :501–521. [1.3.3](#)
- Ohta, T. and Kimura, M. (1981). Some calculations on the amount of selfish DNA. *Proceedings of the National Academy of Sciences*, 78(2) :1129–1132. [1.3.3](#)
- Ohzeki, J.-i., Nakano, M., Okada, T., and Masumoto, H. (2002). CENP-B box is required for de novo centromere chromatin assembly on human alphoid DNA. *The Journal of cell biology*, 159(5) :765–775. [1.2.1](#)
- Okada, M., Cheeseman, I. M., Hori, T., Okawa, K., McLeod, I. X., Yates, J. R., Desai, A., and Fukagawa, T. (2006). The CENP-H-I complex is required for the efficient incorporation of newly synthesized CENP-A into centromeres. *Nature cell biology*, 8(5) :446–457. [1.2.1](#)
- O'Keefe, C. L. and Matera, A. G. (2000). Alpha satellite DNA variant-specific oligoprobes differing by a single base can distinguish chromosome 15 homologs. *Genome research*, 10(9) :1342–1350. [2.1.3](#)
- Ollion, J., Loll, F., Cochenne, J., Boudier, T., and Escudé, C. (2015). Proliferation-dependent positioning of individual centromeres in the interphase nucleus of human lymphoblastoid cell lines. *Molecular biology of the cell*, 26(13) :2550–2560. [1.1.3](#), [2.1.3](#), [2.1.6](#), [2.2.6](#), [3.1.2](#)
- O'Neill, R., Eldridge, M., and Metcalfe, C. (2004). Centromere dynamics and chromosome evolution in marsupials. *Journal of Heredity*, 95(5) :375–381. [1.2.2](#)
- Page, S. and Goodman, M. (2001). Catarrhine phylogeny : noncoding DNA evidence for a di-phyletic origin of the mangabeys and for a human-chimpanzee clade. *Molecular Phylogenetics and Evolution*, 18(1) :14–25. [2.3.4](#)
- Pardue, M. L. and Gall, J. G. (1970). Chromosomal localization of mouse satellite DNA. *Science*, 168(3937) :1356–1358. [1.2.2](#)
- Pearson, P., Bobrow, M., Vosa, C., and Barlow, P. (1971). Quinacrine fluorescence in mammalian chromosomes. [1.4.3](#)
- Pech, M., Streeck, R. E., and Zachau, H. G. (1979). Patchwork structure of a bovine satellite DNA. *Cell*, 18(3) :883–893. [1.2.2](#)
- Perelman, P., Johnson, W. E., Roos, C., Seuánez, H. N., Horvath, J. E., Moreira, M. A., Kessing, B., Pontius, J., Roelke, M., Rumpler, Y., et al. (2011). A molecular phylogeny of living primates. *PLoS Genet*, 7(3) :e1001342. [3.1](#), [1.4.2](#), [2.3.2](#), [2.3.4](#), [2.3.4](#), [2.3.6](#), [3.2.2](#), [3.3](#)

- Perelson, A. S. and Bell, G. I. (1977). Mathematical models for the evolution of multigene families by unequal crossing over. *Nature*, 265(5592) :304. [1.3.3](#)
- Perez-Castro, A. V., Shamanski, F. L., Meneses, J. J., Lovato, T. L., Vogel, K. G., Moyzis, R. K., and Pedersen, R. (1998). Centromeric protein B null mice are viable with no apparent abnormalities. *Developmental biology*, 201(2) :135–143. [1.2.1](#)
- Pertile, M. D., Graham, A. N., Choo, K. A., and Kalitsis, P. (2009). Rapid evolution of mouse Y centromere repeat DNA belies recent sequence stability. *Genome research*, 19(12) :2202–2213. [1.2.2](#), [2.1.4](#)
- Peters, A. H., O'Carroll, D., Scherthan, H., Mechtler, K., Sauer, S., Schöfer, C., Weipoltshammer, K., Pagani, M., Lachner, M., Kohlmaier, A., et al. (2001). Loss of the Suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability. *Cell*, 107(3) :323–337. [1.2.1](#)
- Petitpierre, E., Juan, C., Pons, J., Plohl, M., and Ugarkovic, D. (1995). Satellite DNA and constitutive heterochromatin in tenebrionid beetles. In *Kew Chromosome Conference IV. Royal Botanic Gardens, London*, pages 351–362. [1.1.2](#)
- Piazza, A., Serero, A., Boule, J.-B., Legoix-Ne, P., Lopes, J., and Nicolas, A. (2012). Stimulation of gross chromosomal rearrangements by the human CEB1 and CEB25 minisatellites in *Saccharomyces cerevisiae* depends on G-quadruplexes or Cdc13. *PLoS Genet*, 8(11) :e1003033. [1.1.2](#)
- Picard, G., Bregliano, J., Bucheton, A., Lavige, J., Pelisson, A., and Kidwell, M. (1978). Non-mendelian female sterility and hybrid dysgenesis in *drosophila melanogaster*. *Genetical research*, 32(03) :275–287. [1.1.1](#)
- Pike, L. M., Carlisle, A., Newell, C., Hong, S.-B., and Musich, P. R. (1986). Sequence and evolution of rhesus monkey alphoid DNA. *Journal of molecular evolution*, 23(2) :127–137. [1.3.1](#), [1.3.1](#)
- Pinkel, D., Landegent, J., Collins, C., Fuscoe, J., Segraves, R., Lucas, J., and Gray, J. (1988). Fluorescence in situ hybridization with human chromosome-specific libraries : detection of trisomy 21 and translocations of chromosome 4. *Proceedings of the National Academy of Sciences*, 85(23) :9138–9142. [1.4.3](#)
- Piras, F. M., Nergadze, S. G., Magnani, E., Bertoni, L., Attolini, C., Khoriauli, L., Raimondi, E., and Giulotto, E. (2010). Uncoupling of satellite DNA and centromeric function in the genus *Equus*. *PLoS Genet*, 6(2) :e1000845. [1.2.2](#), [2.3.2](#)
- Plohl, M., Luchetti, A., Meštrović, N., and Mantovani, B. (2008). Satellite DNAs between selfishness and functionality : structure, genomics and evolution of tandem repeats in centromeric (hetero) chromatin. *Gene*, 409(1) :72–82. [2.1.2](#), [2.2.4](#)

- Plohl, M., Meštrović, N., and Mravinac, B. (2012). Satellite DNA evolution. In *Repetitive DNA*, volume 7, pages 126–152. Karger Publishers. [2.1.2](#), [2.2.2](#), [2.3.2](#)
- Popescu, P., Hayes, H., Dutrillaux, B., et al. (1998). Animal cytogenetic techniques. *Techniques de cytogénétique animale*. [4.1](#), [1.4.3](#), [1.4.3](#)
- Prakhongcheep, O., Hirai, Y., Hara, T., Srikulnath, K., Hirai, H., and Koga, A. (2013). Two types of alpha satellite DNA in distinct chromosomal locations in Azara's owl monkey. *DNA research*, page dst004. [3.1](#), [1.3.1](#), [1.3.2](#), [2.2.2](#)
- Prassolov, V., Kuchino, Y., Nemoto, K., and Nishimura, S. (1986). Nucleotide sequence of the Bamhi repetitive sequence, including the hindiii fundamental unit, as a possible mobile element from the Japanese monkey Macaca fuscata. *Journal of molecular evolution*, 23(3) :200–204. [1.3.1](#)
- Pritham, E. J. and Feschotte, C. (2007). Massive amplification of rolling-circle transposons in the lineage of the bat Myotis lucifugus. *Proceedings of the National Academy of Sciences*, 104(6) :1895–1900. [1.1.1](#)
- Pruetz, J. D. and Isbell, L. A. (2000). Correlations of food distribution and patch size with agonistic interactions in female vervets (Chlorocebus aethiops) and patas monkeys (Erythrocebus patas) living in simple habitats. *Behavioral Ecology and Sociobiology*, 49(1) :38–47. [1.4.1](#)
- Purgato, S., Belloni, E., Piras, F. M., Zoli, M., Badiale, C., Cerutti, F., Mazzagatti, A., Perini, G., Della Valle, G., Nergadze, S. G., et al. (2015). Centromere sliding on a mammalian chromosome. *Chromosoma*, 124(2) :277–287. [1.2.2](#)
- Quris, R. (1976). Données comparatives sur la socio-écologie de huit espèces de Cercopithecidae vivant dans une même zone de forêt primitive périodiquement inondée (Nord-Est du Gabon). [1.4.1](#)
- R Core Team (2013). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. [2.1.6](#), [2.2.6](#)
- Rens, W., O'Brien, P., Fairclough, H., Harman, L., Graves, J., and Ferguson-Smith, M. (2004). Reversal and convergence in marsupial chromosome evolution. *Cytogenetic and genome research*, 102(1-4) :282–290. [1.2.2](#)
- Ribeiro, S. A., Vagnarelli, P., Dong, Y., Hori, T., McEwen, B. F., Fukagawa, T., Flors, C., and Earnshaw, W. C. (2010). A super-resolution map of the vertebrate kinetochore. *Proceedings of the National Academy of Sciences*, 107(23) :10484–10489. [1.2.1](#)
- Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS : the European molecular biology open software suite. *Trends in genetics*, 16(6) :276–277. [2.1.6](#), [2.2.6](#)

- Richard, F., Lombard, M., and Dutrillaux, B. (1996). ZOO-FISH suggests a complete homology between human and capuchin monkey (Platyrrhini) euchromatin. *Genomics*, 36(3) :417–423. [1.4.3](#)
- Richard, G.-F., Kerrest, A., and Dujon, B. (2008). Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiology and Molecular Biology Reviews*, 72(4) :686–727. [1.1](#), [1.1.2](#), [1.1.2](#)
- Rieseberg, L. H. (2001). Chromosomal rearrangements and speciation. *Trends in Ecology & Evolution*, 16(7) :351–358. [1.4.3](#)
- Rieseberg, L. H., Whitton, J., and Gardner, K. (1999). Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics*, 152(2) :713–727. [1.4.3](#)
- Roberto, R., Capozzi, O., Wilson, R. K., Mardis, E. R., Lomiento, M., Tuzun, E., Cheng, Z., Mootnick, A. R., Archidiacono, N., Rocchi, M., et al. (2007). Molecular refinement of gibbon genome rearrangements. *Genome research*, 17(2) :249–257. [1.4.3](#)
- Rocchi, M., Archidiacono, N., Antonacci, R., Finelli, P., D'Aiuto, L., Carbone, R., Lindsay, E., and Baldini, A. (1994). Cloning and comparative mapping of recently evolved human chromosome 22-specific alpha satellite DNA. *Somatic cell and molecular genetics*, 20(5) :443–448. [1.3.2](#)
- Rocchi, M., Archidiacono, N., Schempp, W., Capozzi, O., and Stanyon, R. (2012). Centromere repositioning in mammals. *Heredity*, 108(1) :59–67. [2.3.2](#), [2.3.4](#)
- Rodrigues, A., Douzery, E., et al. (2009). Patterns of macroevolution among Primates inferred from a supermatrix of mitochondrial and nuclear DNA. *Molecular phylogenetics and evolution*, 53(3) :808–825. [2.3.4](#)
- Rofe, R. (1978). G-banded chromosomes and the evolution of Macropodidae. *Aust Mammal*, 2 :53–63. [1.2.2](#)
- Roizès, G. (2006). Human centromeric alphoid domains are periodically homogenized so that they vary substantially between homologues. Mechanism and implications for centromere functioning. *Nucleic acids research*, 34(6) :1912–1924. [1.3.3](#), [2.1.2](#)
- Rojo, V., Martínez-Lage, A., Giovannotti, M., González-Tizón, A. M., Cerioni, P. N., Barucchi, V. C., Galán, P., Olmo, E., and Naveira, H. (2015). Evolutionary dynamics of two satellite DNA families in rock lizards of the genus Iberolacerta (Squamata, Lacertidae) : different histories but common traits. *Chromosome Research*, 23(3) :441–461. [2.1.4](#)
- Romanova, L., Deriagin, G., Mashkova, T., Tumeneva, I., Mushegian, A., Kisseelev, L., and Alexandrov, I. (1996). Evidence for selection in evolution of alpha satellite DNA : the central role of CENP-B/pJ α binding region. *Journal of molecular biology*, 261(3) :334–340. [3.1](#), [1.3.1](#), [1.3.1](#), [1.3.2](#), [2.2.2](#), [2.2.4](#)

- Rosandić, M., Paar, V., Basar, I., Glunčić, M., Pavin, N., and Pilaš, I. (2006). CENP-B box and pJ α sequence distribution in human alpha satellite higher-order repeats (HOR). *Chromosome research*, 14(7) :735–753. [2.1.3](#), [2.1.6](#), [2.2.6](#)
- Rosenberg, H., Singer, M., and Rosenberg, M. (1978). Highly reiterated sequences of SIMIAN-SIMIANSIMIANSIMIANSIMIAN. *Science*, 200(4340) :394–402. [2.1.3](#), [2.2.4](#)
- Ross, M. T., Grafham, D. V., Coffey, A. J., Scherer, S., McLay, K., Muzny, D., Platzer, M., Howell, G. R., Burrows, C., Bird, C. P., et al. (2005). The DNA sequence of the human X chromosome. *Nature*, 434(7031) :325–337. [1.3.3](#)
- Ross, R., Hankeln, T., and Schmidt, E. R. (1997). Complex evolution of tandem-repetitive DNA in the Chironomus thummi species group. *Journal of molecular evolution*, 44(3) :321–326. [1.1.2](#)
- Rowell, T. (1971). Organization of caged groups of Cercopithecus monkeys. *Animal behaviour*, 19(4) :625–645. [1.4.1](#)
- Rowell, T. and Dixson, A. (1975). Changes in social organization during the breeding season of wild talapoin monkeys. *Journal of reproduction and fertility*, 43(3) :419–434. [1.4.1](#)
- Rowell, T. E. (1988). The social system of guenons, compared with baboons, macaques and mangabeys. *A primate radiation : Evolutionary biology of the African guenons*, pages 439–451. [1.4.1](#), [1.4.1](#)
- Royle, N. J., Baird, D. M., and Jeffreys, A. J. (1994). A subterminal satellite located adjacent to telomeres in chimpanzees is absent from the human genome. *Nature genetics*, 6(1) :52–56. [1.3.2](#)
- Rubin, C. M., Deininger, P. L., Houck, C. M., and Schmid, C. W. (1980). A dimer satellite sequence in bonnet monkey DNA consists of distinct monomer subunits. *Journal of molecular biology*, 136(2) :151–167. [1.3.1](#), [1.3.1](#)
- Rubin, C. M., Kimura, R. H., and Schmid, C. W. (2002). Selective stimulation of translational expression by Alu RNA. *Nucleic acids research*, 30(14) :3253–3261. [1.1.3](#)
- Rubin, G. M. and Spradling, A. C. (1982). Genetic transformation of Drosophila with transposable element vectors. *Science*, 218(4570) :348–353. [1.1.1](#)
- Rudd, M. K. and Willard, H. F. (2004). Analysis of the centromeric regions of the human genome assembly. *Trends in genetics*, 20(11) :529–533. [1.2.2](#), [1.3.2](#), [1.3.3](#), [1.3.3](#), [3.10](#), [2.1.2](#), [2.2.2](#), [2.2.4](#), [2.3.2](#)
- Rudd, M. K., Wray, G. A., and Willard, H. F. (2006). The evolutionary dynamics of α -satellite. *Genome research*, 16(1) :88–96. [1.3.2](#), [2.1.2](#), [2.2.2](#), [2.2.4](#), [2.3.2](#)

- Ruiz-Ruano, F. J., López-León, M. D., Cabrero, J., and Camacho, J. P. M. (2016). High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Scientific Reports*, 6. [2.1.4](#)
- Rumpler, Y. and Dutrillaux, B. (1989). Chromosomal evolution and speciation in primates. *Revisiones sobre biología celular : RBC*, 23 :1–112. [4.16](#)
- Saksouk, N., Simboeck, E., and Déjardin, J. (2015). Constitutive heterochromatin formation and transcription in mammals. *Epigenetics & chromatin*, 8(1) :1. [1.2.1](#), [1.2.1](#)
- Sakuno, T. and Watanabe, Y. (2009). Studies of meiosis disclose distinct roles of cohesion in the core centromere and pericentromeric regions. *Chromosome research*, 17(2) :239–249. [1.2](#)
- Sangpakdee, W., Tanomtong, A., Fan, X., Pinthong, K., Weise, A., and Liehr, T. (2016). Application of multicolor banding combined with heterochromatic and locus-specific probes identify evolutionary conserved breakpoints in *Hylobates pileatus*. *Molecular cytogenetics*, 9(1) :1. [1.4.3](#)
- Sangwan, I. and O'Brian, M. R. (2002). Identification of a soybean protein that interacts with GAGA element dinucleotide repeat DNA. *Plant Physiology*, 129(4) :1788–1794. [1.1.3](#)
- Sattler, G. D. and Braun, M. J. (2000). Morphometric variation as an indicator of genetic interactions between Black-capped and Carolina chickadees at a contact zone in the Appalachian mountains. *The Auk*, 117(2) :427–444. [1.4.2](#)
- Schindelhauer, D. and Schwarz, T. (2002). Evidence for a fast, intrachromosomal conversion mechanism from mapping of nucleotide variants within a homogeneous α -satellite DNA array. *Genome research*, 12(12) :1815–1826. [1.3.3](#), [2.1.2](#)
- Schlötterer, C. (2000). Evolutionary dynamics of microsatellite DNA. *Chromosoma*, 109(6) :365–371. [1.1.2](#)
- Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., et al. (2009). The B73 maize genome : complexity, diversity, and dynamics. *science*, 326(5956) :1112–1115. [1.1.1](#)
- Schrock, E., Du Manoir, S., Veldman, T., Schoell, B., et al. (1996). Multicolor spectral karyotyping of human chromosomes. *Science*, 273(5274) :494. [1.4.3](#), [4.17](#), [4.18](#)
- Schueler, M. G., Dunn, J. M., Bird, C. P., Ross, M. T., Viggiano, L., Rocchi, M., Willard, H. F., Green, E. D., Program, N. C. S., et al. (2005). Progressive proximal expansion of the primate X chromosome centromere. *Proceedings of the National Academy of Sciences of the United States of America*, 102(30) :10563–10568. [1.2.2](#), [1.3.2](#), [1.3.3](#), [3.11](#), [2.1.2](#), [2.1.4](#), [2.2.2](#), [2.2.4](#), [2.3.2](#), [2.3.4](#), [3.2.2](#)
- Schueler, M. G., Higgins, A. W., Rudd, M. K., Gustashaw, K., and Willard, H. F. (2001). Genomic and genetic definition of a functional human centromere. *Science*, 294(5540) :109–115. [1.2.2](#), [1.3.2](#), [1.3.3](#), [1.3.3](#), [3.12](#), [2.1.2](#), [2.1.4](#), [2.2.2](#), [2.2.4](#), [2.3.2](#), [2.3.4](#), [2.3.4](#), [3.2.2](#)

- Schueler, M. G. and Sullivan, B. A. (2006). Structural and functional dynamics of human centromeric chromatin. *Annu. Rev. Genomics Hum. Genet.*, 7 :301–313. [3.2](#), [2.1.2](#), [2.2.2](#)
- Seabright, M. (1971). A rapid banding technique for human chromosomes. *The Lancet*, 298(7731) :971–972. [1.4.3](#)
- Sequencing, T. C., Consortium, A., et al. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055) :69–87. [1.3.3](#)
- She, X., Horvath, J. E., Jiang, Z., Liu, G., Furey, T. S., Christ, L., Clark, R., Graves, T., Gulden, C. L., Alkan, C., et al. (2004). The structure and evolution of centromeric transition regions within the human genome. *Nature*, 430(7002) :857–864. [1.3.3](#), [2.1.2](#), [2.3.2](#)
- Shepelev, V. A., Alexandrov, A. A., Yurov, Y. B., and Alexandrov, I. A. (2009). The evolutionary origin of man can be traced in the layers of defunct ancestral alpha satellites flanking the active centromeres of human chromosomes. *PLoS Genet*, 5(9) :e1000641. [3.1](#), [1.3.1](#), [3.4](#), [1.3.1](#), [1.3.2](#), [1.3.2](#), [3.6](#), [1.3.2](#), [1.3.3](#), [1.3.3](#), [3.2](#), [2.1.2](#), [2.1.3](#), [2.1.4](#), [2.1.6](#), [2.2.2](#), [2.2.4](#), [2.3.2](#), [2.3.4](#), [3.2.1](#), [3.2.3](#)
- Sheppard, D. J. (2000). *Ecology of the Budongo Forest redtail : Patterns of habitat use and population density in primary and regenerating forest sites*. University of Calgary. [1.4.1](#)
- Shiwago, P. (1939). Recherches sur le caryotype du Rhesus macacus. *Bull Biol Med Exp (USSR)*, 9 :3–8. [1.4.3](#)
- Sienski, G., Dönertas, D., and Brennecke, J. (2012). Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. *Cell*, 151(5) :964–980. [1.1.3](#), [2.2.2](#)
- Silahtaroglu, A., Pfundheller, H., Koshkin, A., Tommerup, N., and Kauppinen, S. (2004). Lna-modified oligonucleotides are highly efficient as fish probes. *Cytogenetic and genome research*, 107(1-2) :32–37. [2.1.3](#)
- Silva, J., Shabalina, S., Harris, D., Spouge, J., and Kondrashov, A. (2003). Conserved fragments of transposable elements in intergenic regions : evidence for widespread recruitment of MIR-and L2-derived sequences within the mouse and human genomes. *Genetical research*, 82(01) :1–18. [1.1.3](#)
- Sineo, L. (2012). The still under construction Cercopithecinae phylogeny. [2.3.2](#)
- Smith, G. P. (1976). Evolution of repeated DNA sequences by unequal crossover. *Science*, 191(4227) :528–535. [1.3.3](#), [2.1.4](#)
- Song, X., Sui, A., and Garen, A. (2004). Binding of mouse VL30 retrotransposon RNA to PSF protein induces genes repressed by PSF : effects on steroidogenesis and oncogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 101(2) :621–626. [1.1.3](#)

- Stanyon, R., Archidiacono, N., and Rocchi, M. (2012). Comparative primate molecular cytogenetics : Revealing ancestral genomes, marker order, and evolutionary new centromeres. In *Post-Genome Biology of Primates*, pages 193–216. Springer. [2.3.4](#)
- Stanyon, R., Arnold, N., Koehler, U., Bigoni, F., and Wienberg, J. (1995). Chromosomal painting shows that “marked chromosomes” in lesser apes and Old World monkeys are not homologous and evolved by convergence. *Cytogenetic and Genome Research*, 68(1-2) :74–78. [1.4.3](#)
- Stanyon, R., Bruening, R., Stone, G., Shearin, A., and Bigoni, F. (2004). Reciprocal painting between humans, De Brazza’s and patas monkeys reveals a major bifurcation in the Cercopithecini phylogenetic tree. *Cytogenetic and genome research*, 108(1-3) :175–182. [1.4.3](#), [2.3.2](#)
- Stanyon, R., Rocchi, M., Capozzi, O., Roberto, R., Misceo, D., Ventura, M., Cardone, M., Bigoni, F., and Archidiacono, N. (2008). Primate chromosome evolution : ancestral karyotypes, marker order and neocentromeres. *Chromosome Research*, 16(1) :17–39. [1.4.3](#), [4.20](#), [2.3.2](#)
- Stanyon, R. and Stone, G. (2008). Phylogenomic analysis by chromosome sorting and painting. *Phylogenomics*, pages 13–29. [1.4.3](#)
- Stanyon, R., Wienberg, J., Romagno, D., Bigoni, F., Jauch, A., and Cremer, T. (1992). Molecular and classical cytogenetic analyses demonstrate an apomorphic reciprocal chromosomal translocation in Gorilla gorilla. *American journal of physical anthropology*, 88(2) :245–250. [1.4.3](#)
- Steiner, F. A. and Henikoff, S. (2015). Diversity in the organization of centromeric chromatin. *Current opinion in genetics & development*, 31 :28–35. [1.2](#), [2.1](#)
- Stephan, W. (1986). Recombination and the evolution of satellite DNA. *Genetical research*, 47(03) :167–174. [1.3.3](#)
- Stephens, A. D., Haase, J., Vicci, L., Taylor, R. M., and Bloom, K. (2011). Cohesin, condensin, and the intramolecular centromere loop together generate the mitotic chromatin spring. *The Journal of cell biology*, 193(7) :1167–1180. [1.2.1](#)
- Stock, A. D. and Hsu, T. (1973). Evolutionary conservatism in arrangement of genetic material. *Chromosoma*, 43(2) :211–224. [1.4.3](#)
- Struhsaker, T. T. (1967). Auditory communication among vervet monkeys (*Cercopithecus aethiops*). *Social communication among primates*, pages 281–324. [1.4.1](#)
- Struhsaker, T. T. (1981). Polyspecific associations among tropical rain-forest primates. *Zeitschrift für Tierpsychologie*, 57(3-4) :268–304. [1.4.1](#)
- Struhsaker, T. T. (2000). The effects of predation and habitat quality on the socioecology of African monkeys : Lessons from the islands of Bioko and Zanzibar. *Old world monkeys*, pages 393–430. [1.4.1](#)

- Struhsaker, T. T. and Leakey, M. (1990). Prey selectivity by crowned hawk-eagles on monkeys in the Kibale Forest, Uganda. *Behavioral Ecology and Sociobiology*, 26(6) :435–443. [1.4.1](#)
- Struhsaker, T. T. and Leland, L. (1979). Socioecology of five sympatric monkey species in the Kibale Forest, Uganda. *Advances in the Study of Behavior*, 9 :159–228. [1.4.1](#)
- Struhsaker, T. T. and Leland, L. (1988). Group fission in redtail monkeys (*Cercopithecus ascanius*) in the Kibale Forest, Uganda. *A primate radiation : Evolutionary biology of the African guenons*, pages 364–388. [1.4.1](#)
- Su, M., Han, D., Boyd-Kirkup, J., Yu, X., and Han, J.-D. J. (2014). Evolution of Alu elements toward enhancers. *Cell reports*, 7(2) :376–385. [1.1.3](#)
- Sujiwattanarat, P., Thapana, W., Srikulnath, K., Hirai, Y., Hirai, H., and Koga, A. (2015). Higher-order repeat structure in alpha satellite DNA occurs in New World monkeys and is not confined to hominoids. *Scientific reports*, 5. [1.3.1](#), [2.1.4](#), [2.2.2](#), [2.2.4](#)
- Sullivan, B. A. and Karpen, G. H. (2004). Centromeric chromatin exhibits a histone modification pattern that is distinct from both euchromatin and heterochromatin. *Nature structural & molecular biology*, 11(11) :1076–1083. [1.2.1](#)
- Sullivan, K. F., Hechenberger, M., and Masri, K. (1994). Human CENP-A contains a histone H3 related histone fold domain that is required for targeting to the centromere. *The Journal of cell biology*, 127(3) :581–592. [1.2.1](#)
- Sumner, A. (1991). Scanning electron microscopy of mammalian chromosomes from prophase to telophase. *Chromosoma*, 100(6) :410–418. [1.2](#)
- Suzuki, N., Nakano, M., Nozaki, N., Egashira, S.-i., Okazaki, T., and Masumoto, H. (2004). CENP-B interacts with CENP-C domains containing Mif2 regions responsible for centromere localization. *Journal of Biological Chemistry*, 279(7) :5934–5946. [1.2.1](#)
- Takahata, N. (1981). Genetic variability and rate of gene substitution in a finite population under mutation and fluctuating selection. *Genetics*, 98(2) :427–440. [1.3.3](#)
- Takeuchi, K., Nishino, T., Mayanagi, K., Horikoshi, N., Osakabe, A., Tachiwana, H., Hori, T., Kurumizaka, H., and Fukagawa, T. (2014). The centromeric nucleosome-like CENP-T-W-S-X complex induces positive supercoils into DNA. *Nucleic acids research*, 42(3) :1644–1655. [1.2.1](#)
- Tantravahi, R., Dev, V., Firschein, I., Miller, D., and Miller, O. (1975). Karyotype of the gibbons *Hylobates lar* and *H. moloch* : inversion in chromosome 7. *Cytogenetic and Genome Research*, 15(2) :92–102. [1.4.3](#)
- Tapanes, E., Detwiler, K. M., and Cords, M. (2016). Bat predation by *Cercopithecus* monkeys : Implications for zoonotic disease transmission. *EcoHealth*, pages 1–5. [1.4.1](#)

- Telenius, H., Ponder, B. A., Tunnacliffe, A., Pelmear, A. H., Carter, N. P., Ferguson-Smith, M. A., Behmel, A., Nordenskjöld, M., and Pfragner, R. (1992). Cytogenetic analysis by chromosome painting using dop-PCR amplified flow-sorted chromosomes. *Genes, Chromosomes and Cancer*, 4(3) :257–263. [1.4.3](#)
- Terada, S., Hirai, Y., Hirai, H., and Koga, A. (2013). Higher-order repeat structure in alpha satellite DNA is an attribute of hominoids rather than hominids. *Journal of human genetics*, 58(11) :752–754. [1.3.1](#), [2.2.2](#), [2.2.4](#)
- Toews, D. P. and Brelsford, A. (2012). The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology*, 21(16) :3907–3930. [1.4.2](#)
- Tosi, A. J. (2008). Forest monkeys and Pleistocene refugia : a phylogeographic window onto the disjunct distribution of the Chlorocebus lhoesti species group. *Zoological Journal of the Linnean Society*, 154(2) :408–418. [4.10](#), [1.4.2](#), [2.2.2](#), [2.3.2](#), [2.3.4](#), [2.3.4](#), [3.2.3](#)
- Tosi, A. J., Buzzard, P. J., Morales, J. C., and Melnick, D. J. (2002). Y-chromosome data and tribal affiliations of Allenopithecus and Miopithecus. *International journal of primatology*, 23(6) :1287–1299. [1.4.2](#)
- Tosi, A. J., Detwiler, K. M., and Disotell, T. R. (2005). X-chromosomal window into the evolutionary history of the guenons (Primates : Cercopithecini). *Molecular phylogenetics and evolution*, 36(1) :58–66. [1.4.2](#)
- Tosi, A. J., Disotell, T. R., Morales, J. C., and Melnick, D. J. (2003). Cercopithecine Y-chromosome data provide a test of competing morphological evolutionary hypotheses. *Molecular phylogenetics and evolution*, 27(3) :510–521. [1.4.2](#)
- Tosi, A. J., Melnick, D. J., and Disotell, T. R. (2004). Sex chromosome phylogenetics indicate a single transition to terrestriality in the guenons (tribe Cercopithecini). *Journal of human evolution*, 46(2) :223–237. [2.3.2](#), [2.3.4](#)
- Tsingalia, H. and Rowell, T. E. (1984). The behaviour of adult male blue monkeys. *Zeitschrift für Tierpsychologie*, 64(3-4) :253–268. [1.4.1](#)
- Turleau, C. and De Grouchy, J. (1972). Caryotypes de l'homme et du chimpanzé. comparaison de la topographie des bandes. Mécanismes évolutifs possibles. *CR Acad. Sci. (Paris)*, 274 :2355–2357. [1.4.3](#)
- Tyler-Smith, C. and Brown, W. R. (1987). Structure of the major block of alphoid satellite DNA on the human Y chromosome. *Journal of molecular biology*, 195(3) :457–470. [1.3.1](#)
- Ugarković, Đ. i. (2009). *Centromere-competent DNA : structure and evolution*. Springer. [3.13](#), [2.2.4](#), [2.3.4](#)
- van der Kuyl, A. C., Kuiken, C. L., Dekker, J. T., and Goudsmit, J. (1995). Phylogeny of African monkeys based upon mitochondrial 12S rRNA sequences. *Journal of Molecular Evolution*, 40(2) :173–180. [1.4.2](#)

- van Steensel, B., Delrow, J., and Bussemaker, H. J. (2003). Genomewide analysis of *Drosophila* GAGA factor target genes reveals context-dependent DNA binding. *Proceedings of the National Academy of Sciences*, 100(5) :2580–2585. [1.1.3](#)
- Vansant, G. and Reynolds, W. F. (1995). The consensus sequence of a major Alu subfamily contains a functional retinoic acid response element. *Proceedings of the National Academy of Sciences*, 92(18) :8229–8233. [1.1.3](#)
- Ventura, M., Antonacci, F., Cardone, M. F., Stanyon, R., D'Addabbo, P., Cellamare, A., Sprague, L. J., Eichler, E. E., Archidiacono, N., and Rocchi, M. (2007). Evolutionary formation of new centromeres in macaque. *Science*, 316(5822) :243–246. [1.2.2](#), [1.4.3](#), [2.3.2](#)
- Ventura, M., Archidiacono, N., and Rocchi, M. (2001). Centromere emergence in evolution. *Genome Research*, 11(4) :595–599. [4.1.9](#), [1.4.3](#), [2.3.2](#), [3.2.2](#)
- Verdel, A., Jia, S., Gerber, S., Sugiyama, T., Gygi, S., Grewal, S. I., and Moazed, D. (2004). RNAi-mediated targeting of heterochromatin by the RITS complex. *Science*, 303(5658) :672–676. [1.1.3](#)
- Vergnaud, G. and Denoeud, F. (2000). Minisatellites : mutability and genome architecture. *Genome Research*, 10(7) :899–907. [1.1.2](#)
- Vershinin, A., Alkhimova, E., and Heslop-Harrison, J. (1996). Molecular diversification of tandemly organized DNA sequences and heterochromatic chromosome regions in some Triticeae species. *Chromosome Research*, 4(7) :517–525. [1.1.2](#)
- Vilà, C., Leonard, J. A., Götherström, A., Marklund, S., Sandberg, K., Lidén, K., Wayne, R. K., and Ellegren, H. (2001). Widespread origins of domestic horse lineages. *Science*, 291(5503) :474–477. [1.1.2](#)
- Vissel, B. and Choo, K. (1991). Four distinct alpha satellite subfamilies shared by human chromosomes 13, 14 and 21. *Nucleic acids research*, 19(2) :271–277. [1.3.2](#)
- Vissel, B. and Choo, K. H. (1989). Mouse major (γ) satellite DNA is highly conserved and organized into extremely long tandem arrays : implications for recombination between non-homologous chromosomes. *Genomics*, 5(3) :407–414. [1.2.2](#)
- Volpe, T. A., Kidner, C., Hall, I. M., Teng, G., Grewal, S. I., and Martienssen, R. A. (2002). Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *science*, 297(5588) :1833–1837. [1.1.3](#), [2.2.2](#)
- Wade, C., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., Lear, T., Adelson, D., Bailey, E., Bellone, R., et al. (2009). Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science*, 326(5954) :865–867. [1.2.2](#), [2.3.2](#)
- Walsh, J. B. (1987). Persistence of tandem arrays : implications for satellite and simple-sequence DNAs. *Genetics*, 115(3) :553–567. [1.3.3](#)

- Warburton, P. and Willard, H. (1996). Evolution of centromeric alpha satellite DNA : molecular organization within and between human and primate chromosomes. *Human genome evolution*, pages 121–145. [1.3.3](#)
- Warburton, P. E., Haaf, T., Gosden, J., Lawson, D., and Willard, H. F. (1996). Characterization of a chromosome-specific chimpanzee alpha satellite subset : evolutionary relationship to subsets on human chromosomes. *Genomics*, 33(2) :220–228. [1.3.1](#), [1.3.2](#), [2.1.2](#)
- Warburton, P. E., Hasson, D., Guillem, F., Lescalle, C., Jin, X., and Abrusan, G. (2008). Analysis of the largest tandemly repeated DNA families in the human genome. *BMC genomics*, 9(1) :1. [2.1.4](#), [2.3.4](#)
- Warburton, P. E. and Willard, H. F. (1990). Genomic analysis of sequence variation in tandemly repeated DNA : Evidence for localized homogeneous sequence domains within arrays of α -satellite DNA. *Journal of molecular biology*, 216(1) :3–16. [2.1.2](#)
- Warburton, P. E. and Willard, H. F. (1995). Interhomologue sequence variation of alpha satellite DNA from human chromosome 17 : Evidence for concerted evolution along haplotypic lineages. *Journal of molecular evolution*, 41(6) :1006–1015. [1.3.3](#), [2.1.2](#), [2.2.2](#)
- Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301) :236–244. [2.1.6](#), [2.2.6](#)
- Watanabe, H., Fujiyama, A., Hattori, M., Taylor, T., Toyoda, A., Kuroki, Y., Noguchi, H., BenKahla, A., Lehrach, H., Sudbrak, R., et al. (2004). DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature*, 429(6990) :382–388. [1.3.3](#)
- Watanabe, T., Tomizawa, S.-i., Mitsuya, K., Totoki, Y., Yamamoto, Y., Kuramochi-Miyagawa, S., Iida, N., Hoki, Y., Murphy, P. J., Toyoda, A., et al. (2011). Role for piRNAs and non-coding RNA in de novo DNA methylation of the imprinted mouse Rasgrf1 locus. *Science*, 332(6031) :848–852. [1.1.3](#)
- Watanabe, Y. (2005). Shugoshin : guardian spirit at the centromere. *Current opinion in cell biology*, 17(6) :590–595. [1.2.1](#)
- Waye, J. and Willard, H. (1986). Structure, organization, and sequence of alpha satellite DNA from human chromosome 17 : evidence for evolution by unequal crossing-over and an ancestral pentamer repeat shared with the human X chromosome. *Molecular and cellular biology*, 6(9) :3156–3165. [1.3.1](#), [1.3.2](#)
- Waye, J. S. and Willard, H. F. (1989). Concerted evolution of alpha satellite DNA : evidence for species specificity and a general lack of sequence conservation among alphoid sequences of higher primates. *Chromosoma*, 98(4) :273–279. [1.3.3](#), [2.2.4](#)
- Weber, S. A., Gerton, J. L., Polancic, J. E., DeRisi, J. L., Koshland, D., and Megee, P. C. (2004). The kinetochore is an enhancer of pericentric cohesin binding. *PLoS Biol*, 2(9) :e260. [1.2.1](#)

- Weierich, C., Brero, A., Stein, S., von Hase, J., Cremer, C., Cremer, T., and Solovei, I. (2003). Three-dimensional arrangements of centromeres and telomeres in nuclei of human and murine lymphocytes. *Chromosome research*, 11(5) :485–502. [1.1.3](#)
- Weisblum, B. and De Haseth, P. L. (1972). Quinacrine, a chromosome stain specific for deoxyadenylate-deoxythymidylate-rich regions in DNA. *Proceedings of the National Academy of Sciences*, 69(3) :629–632. [1.4.3](#)
- Wevrick, R. and Willard, H. F. (1989). Long-range organization of tandem arrays of alpha satellite DNA at the centromeres of human chromosomes : high-frequency array-length polymorphism and meiotic stability. *Proceedings of the National Academy of Sciences*, 86(23) :9394–9398. [1.2.2](#), [1.3.1](#), [1.3.2](#)
- Whitten, P. L. (1983). Diet and dominance among female vervet monkeys (*Cercopithecus aethiops*). *American Journal of Primatology*, 5(2) :139–159. [1.4.1](#)
- Wienberg, J., Jauch, A., Stanyon, R., and Cremer, T. (1990). Molecular cytotaxonomy of primates by chromosomal in situ suppression hybridization. *Genomics*, 8(2) :347–350. [1.4.3](#)
- Wienberg, J. and Stanyon, R. (1998). Comparative chromosome painting of primate genomes. *ILAR Journal*, 39(2-3) :77–91. [4.17](#), [4.18](#)
- Wienberg, J., Stanyon, R., Jauch, A., and Cremer, T. (1992). Homologies in human and *Macaca fuscata* chromosomes revealed by in situ suppression hybridization with human chromosome specific DNA libraries. *Chromosoma*, 101(5-6) :265–270. [1.4.3](#)
- Willard, H. F. (1985). Chromosome-specific organization of human alpha satellite DNA. *American journal of human genetics*, 37(3) :524. [1.3.2](#), [1.3.3](#)
- Willard, H. F. (1990). Centromeres of mammalian chromosomes. *Trends in Genetics*, 6 :410–416. [1.2.2](#)
- Willard, H. F. (1991). Evolution of alpha satellite. *Current opinion in genetics & development*, 1(4) :509–514. [1.3.3](#), [2.1.2](#), [2.2.2](#), [2.3.2](#)
- Willard, H. F. and Waye, J. S. (1987). Hierarchical order in chromosome-specific human alpha satellite DNA. *Trends in Genetics*, 3 :192–198. [1.3.1](#), [1.3.2](#), [1.3.3](#), [2.1.2](#), [2.2.2](#), [2.3.2](#)
- Wilson, D. E. and Reeder, D. M. (2005). *Mammal species of the world : a taxonomic and geographic reference*, volume 1. JHU Press. [1.3](#), [1.4.1](#), [2.1.2](#), [2.3.2](#)
- Windfelder, T. L. and Lwanga, J. S. (2004). Group fission in red-tailed monkeys (*Cercopithecus ascanius*) in Kibale National Park, Uganda. In *The guenons : Diversity and adaptation in African monkeys*, pages 147–159. Springer. [1.4.1](#)
- Wolfheim, J. H. (1983). *Primates of the world : distribution, abundance and conservation*. Psychology Press. [1.4.1](#)

- Wong, A. and Rattner, J. (1988). Sequence organization and cytological localization of the minor satellite of mouse. *Nucleic acids research*, 16(24) :11645–11661. [1.2.2](#)
- Wong, L. H., Brettingham-Moore, K. H., Chan, L., Quach, J. M., Anderson, M. A., Northrop, E. L., Hannan, R., Saffery, R., Shaw, M. L., Williams, E., et al. (2007). Centromere RNA is a key component for the assembly of nucleoproteins at the nucleolus and centromere. *Genome research*, 17(8) :1146–1160. [1.2.2](#), [2.2.2](#)
- Wood, V., Gwilliam, R., Rajandream, M.-A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S., et al. (2002). The genome sequence of Schizosaccharomyces pombe. *Nature*, 415(6874) :871–880. [1.1.2](#)
- Wrangham, R. and Waterman, P. (1981). Feeding behaviour of vervet monkeys on Acacia tortilis and Acacia xanthophloea : with special reference to reproductive strategies and tannin production. *The Journal of Animal Ecology*, pages 715–731. [1.4.1](#)
- Wyman, A. R. and White, R. (1980). A highly polymorphic locus in human DNA. *Proceedings of the National Academy of Sciences*, 77(11) :6754–6758. [1.1.2](#)
- Xing, J., Wang, H., Zhang, Y., Ray, D. A., Tosi, A. J., Disotell, T. R., and Batzer, M. A. (2007). A mobile element-based evolutionary history of guenons (tribe Cercopithecini). *BMC biology*, 5(1) :1. [1.4.2](#), [2.3.2](#), [2.3.4](#)
- Yamagishi, Y., Sakuno, T., Shimura, M., and Watanabe, Y. (2008). Heterochromatin links to centromeric protection by recruiting shugoshin. *Nature*, 455(7210) :251–255. [1.2](#)
- Yoda, K., Nakamura, T., Masumoto, H., Suzuki, N., Kitagawa, K., Nakano, M., Shinjo, A., and Okazaki, T. (1996). Centromere protein B of African green monkey cells : gene structure, cellular expression, and centromeric localization. *Molecular and cellular biology*, 16(9) :5169–5177. [1.2.2](#), [1.3.1](#), [1.3.1](#), [2.1.4](#), [3.2.1](#)
- Youn, B.-S., Lim, C. L., Shin, M. K., Hill, J. M., and Kwon, B. S. (2002). An intronic silencer of the mouse perforin gene. *Molecules and cells*, 13(1) :61–68. [1.1.3](#)
- Yunis, J. J. and Yasmineh, W. G. (1971). Heterochromatin, satellite DNA, and cell function. *Science*, 174(4015) :1200–1209. [1.1.3](#), [2.1.2](#)
- Zeeve, S. R. (1991). *Behavior and ecology of primates in the Lomako Forest, Zaire*. PhD thesis. [1.4.1](#), [1.4.1](#)
- Zeitlin, S. G., Baker, N. M., Chapados, B. R., Soutoglou, E., Wang, J. Y., Berns, M. W., and Cleveland, D. W. (2009). Double-strand DNA breaks recruit the centromeric histone CENP-A. *Proceedings of the National Academy of Sciences*, 106(37) :15762–15767. [2.3.4](#)
- Zhou, K., Aertsen, A., and Michiels, C. W. (2014). The role of variable DNA tandem repeats in bacterial adaptation. *FEMS microbiology reviews*, 38(1) :119–141. [3.7](#)

- Zhou, Y.-H., Zheng, J. B., Gu, X., Li, W.-H., and Saunders, G. F. (2000). A novel Pax-6 binding site in rodent B1 repetitive elements : coevolution between developmental regulation and repeated elements ? *Gene*, 245(2) :319–328. [1.1.3](#)
- Zhou, Y.-H., Zheng, J. B., Gu, X., Saunders, G. F., and Yung, W.-K. A. (2002). Novel PAX6 binding sites in the human genome and the role of repetitive elements in the evolution of gene regulation. *Genome research*, 12(11) :1716–1722. [1.1.3](#)
- Zucker, E. L. and Kaplan, J. (1981). Allomaternal behavior in a group of free-ranging patas monkeys. *American Journal of Primatology*, 1(1) :57–64. [1.4.1](#)