Page de garde

Page Plagia

Remerciement…

Page de garde

Table des figures

Liste de tableau ??

Liste des abbréviations

Table des matières

**Word n'a pas trouvé d'entrées pour votre table des matières.**

Introduction

Scientific Background

In eukaryotic cell, proper chromosome segregation during cell division is ensured by a chromatin locus known as the centromere, first described by Walther Flemming in 1882 as primary constriction. This specialized locus enables the assembly of the kinetochore, a multi-protein structure, which attaches chromosomes to spindle microtubules to ensure the equal partitioning of genetic material between daughter cells of a dividing cell. Defects or errors in chromosome segregation lead to aneuploidy or polyploidy cells (The Consequences of Chromosome Segregation Errors in Mitosis and Meiosis, Tamara Potapova , 2017). While the proteins associated with this locus are conserved among organisms, DNA sequences are dissimilar among organisms and often within the same organism. (The Past, Present, and Future of Human Centromere Genomics, Megan E. Aldrup-MacDonald, 2014) Human centromeres are defined by alpha satellite DNA. Alpha satellites are a family of AT-rich repetitive sequences and are based on a 171-bp monomeric subunit. Different monomers within the family sharing 50%-70% sequence similarity are tandemly organised in a head-to-tail pattern to form a higher-order repeat (HOR) that is itself reiterated hundreds or thousands of time and form extensive arrays. At chromosome-specific centromere, monomers within a HOR has a high sequence conservation (>99%) (Chromosome-specific subsets of human alpha satellite DNA: analysis of sequence divergence within and between chromosomal subsets and evidence for an ancestral pentameric repeat. Willard HF, Waye JS, 1987). The length of these alpha satellite arrays and thereby centromeric loci varies from hundred kilobases to megabases.( Greig, G.M.; Parikh, S.; George, J.; Powers, V.E.; Willard, H.F. Molecular cytogenetics of alpha satellite DNA from chromosome 12: Fluorescence in situ hybridization and description of DNA and array length polymorphisms. Cytogenet. Cell Genet. 1991, 56, 144–148. ; Abruzzo, M.A.; Griffin, D.K.; Millie, E.A.; Sheean, L.A.; Hassold, T.J. The effect of Y-chromosome alpha-satellite array length on the rate of sex chromosome disomy in human sperm. Hum. Genet. 1996, 97, 819–823. ; Oakey, R.; Tyler-Smith, C. Y chromosome DNA haplotyping suggests that most European and Asian men are descended from one of two males. Genomics 1990, 7, 325–330.)

Aneuploidy

Chez les humains, et la plupart des primates, les centromeres sont composes d’ADN satellite alpha. Les satellites alpha sont une famille de sequences répétées riche en nucleotides A et T. Il s’agit de monomers d’approximativement 171pb qui sont répétées en tandem en modèle tête à queue. Différent monomers au sein de la famille partage 50 à 70% de similarité de sequence. Cependant, au centromere, a sous ensemble de monomers chromosome-spécifique forme un Higher Order Repeat qui est lui-même répété des centaines ou des milliers de fois

Centromeric chromatin is defined in nucleosomes which contains, a specialized variant of the core histone H3, CenH3 or CENP-A which can associate with the following histones : H2A, H2B and H4 (Sullivan et al., 1994). Un nucleosome sur 6 à 8 au centromere serait CenH3 et entourés de nucleosomes canoniques (Joglekar et al.,2008; Johnston et al., 2010; Fukagawa and Earnshaw,2014)

Centromeric chromatin is defined and enriched in nucleosomes containing a specialized variant of the core histone H3, CenH3

Centromeres are defined and enriched by a variant

.

In eukaryotic cell, proper chromosome segregation during cell division is directed by a chromatin locus known as the centromere, first term coiled in 1936 (Darlington - 1936 - The External Mechanics of the Chromosomes.). First describe by Walther Flemming in 1882, human centromeres are visible as primary constriction.

In humans, and most primates, centromeres are composed of alpha satellite DNA. Alpha satellites are a family of AT-rich repetitive sequences: monomers of ~171 bp that are tandemly repeated in a head-to-tail pattern. Different monomers within the family share 50% to 70% sequence similarity. However, at centromeres, a chromosome- specific subset of monomers forms a higher-order repeat (HOR) that is itself reiterated hundreds or thousands of times. Centromeric loci are hence characterized by a megabase-sized, homogeneous array of HORs that are near- identical (>99% similarity). The constituent monomers of a HOR have different sequences and come in different chromosome-specific combinations but, within an array, HORs are virtually indistinguishable. This makes it extremely challenging to characterize centromeres at the genetic level, and the reference assembly of the human genome only recently started to include *in silico* models of centromeric sequences.

Regional centromeres that are hundreds of kilobases long, consisting of higher order arrays of AT-rich alpha satellite repeats assembled in higher order repeats

Specifically, centromeric chromatin is enriched in nucleosomes containing a specialized variant of the core histone H3, CenH3

La chromatine centromérique est différente du reste des chromosomes : elle forme le centromère qui est la constriction primaire des chromosomes. Ce centromère, qui joint les deux bras des chromosomes, joue un rôle fondamental dans la ségrégation de ceux-ci au cours de la mitose et de la méiose. C’est en effet au niveau du centromère que se forme le kinétochore, la structure protéique responsable de l’attachement des microtubules qui permet le mouvement des chromosomes vers les deux cellules filles (Vagnarelli, Ribeiro et al. 20

08). Des erreurs de répartition des chromosomes peuvent avoir des conséquences dramatiques pour la survie de la cellule ou de l’organisme.

Le kinétochore s’assemble sur une structure chromatinienne particulière localisée au niveau du centromère. Cette caractéristique, conservée au cours de l’évolution, laissait penser à l’existence au sein du centromère d’une séquence génomique spécifique. Celle-ci a longtemps été recherchée en vain. En effet, même si la séquence génomique centromérique est nécessaire et suffisante à la formation du kinétochore chez S.cerevisiae, ce n’est pas le cas chez les autres eucaryotes (Amor, Kalitsis et al. 2004). Cependant, à l’exception de

S.cerevisiae, les eucaryotes contiennent au niveau de leur centromère un ADN rép

été (Tyler-Smith and Floridia 2000). Ces séquences, au sein desquelles on retrou

ve également des séquences répétées dispersées, sont appelées :

alpha satellites chez l’Homme

satellites mineurs chez la souris

régions AATAT et TTCTC chez la drosophile

central core chez la levure

S.pombe

Toutefois, la présence de séquences répétées ne suffit pas à assurer la fonction

centromérique (Craig, Wong et al. 2003). En effet des études portant sur des

chromosomes anormaux, caractérisés par une perte ducentromère, montrent que de

nouveaux centromères peuvent se former sur des régions dépourvues d’alpha

satellites mais présentant néanmoins une séquence riche en AT (du Sart, Cancilla et

al. 1997). La richesse des centromères en AT est untrait commun à toutes les

espèces. En effet, si le centromère de S.cerevisiae n’est pas composé de séquences

répétées, les trois éléments Cde qui forment son centromère sont enrichis en AT

(Clarke and Carbon 1985).

La structure des centromères ne reposant pas exclusivement sur la présence de

séquences répétées, les regards se sont alors tourn

és vers l’implication possible de

mécanismes épigénétiques. Il existe en effet un var

iant d’histone spécifique des

centromères, CenH3. Il est nommé de manière différe

nte selon les espèces

considérées :

hCENPA chez l’Homme

mCENPA chez la souris

cid chez la drosophile

CNP1 chez

S.pombe

Cse4 chez

S.cerevisiae

C’est la seule protéine centromérique qui existe ch

ez toutes les espèces. La

déplétion de CenH3 chez la levure, la drosophile, l

a souris ou l’Homme conduit à des

défauts sévères de ségrégation des chromosomes et d

e localisation centromérique des

autres protéines du kinétochore (Stoler, Keith et a

l. 1995; Valdivia, Figueroa et al.

1998; Howman, Fowler et al. 2000; Blower and Karpen

2001). De plus, des souris

déficientes pour CenH3 meurent au stade embryonnair

e indiquant que CenH3 aurait

un rôle fondamental dans la structure du kinétochor

e. Bien que CenH3 soit nécessaire

au recrutement d’autres protéines du kinétochore co

mme CENPC, elle n’est pas

suffisante pour la formation d’un centromère foncti

onnel. En effet, lorsque CenH3 est

surexprimée, elle se localise de manière ectopique

sur de la chromatine non

centromérique et recrute CENPC ainsi que d’autres p

rotéines du kinétochore.

Cependant, cela n’est pas suffisant pour créer des

centromères ectopiques

fonctionnels (Van Hooser, Ouspenski et al. 2001).

Néanmoins, il est important de noter que l’incorpor

ation de CENH3 dans la

chromatine centromérique n’est pas exclusive. En ef

fet, une fraction de l’histone H3

Chromosomes undergo major changes in structure and organization during the cell cycle (FIG. 1). They condense during mitosis, and during this stage, as first described by Walther Flemming in 1882 (REF. 1), human centromeres become visible as chromosomal constrictions. The spe- cialized nature and environment of centromeric chro- matin enables the assembly of the kinetochore, which is a large, multi-protein complex that attaches to micro- tubules during cell division (for reviews, see REFS 2,3), thereby ensuring equal partitioning of genetic material between daughter cells. Following each cell division, chromatin decondenses, the structure and biochemical composition of centromeres change, and kinetochores disassemble. During mitosis, this decompaction is vis- ualized by weak DNA staining on individual chromo- somes (FIG. 1). In interphase, specialized densely stained chromocentres become visible in mouse cells. They cor- respond to pericentric heterochromatin (PHC) bringing several chromosomes together.

The ability of cells to properly segregate a complete set of chromosomes

to each daughter cell during mitosis is dependent on

a unique chromatin domain known as the centromere. Human

centromeres are located on 1- to 4- Mb chromosomal regions

that are composed of tandemly repeated arrays of a ∼171-bp

element, termed α-satellite DNA repeats (Willard, 1985; Cleveland

et al., 2003). In a proportion of centromeric DNA (Blower

et al., 2002; Sullivan and Karpen, 2004), the canonical histone

H3 is replaced with a histone variant initially identified in humans

and named CENP-A (Earnshaw and Rothfield, 1985;

Palmer et al., 1987). Paradoxically, except in budding yeast,

this basic unit of chromosome inheritance is not defined by the

DNA sequence, as α-satellite DNA sequences are neither sufficient

nor essential for centromere identity (Karpen and Allshire,

1997; Cleveland et al., 2003; Stimpson and Sullivan, 2010).

Rather, centromeres are defined and maintained epigenetically

by CENP-A–containing chromatin that has been shown to identify,

maintain, and propagate centromere function indefinitely

in human cells and fission yeast (Fachinetti et al., 2013).

The chromatin composition of human centromeres has

proven challenging to elucidate given their large (1–4 Mb) size

and highly repetitive nature (Sullivan et al., 2001). Although

it is known that centromeric chromatin is composed of both

CENP-A– and H3-containing chromatin (Blower et al., 2002;

Sullivan and Karpen, 2004), the precise nature and composition

of CENP-A–containing chromatin in vivo has remained highly

controversial. There is agreement that CENP-A readily assembles

in vitro into octameric nucleosomes containing two molecules

each of CENP-A and histones H2A, H2B, and H4 (Black

et al., 2007; Falk et al., 2016) and forms a structure similar to

that assembled with histone H3 (Tachiwana et al., 2011). However,

the precise nature and composition of CENP-A–containing

chromatin in vivo at multiple cell cycle points has not been settled.

One prominent proposal is that human CENP-A chromatin

oscillates in a cell cycle–dependent manner between two forms:

a hemisome (containing only one molecule of histones H2A,

(Human centromeric CENP-A chromatin is

a homotypic, octameric nucleosome at all

cell cycle points)

The centromere is part of a chromosome essential for correct chromosome segregation during cell division,

serving as the point to which the spindle fiber attaches via the kinetochore. Centromeres of higher

eukaryotes generally contain large amounts of tandem repeat DNA. Alpha satellite DNA (AS) is the

most abundant tandem repeat DNA of primate centromeres1,2, although this may not be true of suborder

Strepsirrhini, one member of which (the aye-aye, *Daubentonia madagascariensis*) is known to carry

other unrelated tandem repeat DNA as main components of its centromeres3. Strepsirrhini is a taxon

that diverged from other groups in an early stage of the primate evolution. The length of the repeat

units of AS is approximately 170 bp in parvorder Catarrhini, which includes hominoids (superfamily

Hominoidea; humans, great apes, and small apes) and Old World monkeys (family Cercopithecidae;

macaques, baboons, and related monkeys found in Africa and Asia)4–6. In New World monkeys

(parvorder Platyrrhini; monkeys inhabiting Central and South America), the length of the repeat units

is approximately 340 bp6,7. Sequence analyses have suggested that this difference is due to an ancient

dimeric structure of AS of New World monkeys6. The parvorders Catarrhini and Platyrrhini constitute

the infraorder Simiiformes (simian primates).

(Higher-order repeat structure in

alpha satellite DNA occurs in New

World monkeys and is not confined

to hominoids)

In humans, and most primates, centromeres are composed of alpha satellite DNA. Alpha satellites are a family of AT-rich repetitive sequences: monomers of ~171 bp that are tandemly repeated in a head-to-tail pattern. Different monomers within the family share 50% to 70% sequence similarity. However, at centromeres, a chromosome- specific subset of monomers forms a higher-order repeat (HOR) that is itself reiterated hundreds or thousands of times. Centromeric loci are hence characterized by a megabase-sized, homogeneous array of HORs that are near- identical (>99% similarity). The constituent monomers of a HOR have different sequences and come in different chromosome-specific combinations but, within an array, HORs are virtually indistinguishable. This makes it extremely challenging to characterize centromeres at the genetic level, and the reference assembly of the human genome only recently started to include *in silico* models of centromeric sequences.

Centromeres are better (or more easily) defined at a functional level, being the specialized loci that mediate spindle attachment and ensure proper chromosome segregation during cell division. Functional centromeres are defined by the presence of inner centromere proteins that epigenetically mark the site of kinetochore assembly. Specifically, centromeric chromatin is enriched in nucleosomes containing a specialized variant of the core histone H3, CenH3 (historically known as CENP-A). CenH3 incorporation is mediated by its dedicated histone chaperone HJURP, and its presence is required for proper kinetochore assembly. New CenH3 is specifically deposited at centromeres, hence most CenH3 is bound to alpha satellite DNA. No sequence preference or consensus motif has however been identified, and its deposition thought to occur independently of DNA sequence. This is in contrast to e.g. CENP-B, an inner centromere protein that instead recognizes a 17 bp motif (CENP-B box) present in some of the constituent monomers of centromeric HORs.

Centromeres are essential loci that ensure an even redistribution of the proper set of chromosomes to each daughter of a dividing cell. Given their crucial role in cell biology and disease, human centromeres are relatively well characterized and studied extensively at a biochemical and functional level. However, due to the repetitive nature of the underlying sequence, little is known about centromere diversity at the DNA level. Centromeres can exhibit different types of polymorphisms - varying in total array size, HOR length and/or composition, or in terms of sequence of constituent monomers. The basis, extent, and potential implications of this genetic variation are largely unknown.

In order to harness the wealth of existing genomics data and start probing these questions, we need an efficient and reliable method to quantify centromere diversity. This will be developed using simulated data as a benchmark, and applied to large genomics cohorts to evaluate within- and between-individual variation and explore global associations. If applicable, implemented methods and data-driven predictions will be validated experimentally.

À la fin des années 1990, l’émergence des puces à ADN puis des technologies de séquençages à ultra-débit à partir du milieu des années 2000 a fait émerger de nouvelles questions relatives à l’accumulation des grands volumes de données en biologie (Barrett and Edgar 2006; Reimand et al. 2016). Cette évolution quasi-exponentielle (Barrett et al. 2011) a conduit plusieurs grandes institutions, telles que la European Bioinformatics Institute (EBI, http://www.ebi.ac.uk) et le National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov) à développer de nouveaux outils de stockage, aussi appelés espaces de dépôts (ou repositories en anglais), qui permettent de centraliser et organiser les données « omiques », telles que transcript-, proté-, gén-, métabol-omiques. Ainsi, ces repositories ont pour objectifs de stocker de manière pérenne les informations générées par les chercheurs et de les rendre accessible à la communauté scientifique. Avant de soumettre une publication, il est généralement imposé aux auteurs de déposer les données brutes ou prétraitées (normalisées par exemple) au sein de ces espaces de dépôts.

Contrairement aux banques et repositories « spécialisés » qui ont pour vocation de centraliser les données d’un même sujet biologique, tels que PubChem (Kim et al. 2016) pour les substances chimiques, OpenfMRI (Poldrack et al. 2013) en neuroscience ou encore BioModels Database (Chelliah, Laibe, and Le Novère 2013) pour les modèles biologiques, certains espaces de dépôts sont qualifiés de « généralistes ». Ceux-ci ont pour fonction le stockage de données brutes sans aucune restriction sur la thématique étudiée. Plusieurs repositories sont ainsi disponibles selon les différentes disciplines «omiques». Parmi eux, GEO (Gene Expression Omnibus) (Edgar, Domrachev, and Lash 2002) et ArrayExpress (Parkinson et al. 2005) font figure de précurseurs et de références dans le domaine. Alors que GEO est développé et maintenu par le NCBI, ArrayExpress dépend de l’EBI. Ces deux repositories permettent aux scientifiques de déposer leurs données brutes de transcriptomique et génomique générées issues d’expériences de puces à ADN et de séquençage à haut-débit (RNA-seq ou ChIP-seq par exemple). Ils fournissent également un certain nombre d’outils d’analyse bio-informatique permettant aux utilisateurs, par exemple, de consulter le profil d’expression de gènes d’intérêts. A l’instar de ces deux exemples, PRIDE (PRoteomics Identifications Database) est un des espaces de dépôts de référence pour la protéomique et notamment pour le stockage des données de spectrométrie de masse (Vizcaíno et

￼￼1

￼￼

al. 2016). Créée en 2004 par l’EBI, PRIDE permet de stocker des séquences protéiques et peptidiques ainsi que des modifications post-traductionnelles.

Bien que développés initialement pour les biologistes, ces espaces de dépôts présentent néanmoins une restriction majeure : l’exploitation des données qu’ils contiennent nécessitent des compétences bio-informatiques et statistiques avancées. Ce qui constitue un frein pour la plupart des chercheurs non-génomistes. De plus, dans la majorité des cas, ceux-ci ne souhaitent pas ré-analyser les données brutes disponible mais désirent « simplement » comparer leurs propres résultats issus d’analyses « omiques » avec ceux publiés par leurs confrères. Ces résultats se présentent habituellement sous la forme de listes d’entités biologiques (gènes, transcrits, protéines, sondes nucléotidiques, ...) générées par les différentes étapes de filtration des données décrites dans les publications. Ainsi, ces listes peuvent représenter par exemple, des listes de sondes nucléotidiques d’une puce à ADN sur- ou sous-exprimées, ou encore une liste de protéines différentielles identifiées par spectrométrie de masse (LC-MS/MS par exemple) entre deux conditions expérimentales à comparer. Bien que ces listes d’entités biologiques constituent le cœur même des résultats présentés dans les publications scientifiques, celles-ci sont la plupart du temps disponibles dans les publications scientifiques sous la forme de tableaux souvent perdus dans des fichiers supplémentaires dans des formats (PDF par exemple) difficilement exploitables à grande échelle. Certaines banques de données, telles que MSigDB (Molecular Signature DataBase, http://software.broadinstitute.org/gsea/msigdb) développé par le Broad Institut (Subramanian et al. 2005), permettent d’héberger des listes de gènes annotés (ou gene sets en anglais). Néanmoins, ces banques ne constituent pas des espaces de dépôts en ce sens qu’elles ne permettent pas aux utilisateurs de déposer eux-mêmes leurs listes d’entités biologiques. Ainsi, contrairement aux données brutes, aucun repository n’est dédié au stockage de ces informations précieuses décrivant les différentes étapes du processus d’analyse et des listes d’entités biologiques résultantes. C’est face à ce constat que mon encadrant de stage, Frédéric CHALMEL, a initié le projet GeneULike.