University of Trás-os-Montes and Alto Douro

# Unlocking Robertsonian translocated chromosomes in its repetitive DNA content to tackle genomic and mechanistic issues

Master's Degree in Comparative, Technological and Molecular Genetics

**Mariana Faria Lopes**

**Supervisors:**

Professora Doutora Raquel Maria Garcia dos Santos Chaves

Professora Doutora Margarida Henriques da Gama Carvalho


Vila Real, 2019

University of Trás-os-Montes and Alto Douro

# Unlocking Robertsonian translocated chromosomes in its repetitive DNA content to tackle genomic and mechanistic issues

Master's Degree in Comparative, Technological and Molecular Genetics

Mariana Faria Lopes

Supervisors:

Professora Doutora Raquel Maria Garcia dos Santos Chaves

_____

Professora Doutora Margarida Henriques da Gama Carvalho

_____

Composition of the jury:

_____

_____

_____

Vila Real, 2019

I declare for all purposes that this dissertation meets the technical and scientific standards required by the regulations of the University of Trás-os-Montes and Alto Douro. The presented doctrines are the exclusive responsibility of the author.

This dissertation was specifically prepared to obtain the Master´s degree in Technological, Comparative and Molecular Genetics.

***Questions you cannot answer are usually far better for you than answers you cannot question.***

Yuval Noah Harar

## Agradecimentos

Quando uma etapa é concluída o mérito nunca é individual. Este é o tempo e o espaço para agradecer, formal e informalmente, a quem, de uma forma ou de outra, contribuiu para que tudo o que se segue ganhasse forma.

À Universidade de Trás-os-Montes e Alto Douro, na pessoa do Magnífico Reitor Professor Doutor António Fontaínhas Fernandes, por fornecer as instalações e equipamentos necessários à minha formação académica, bem como as condições para a realização desta dissertação de mestrado.

À Escola de Ciências da Vida e Ambiente, na pessoa do presidente Professor Doutor Artur Agostinho de Abreu e Sá, igualmente pela disponibilização das condições necessárias à realização deste trabalho.

Ao Departamento de Genética e Biotecnologia, na pessoa da Professora Doutora Maria Manuela do Outeiro Correia de Matos, por todo o trabalho desenvolvido em prol dos alunos, vital para a nossa formação.

À Diretora de Curso, na pessoa da Professora Doutora Raquel Maria Garcia dos Santos Chaves por tudo o que sempre fez e fará pelos alunos em geral, para que a formação nunca fosse unidimensional, mas sim rica e desafiante. Quanto a mim em particular, segue o agradecimento posteriormente.

À Vice-diretora de Curso, Professora Doutora Paula Filomena Martins Lopes, por ter sido ininterruptamente uma professora presente, primando sempre por nós, GMCT, enquanto futuros investigadores. É com certeza uma professora a nunca esquecer.

À Professora Doutora Raquel Maria Garcia dos Santos Chaves, minha orientadora, por ter confiado em mim e por ter permitido o desenvolvimento deste trabalho, mas, talvez mais importante ainda, por ter contribuído para o meu desenvolvimento pessoal. Obrigada pelo seu constante entusiasmo, pelo seu "brainstorming" e por tudo o que me ensinou sendo quem é. Obrigada por me incutir a paixão pelo conhecimento *per si*, não estaria melhor integrada noutro lugar.

À Professora Doutora Margarida Henriques da Gama Carvalho, minha co-orientadora, por me aceitar sem reversas no seu espaço e trabalho. Obrigada pela vontade em fazer ciência, obrigada por sempre me ter feito sentir incluída.

À Professora Doutora Maria Filomena Lopes Adega, por todo o apoio durante todo o tempo em que tive o prazer de trabalhar perto de si. Obrigada por tudo o fez e por reservar sempre um carinho especial para os seus alunos (saiba que é recíproco).

À Doutora Sandra Louzada por ter confiado em mim para uma próxima etapa, vou tentar estar à altura e aprender muito.

À Escudeiro, porque ela foi tudo o que podia pedir como mentora. Obrigada, do fundo do coração. A maneira como olhas para a vida faz qualquer um querer superar-se. Obrigada por todo este ano que se resumiu, metaforicamente, na tua ajuda constante para que eu deixasse de usar "bumpers" e começasse a encarar a caixa. Obrigada pela tua amizade e por tudo o que me ensinaste.

À Daniela, porque tem sempre uma palavra reconfortante e amiga, sem deixar de ser sincera e direta. Vais ser sempre alguém por quem terei um carinho especial, pessoal e profissionalmente. Obrigada por teres confiado em mim e pela tua capacidade inerente de te colocares no lugar do outro, ensinou-me muito também.

Aos meus amigos, de cá e lá, por todos os desabafos e cafés e por nunca deixarem de pertencer a quem sou. Obrigada à Telma, ao Mário, à Verónica e à Sabença pelos jogos de tabuleiro para desanuviar e por serem quem eu quero manter para a posteridade. Obrigada ao Odin por todos os abraços fofinhos e cheios de pelo que me fazem sorrir logo pela manhã. Obrigada ao Fi, primeiro porque me quis dar o Odin, e depois porque ninguém me ouviu mais do que ele durante todo este processo. Obrigada por me conheceres tão bem e por seres aquela constante inabalável.

Por fim, obrigada aos meus pais, por todos estes anos e por tudo o que sempre fizeram por mim. Dou muito valor a quem são e àquilo que me incutiram. Obrigada por acreditarem sempre em mim, mais do que ninguém. Obrigada por viverem para que eu seja a filha mais realizada do mundo, acreditem que são maravilhosos.

## Abstract

(Peri)centromeric repetitive sequences, and more specifically satellite DNA (SatDNA) sequences are a major heterochromatic genomic component, knowingly involved in Robertsonian translocations (ROBs). ROBs occur nonrandomly between acrocentric chromosomes (human chromosomes 13, 14, 15, 21 and 22). Presently, ROB mechanism and breakpoint location are still major issues, remaining not completely understood, essentially due to related assembly issues. The present work aims to inform about satellite arrangement (order and composition) in the context of the most common ROBs and, more precisely, Down-associated rob(14;21), by comparing up-to-date genome information and physical cytogenetic mapping with (peri)centromeric satellites (I, II, III, α and β).

Between human classical satellites and in opposition to the greatly approached α satellite (αSat), satellite I (SatI) has been overlooked in respect to its presence, organization and significance. Thought to locate at the (peri)centromeric regions of chromosomes 3, 4 and acrocentric chromosomes, this satellite could be involved in the breakpoint of rob(14;21), being noteworthy for the context of ROB formation. *In silico* analysis in the framework of SatI is presented, along with physical mapping, allowing to identify assembly gaps related with this satellite: *in silico* mapping still did not correspond to physical mapping results. Accordingly, SatI should be considered when addressing (peri)centromere-related ROBs, since the associated information gap is inevitable in the attempt to fully understand this rearrangement.

In addition, (peri)centromeric SatDNAs were physically and *in silico* mapped, in order to analyze their presence and organization. Physical maps for chromosomes 14, 21 and der(14;21) allowed us to infer about the etiological and mechanistic origin of ROBs and the possible chain of events leading to this rearrangement. Physical mapping information was compared with *in silico* analysis, leading to the recognition of a substantial number of assembly gaps in the human reference genome. Therefore, it is possible to acknowledge the preserved utility of physically mapping satellite probes. Furthermore, the present work also demonstrates the uneven representation of satellite families in general comparatively to αSat (greatly represented in HSA14 and HSA21).

The obtained results point out that the study of ROB mechanism and breakpoints still cannot exclusively rely on genomic technologies. Physical mapping should continue a prevailing player in the attempt to achieve accurate maps for pericentromeric/short-arm regions of acrocentric chromosomes. In order to entirely address these genomic regions with recent long-

read technologies, sequential mapping steps must be followed. This work provides a clear mapping basis and, complementarily, a full gathering of current genomic data.

**Keywords:** SatDNA; (Peri)centromere; Acrocentric chromosomes; Physical mapping; *In silico* mapping; ROB mechanism.

## Resumo

As sequências repetitivas de DNA (peri)centromérico e, mais especificamente as sequências de DNA satélite (SatDNA), correspondem a uma componente vital da heterocromatina, estando ativamente envolvidas na formação de translocações robertsonianas (ROBs). As ROBs ocorrem de uma forma não-aleatória entre cromossomas acrocêntricos (cromossomas humanos 13, 14, 15, 21 e 22). Atualmente, o mecanismo inerente à formação de ROBs e a localização dos pontos de quebra associados constituem questões problemáticas em termos de análise, permanecendo associadas a um grande desconhecimento, essencialmente devido a problemas no *assembly* de sequências repetitivas. O presente trabalho teve como objetivo informar sobre a ordem e composição das sequências de SatDNA nas ROBs mais comuns na espécie humana, mais precisamente na rob(14;21), comummente associada à síndrome de Down. A estratégia seguida consistiu em comparar dados genómicos de caráter recente e dados de mapeamento físico citogenético com sondas correspondentes aos satélites humanos I, II, III, α e β.

Em oposição a certas famílias de satélites humanos intensamente estudadas (como o satélite α), o satélite clássico I (SatI) tem sido terminantemente ignorado, no que diz respeito à sua presença, organização e significância. Pensa-se que o SatI se localiza nas regiões (peri)centroméricas dos cromossomas 3, 4 e de todos os cromossomas acrocêntricos humanos, podendo estar significantemente envolvido no ponto de quebra da rob(14;21) e, consequentemente, no contexto de formação desta ROB em específico. A análise *in silico* do SatI é apresentada em conjunto com o mapeamento físico, permitindo a identificação de *assembly gaps*, já que os dados *in silico* não corresponderam na integridade aos dados de mapeamento citogenético. Concordantemente, o SatI não deve ser desconsiderado no estudo de ROBs, efetivamente associadas a repetições (peri)centroméricas, pois a falta de informação associada a este satélite é inevitável na tentativa de analisar esta alteração cromossómica.

Adicionalmente, algumas famílias (peri)centroméricas de SatDNA foram mapeadas fisicamente e *in silico*. A obtenção de mapas físicos para os cromossomas 14, 21 e der(14;21) permitiu inferir sobre a origem etiológica e mecanística das ROBs e sobre a possível cadeia de eventos que culmina nesta translocação. A informação do mapeamento físico foi comparada com a análise *in silico* dos mesmos satélites, levando também ao reconhecimento de um elevado número de *assembly gaps* no genoma de referência presentemente disponível. Assim, torna-se imperativo reconhecer a utilidade preservada de proceder ao mapeamento físico de sondas de SatDNAs. Não obstante, o presente trabalho demonstrou a representação imparcial das famílias

de SatDNA em geral, comparativamente ao satélite α (grandemente representado nos cromossomas HSA14 e HSA21).

Sumariamente, os resultados obtidos apontam para o facto de que o estudo do mecanismo e pontos de quebra das ROBs ainda não pode depender inteiramente do contributo de tecnologias genómicas. O mapeamento físico deve permanecer como um fator primordial aquando da tentativa de obter mapas precisos da região (peri)centromérica e dos braços curtos dos cromossomas acrocêntricos. Com o intuito de abordar integralmente estas regiões genómicas recorrendo a tecnologias de sequenciação através de *reads* longos, é essencial atuar segundo uma abordagem sequencial. O trabalho aqui exposto fornece um mapeamento de caráter basal, dando acesso complementar a uma coleção completa dos dados genómicos mais recentes.

**Palavras-Chave:** SatDNA; (Peri)centrómero; Cromossomas acrocêntricos; Mapeamento físico; Mapeamento *in silico*; Mecanismo ROB.

## General Index

## Image Index

## Table Index

## List of Abbreviations

The abbreviations used throughout the text are listed below. No chemical formulas or symbols of physical and chemical quantities contained in the IUPAC (International Union of Pure and Applied Chemistry) system will be included, as there are international publications by reputable authorities which can be consulted.

**5_TAMRA -** 5-Carboxytetramethylrhodamine

**BAC** - Bacterial Artificial Chromosome

**bp -** base pair

**BLAST -** Basic Local Alignment Search Tool

**CENP** - Centromeric Protein

**CH** - Constitutive Heterochromatin

**DAPI -** 4,6-diamino-2-phenylindole dihydrochloride

**DMEM** - Dulbecco's Modified Eagle's Medium

**EMBOSS -** European Molecular Biology Open Software Suite

**FISH -** Fluorescent *in situ* hybridization

**FITC -** Fluorescein isothiocyanate

**HOR** - Higher Order Repeat

**HSA** - *Homo Sapiens*

**LINE** - Long Interspersed Nuclear Element

**NCBI** - National Center for Biotechnology Information

**NGS -** Next Generation Sequencing

**Non-LTR-retrotransposon** - Non-Long Terminal Repeat-Retrotransposon

**NOR** - Nucleolar Organizer Region

**PacBio** - Pacific Biosciences

**PCR -** Polymerase Chain Reaction

**ROB** - Robertsonian Translocation

**SatDNA** - Satellite DNA

**SINE** - Small Interspersed Nuclear Element

**SSC -** Saline Sodium Citrate

**SV -** Structural Variant

**TE** - Transposable Element

# Chapter I – General Introduction

## I.1. Robertsonian Translocations as a Chromosomal Phenomenon

In a historical point of view, Robertsonian translocations (ROBs) were first described by William Robertson in 1916, while studying grasshoppers (Robertson 1916). Today, our knowledge allows us to place them among the most frequent chromosome alterations in the human population. The early development of human cytogenetics brought the identification of these chromosomal alterations recurring to karyotype analysis (Wilch and Morton 2018). ROBs were indeed the first rearrangements to be reported in humans (Turpin et al. 1959) due to its common and easily identifiable character, especially in individuals with 45 chromosomes (Wilch and Morton 2018).

By definition, ROBs involve the fusion of two acrocentric chromosomes (human chromosome pairs 13, 14, 15, 21 and 22) with breakpoints nearby the centromeric region (Kolgeci et al. 2013). Accordingly, the junction of two acrocentric chromosomes to produce a single metacentric or submetacentric chromosome gives rise to balanced Robertsonian translocation carriers with 45 chromosomes, rather than the usual 46 present in humans. The occurrence of this type of alteration is typically associated with a balanced karyotype, as seen by the fact that most carriers are, at the current knowledge, considered phenotypically normal (Wilch and Morton 2018). During the fusion process, the short p-arms of the involved chromosomes fuse as well, however, the resulting rearrangement is lost in the first rounds of cell division, due to its acentric nature (i.e. lacks stability) (Kaiser-Rogers and Rao 2013). It is recognized that this loss produces no known impact (Morin et al. 2017) given the redundancy of the existing material (essentially, ribosomal RNA genes) (Kaiser-Rogers and Rao 2013). However, throughout the process of meiosis, the fused q-arms form specific structures and their homologues are identified as trivalents. By its turn, the formation of trivalents is responsible for the production of nullisomic or disomic gametes. Understandably, after fertilization, the obtained zygotes can be monosomic or trisomic (Morin et al. 2017). Therefore, only gametes resulting from alternate segregation (Figure I.1) are able to generate embryos with normal or balanced translocation carrier karyotypes (Jin et al. 2010).

ROBs occur more often between non-homologous chromosomes but can also occur between homologous chromosomes, although more rarely (Scriven et al. 2001). ROBs in general occur in the human species with an incidence close to 1 in 1000 individuals (Nielsen and Wohlert 1991). The high observed frequency of ROBs may be associated with NORs (Nucleolar

Organizer Regions) (Page et al. 1996), present at band p12, where each acrocentric chromosome contains ribosomal RNA genes (rDNA) (Ferguson-Smith 1964; Henderson et al. 1972; Evans et al. 1974; Schmickel and Knoller 1977). The propensity of acrocentric chromosome to form translocations involving the centromere and short arms may be an outcome of the favored association of NORs during nucleoli formation (Ferguson-Smith et al. 1961; Ohno et al. 1961; Ferguson-Smith and Handmaker 1963; Ferguson-Smith 1967).

The participation of particular acrocentric chromosomes in ROB formation is considered to be nonrandom. In that basis, ROBs can be classified, according to frequency, in two different groups: common (includes rob(13;14) and rob(14;21)) and rare (includes the remaining nonhomologous ROBs) (Bandyopadhyay et al. 2002). While focusing our scope in the most commonly involved chromosomes, we found the translocation between chromosomes 13 and 14 (73% of all ROBs), usually associated with Patau syndrome. The second most frequent case (10% of the total) is the ROB between chromosomes 14 and 21 (Therman et al. 1989; Scriven et al. 2001). The latter chromosomal translocation is related with Down syndrome: 3% of the affected individuals are a consequence of a translocation event (almost entirely ROBs involving chromosome 21, like rob(14;21)) (Antonarakis and Group* 1991). The meiotic segregation patterns of rob(14;21) carriers explain the possible emergence of the syndrome (Figure I.1).

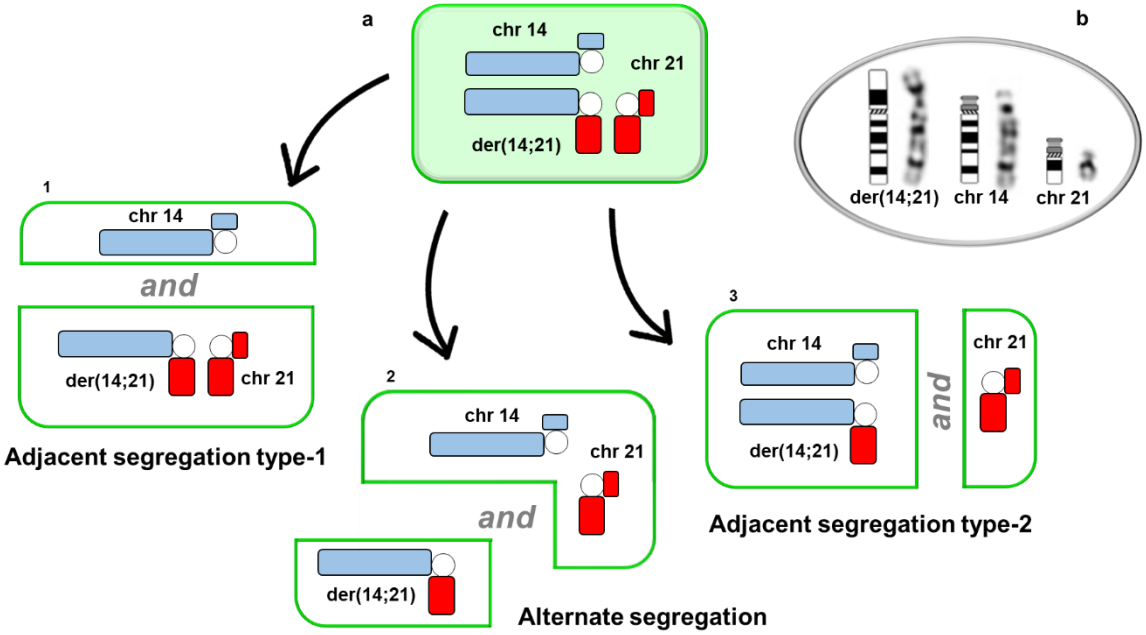



**Figure I.1. The gametogenesis of a rob(14;21) carrier in a simplistic depiction**. **(a)** - Chromosomes 14, 21 and der(14;21) form trivalents during meiosis and this results in 3 different segregation patterns (adjacent segregation type-1, type-2 and alternate segregation): **(a1/a3)** - Production of nullisomic or disomic gametes; **(a2)** - The resulting embryos can be carriers of the translocation or normal. **(b)** - Ideogram representation of chromosomes 14, 21 and der(14;21). Adapted from: Trevisan et al. 2014. G-banded chromosomes are also showed. Adapted from: Padilla et al. 2009.

The centric fusion inherent to the occurrence of a ROB presupposes the active involvement of centromeric sequences. As a matter of fact, repetitive sequences composing constitutive heterochromatin (CH) are considered 'hotspot' regions for chromosomal rearrangements (Chaves et al. 2004; Adega et al. 2006; Adega et al. 2009; Vieira-da-Silva et al. 2015; Escudeiro et al. 2019). The centromere is therefore considered a chromosomal location prompt to a high rate of rearrangements in several biological contexts from cancer to germline chromosomal mutations (Nakagawa and Okita 2019). In the case of ROBs, pericentromeric sequences assume a core role, being frequently involved in breakpoint formation (Gravholt et al. 1992).

### I.1.1. Close mechanistic relationship between centromeric sequences and ROBs

Structurally, the centromere is associated with chromosome segregation during mitosis and meiosis and with the consequent preservation of genomic stability. Centromere/kinetochore functional identity is assured by a complex of centromeric proteins (CENPs), including CENP-A, CENP-B and CENP-T (Aldrup-MacDonald and Sullivan 2014). Together with the presence of centromeric satellite arrays, human centromere is defined by the presence of the histone H3 variant CENP-A (McNulty et al. 2017), loaded into centromeric chromatin in each cell cycle (Foltz et al. 2009). Centromeric satellite DNAs (SatDNAs) have a highly conserved 17-bp-long CENP-B binding motif, which is essential to centromere formation (Masumoto et al. 1989; Ohzeki et al. 2002). Indeed, centromeric chromatin servs as the foundation for the recruitment of the mentioned centromeric proteins and, consequently for kinetochore establishment (McNulty et al. 2017; Musacchio and Desai 2017). The known hallmark of an active centromeres (CENP-A nucleosomes) is defined by open chromatin areas in core centromeric regions. On the contrary, flanking pericentromeric sequences lack the presence of CENP-A, possibly giving this heterochromatic region a conformational or stabilizing role (Schueler and Sullivan 2006).

Understandably, when addressing ROBs as chromosomal rearrangements in a variety of animal species, SatDNAs housed at the (peri)centromere are major dynamic role-players associated with chromosomal instability (Chaves et al. 2000; Chaves et al. 2004; Adega et al. 2006; Paço et al. 2013; Vieira-da-Silva et al. 2015).

The mechanism behind the formation of a ROB has been linked to an event of recombination between homologous SatDNA sequences present in non-homologous chromosomes (Therman et al. 1989; Page et al. 1996), capable of clearing up the nonrandom participation of acrocentric

chromosomes (Denison et al. 2002). Nevertheless, the homology of SatDNAs might not be the sole contributing agent. Centric fusions are preceded by double-strand breaks (DSBs) that can be a result of CENP-B nicking activity, which attributes a possible function to this centromeric protein: leading to a recombination event and, subsequently, to ROB formation (Garagna et al. 2001; Escudeiro et al. 2019).

When two centromeres are placed in close proximity due to a chromosomal rearrangement (as in the case of ROBs), dicentric chromosomes can be created (McNulty and Sullivan 2017). Subsequently, with the purpose of stabilization, these ROB chromosomes undergo centromere inactivation (Niebuhr 1972; Sullivan et al. 1994; Sullivan and Schwartz 1995; Bandyopadhyay et al. 2002; McNulty and Sullivan 2017) and behave afterwards as monocentric chromosomes (Sears and Camara 1952; Therman et al. 1986; Sullivan et al. 1994). The centromeric activity of ROB chromosomes appears to be hierarchically based on a concept similar to centromeric "strength", as some centromeres are more prone to remain active in the derivative chromosome (Sullivan et al. 1994). Succeeding, centromere activity may be evaluated epigenetically by CENP-A occupancy and the presence of centromeric SatDNAs with CENP-B binding boxes (Sullivan et al. 2011; Plohl et al. 2014; McKinley and Cheeseman 2016). Despite being the most frequent dicentric-stabilizing mechanism, centromere inactivation is not always the rule, since some chromosomes can indeed remain functionally dicentric (Sullivan and Schwartz 1995; Page and Shaffer 1998; Sullivan and Willard 1998; Higgins et al. 2005; Stimpson et al. 2010). One of the proposed explanations for this event is that shorter inter-centromeric distances may promote the maintenance of both active centromeres by minimizing the probability of improper chromosome segregation (Stimpson et al. 2012).

Still, the dicentric nature of human ROB chromosomes implies the presence of one active and one latent centromere, as well as the loss of β satellite/rDNA sequences (Hurley and Pathak 1977; Cheung et al. 1990; Gravholt et al. 1992; Wolff and Schwartz 1992; Page et al. 1996; Sullivan et al. 1996). Assuming the dicentric nature, the archetypal model of centric fusion may be acceptable: ROBs may result from crossing-over between satellite sequences (Therman et al. 1989; Cheung et al. 1990; Gravholt et al. 1992; Sullivan et al. 1996), followed by loss of p-arm fragments and reunion (Hurley and Pathak 1977; Stahl et al. 1983). A two-step model, where transitional SatDNA reorganization has to occur, may also be suitable for explaining ROB formation (Chaves et al. 2003; Escudeiro et al. 2019).

In addition to centromere inactivation, ROBs imply the use of mechanisms for DSB repair and the adjustment of CH content over time: a set of strategies used for sustaining chromosomes

viability (Chaves et al. 2003). In this process, (peri)centromeric sequences may be lost, which in some ways assigns a vital stabilizing task to SatDNAs in derived ROB chromosomes (Adega et al. 2009; Iannuzzi et al. 2009; Escudeiro et al. 2019).

## I.2. Centromeric DNA: The intriguing fraction of the genome

When, in 1968, Britten and Kohne revealed the high presence of repetitive sequences in eukaryotic genomes (Britten and Kohne 1968), a new research area was disclosed. Subsequently to this discovery, the biological significance of these sequences was unceasingly overlooked for many years to come, as the repetitive portion of the genome was unquestionably associated with no function (termed simply 'junk DNA'). Repetitive DNA sequences were soon categorized into a major classification related to repeat number, and progressively grouped according to their organization (arrays of tandem repeats or interspersed) (López-Flores and Garrido-Ramos 2012). Tandem repeats are characterized by the contiguous alignment of sequence units in a hierarchically organized manner, while interspersed repeats have a scattered distribution across the genome (McNulty and Sullivan 2018).

Seemingly, the term "heterochromatin" was associated with an inactive transcription nature since its advent in 1928 (essentially due to the related characteristic state of constant compaction), assumption from which repetitive sequences could not escape, being the major heterochromatic component (Podgornaya et al. 2018). Following the historical admission of the repetitive fraction of the genome, a new class of tandemly repeated DNA sequences was identified in the 1970s using cesium chloride density gradients: satellite DNA was born (Yasmineh and Yunis 1974). Fundamentally, SatDNAs constitute the eukaryotic centromeric and pericentromeric genomic regions (Jagannathan and Yamashita 2017), even though they can also be positioned at subtelomeric locations or even at interstitial regions (Henikoff and Dalal 2005; Plohl et al. 2012; Chaves et al. 2017). Despite the clear differences between satellite DNA sequences (specifically in the nucleotide sequence composition, complexity and/or abundance), sharing features can be observed: the capacity to form heterochromatic regions and the intrinsic aptitude to form long tandemly organized arrays (Plohl et al. 2012). Regardless of the apparent impossibility to attribute a straight and identifiable role to repetitive sequences (and therefore to satellite DNA), some studies began to emerge, essentially aiming to address these sequences in terms of functionality (Plohl et al. 2008; Biscotti et al. 2015; Ferreira et al. 2015; Ferreira et al. 2019). The new idea of a possible function arose from a parsimonious point of view: the genomic presence of repetitive sequences was probably not so meaningless (Plohl et al. 2012).

### I.2.1. Human centromere: Families of satellite DNA

Back in the end of the 1960s, and from the clear distinction of three human genomic DNA fractions in $CsSO_4$ gradients, it was feasible to identify and classify the correspondent classical satellite DNAs I, II and III. More precisely, each DNA fraction was known to be composed of a set of repetitive sequences with analogous buoyant densities (Lee et al. 1997). Nevertheless, a new classification was proposed in 1987 given the characteristic sequence heterogeneity of those DNA fractions – a prime family of simple repeats was identified for each fraction (Prosser et al. 1986). The three identified families were referred as satellite DNA families 1, 2 and 3 (accordingly to the enrichment in fractions I, II and III) (Lee et al. 1997).

Satellites I, II and III were first reported to be present in all acrocentric chromosomes, as well as in chromosomes 3 and 4 (Vissel et al. 1992). SatDNA I (SatI) was recognized by the presence of 42 bp repeats, consecutively arranged in constructs of 17 bp and 25 bp repeat units (Lee et al. 1997), that can tandemly arrange to form Higher Order Repeats (HORs) of 2.97 Kb (Kalitsis et al. 1993). Apparently, the amplification of the former sequence arrays arranged in a head-to-tail fashion resulted in the complexity of the SatI DNA family (Meyne et al. 1994). Interestingly, SatI is the most AT-rich fraction of the human genome, being also the less abundant classical satellite (Tagarro et al. 1994).

SatDNA II (SatII) was associated with a poorly conserved repeat unit (ATTCC) and SatDNA III (SatIII) was identified to be composed of pentameric repeats of the same motif (here well-conserved and interspersed with a specific 10 bp sequence) (Jeanpierre 1994; Lee et al. 1997). The inconsistent arrangement of Satellite II/III in complex repeats (in opposition to tandem repeats) has led to a poor characterization of these satellite families (Altemose et al. 2014). SatII and III probably arose from the same pentameric repeat (Prosser et al. 1986), yet today these sequences locate to different genomic regions (for example, a large array of Sat II is specifically present on chromosome 1) (Cooke and Hindley 1979; Altemose et al. 2014).

In 1975, a new discovered human satellite IV was also isolated and characterized (Gosden et al. 1975). However, the classical satellites (amongst the first human satellite families described (Enukashvily and Ponomartsev 2013) became the study focus in the field until the appearance of a new human satellite.

Originally, alpha (α) satellite DNA was isolated from a highly repetitive fraction present in the African green monkey genome (Maio 1971). Later on, α satellite repeats showed to be present in all human centromeres and to be composed of tandem repeats of an AT-rich 171 bp-long monomer. Some monomers within α satellite HORs have a 17 bp sequence motif called

the CENP-B box, recognizable by centromere protein CENP-B (Muro et al. 1992). CENP-B box location is closely associated to the HOR structure and varies depending on the chromosome-specific HOR (McNulty and Sullivan 2018). Alphoid monomers can form the named HORs, responsible for conferring chromosome specificity (Willard 1985; Ohzeki et al. 2002). Each human chromosome contains one or more exclusive α HOR array, except chromosomes 13/21 and 14/22, that share the same HOR (Devilee et al. 1986; Jørgensen et al. 1988; Trowell et al. 1993).

α satellite soon became the most intensively studied satellite DNA family, therefore representing, from thereafter, a model for the hierarchical HOR organization (Lee et al. 1997). Alphoid sequences are acutely related with centromere identity, as they are established as a prerequisite for kinetochore formation and the occurrence of active centromeres (Sullivan et al. 2017). It is possible to distinguish human centromeres based on their α HOR specificity-conferring composition, namely by the number and order of monomers (that share 50-70% of identity) (Sullivan et al. 2017). Through the designation of consensus α monomer sequences it is feasible to conceptualize suprachromosomal groups or families. The main suprachromosomal families (SF1-3) correspond to the kinetochore formation region and are associated to centromere functionality (McNulty and Sullivan 2018). Performing hybridization studies at high stringency allows to map individual HORs to specific chromosomes (Willard 1985), because of sequence polymorphisms found between them (McNulty and Sullivan 2018). At low stringency, subsets of HOR arrays co-hybridize, allowing to study the relation between suprachromosomal families (Waye and Willard 1987; Alexandrov et al. 1993). Beyond the occurrence of α HORs, α monomeric repeats are also present in transitional, array-adjacent regions, conceivably evolving non-homogeneously from homogenous HORs (Schueler et al. 2005; Shepelev et al. 2015). Core CENP-A-associated chromatin represents about 35% of α satellite sequences and the remaining repeats locate at the pericentromere (Nakagawa and Okita 2019).

In addition to α satellite and satellites I, II and III, we can also find gamma (γ) and beta (β) satellites between the diverse families of satellite DNAs (Plohl et al. 2014). γ satellite subfamilies (reported GSAT, GSATX and GSATII with ~60% identity) are GC-rich tandem pericentromeric repeats of a vastly diverged 220 bp monomer (Warburton et al. 2008; Kim et al. 2009) and have been identified in all human chromosomes (Kim et al. 2009) usually forming clusters of 2-10 kb (Lin et al. 1993). Kim et al. proposed that γ satellite repeats may possibly work as barriers for heterochromatin expansion to chromosomal arms, functionally similar to

genomic insulators. By its turn, β satellite consists also of HORs established by tandem arrays of a 68 bp monomer organized in multimeric HORs described to be present in all acrocentric chromosomes and chromosome 9 (Waye and Willard 1989; Choo 1997). Indeed, β satellite was distinguished in two different types of HORs (pB3 and pB4), composed of non-overlapping arrays with unlike genomic locations. pB3 is specifically localized in chromosome 9 and its representation is equivalent to 50-100 times per haploid genome. The second type of HOR, pB4, is 5 times more represented per haploid genome and locates in acrocentric chromosomes, where β satellite was early found to map distally and proximally to rDNA (Waye and Willard 1989).

In contrast to the centromeric ubiquitous presence of α satellite sequences, pericentromeric satellite families can significantly vary in abundance and chromosomal presence (Lee et al. 1997; Rudd et al. 2003), often leading to incongruences about overall existence and location in the human genome (Miga 2017). Table I.1 presents a summary of mentioned information about human satellite families.

**Table I.1. Summary of SatDNA families features, specifically repeat unit size, the possibility of forming HORs and the known chromosomal presence, as well as genome representativity**. *SatII presents large blocks on chromosomes 1 and 16. SatIII is widely represented on chromosome 9. Adapted from: Gosden et al. 1981; Waye and Willard 1989; Kalitsis et al. 1993; Lee et al. 1997; Levy et al. 2007; Kim et al. 2009; Hall et al. 2017; Miga 2017.

| SatDNA | Repeat unit size | HOR formation | Presence in chromosomes | Genome representativity |
|---|---|---|---|---|
| **αSat** | 171 bp | ✓ | All | 3-5% |
| **SatI** | 42 bp | ✓ | 3; 4; All acrocentric | 0,12% |
| **SatII** | 5 bp | - | All but chrs 6, 8 and 20***** | 1,5% (together w/ SatIII) |
| **SatIII** | 5 bp | - | Y; 1; 3-5; 7; **9***; 10; 13-18 ;20-22 | 1,5% (together w/ SatII) |
| **βSat** | 68 bp | ✓ | 9; All acrocentric | 0,02% |
| **γSat** | 220 bp | - | All | 0.13% |

Human centromeres are not only composed of satellite sequences, but also mobile elements (like LINEs and SINEs (Long/Small Interspersed Nuclear Element)), already described both in HOR arrays and monomeric repeats (Miga 2017). The centromeric region of human chromosomes is mostly composed of α HORs, eventually punctuated by Transposable Elements (TEs) (Miga 2015; Jain et al. 2018b). Nevertheless, the insertion of TEs in active HOR arrays is thought to be scarce, due to the binding of centromeric proteins (Schueler et al. 2001) (perhaps TEs presence is selected against to avoid centromere inactivation (Malik and Henikoff 2002)). Unlike the α-rich centromeric region, the pericentromeric chromosomal fraction is often interspaced with other satellite families like γ and SatIII (Figure I.2) (Plohl et al. 2014).

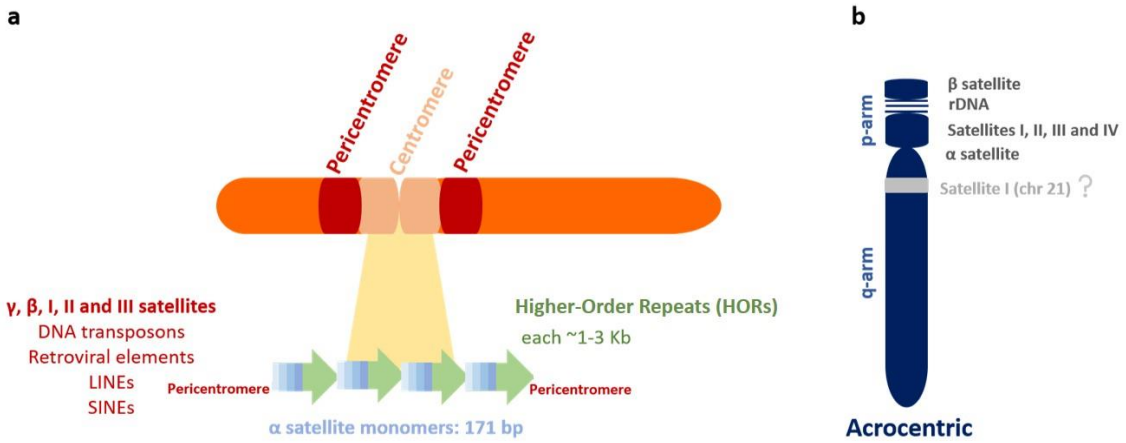



**Figure I.2. Schematic representation of human centromere organization. (a)** - Centromeric chromatin is mainly composed by α satellite HORs flanked by pericentromeric regions rich in other satellites, DNA transposons, retroviral elements, LINEs and SINEs **(b)** - Satellite organization in acrocentric chromosomes (known to date). Allegedly, satellites I-IV are found in p11 and form the α-adjacent pericentromeric region, followed by rDNA and β satellite repeats. The localization of satellite I is somewhat controversial, as some authors place it exclusively in the p-arm pericentric region (Kalitsis et al. 1993; Tagarro et al. 1994; Jarmuz-Szymczak et al. 2014). However, SatI was also described to locate post- α centromeric repeats (q-arm pericentromere) in chromosome 21 (Trowell et al. 1993). Adapted from: Trowell et al. 1993; Jarmuz-Szymczak et al. 2014; Buxton et al. 2017; Hall et al. 2017; Smurova and De Wulf 2018.

### I.2.1.1. Involvement of SatDNA in rob(14;21) formation

In 1996, Page et al. tried to disclose the mechanism of ROB formation, assuming two different protagonists: the upstream events leading to the translocation with a possible causative action (like satellite sequence homology) and the translocation itself, both of them clearly preponderates to assume a single model for the most frequent ROBs.

The p11 band of acrocentric chromosomes is known to be composed of several SatDNA families (detailed characterization above) often involved in ROBs. The translocation event between chromosomes 14 and 21 is deeply connected with sequence homology and consequent recombination (Page et al. 1996). However, and despite clear sequence similarity, the

organization in chromosome 14 seems to be reverse to the one found in chromosome 21 (Therman et al. 1986; Choo et al. 1988). Accordingly, the same situation is observable in rob(13;14), as chromosome 14 shares homologous inversely-oriented sequences with chromosomes 13 and 21. The nature of these sequences corroborates the statistic incidences of ROBs in the population, given that rob(13;14) and rob(14;21) are favored when compared with rob(13;21) (Jarmuz-Szymczak et al. 2014).

Specifically, in rob(14;21) formation, the breakpoint on chromosome 14 seems to localize between satellite III subfamilies pTRS-47 and pTRS-63 (Earle et al. 1992; Han et al. 1994). Furthermore, after the translocation, pTRS-47 appears to be retained and pTRS-63 seems to be lost (Page et al. 1996). By its turn, the breakpoint at 21p11 is distally localized to satellite I pTRI-6 subfamily (Kalitsis et al. 1993), more exactly between pTRI-6 and rDNA (Han et al. 1994; Page et al. 1996). pTRI-6 seems to be maintained on the derivative chromosome (Page et al. 1996) (Figure I.3). When the ROB involves chromosome 13, the breakpoint is comparable to the one on chromosome 21, as it is also thought to occur between pTRI-6 sequences and rDNA (rDNA is lost and pTRI-6, close to α satellite on the original p13 arm, is maintained in the translocated chromosome) (Han et al. 1994; Page et al. 1996).

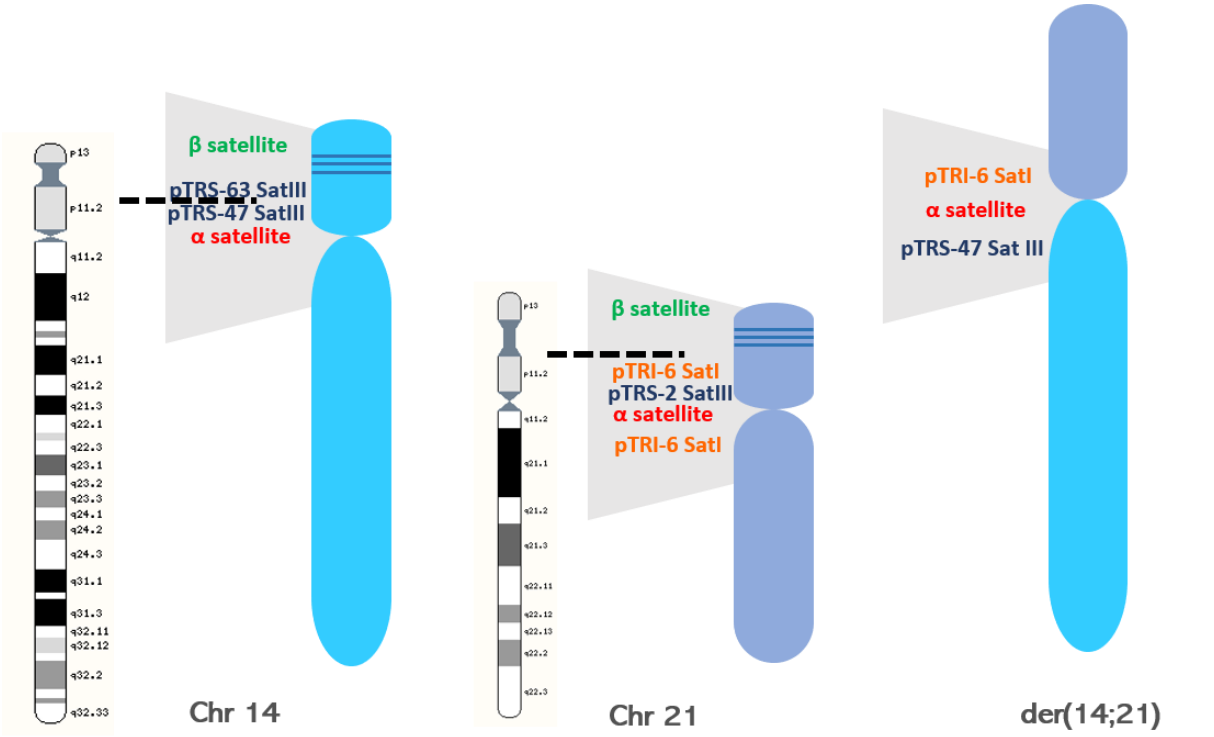



**Figure I.3**. **Breakpoint illustration for rob(14;21) and subsequent sequence organization to the best of current knowledge**. Representation of chromosome 14 and 21 with chromosomal bands (www.ensembl.org). Chromosome 14 is represented with the breakpoint between subfamilies pTRS-47 (adjacent to αSat) and pTRS-63, both in p11 chromosome band. Breakpoint at chromosome 21 is thought to localize distally to subfamily pTRI-6 and close to rDNA. pTRI-6 also shows to hybridize to the q-arm-proximal pericentromeric region. pTRS-2 constitutes a SatIII probe that showed to be present in the short arm of chromosome 21. After the translocation, the der(14;21) displays the presence of some SatDNA sequences, while others are confirmed to be lost. Both initially present in chromosome 14, pTRS-63 is maintained while pTRS-47 is absent. pTRI-6 is retained in the derivative chromosome. At the same time, β satellite repeats and ribosomal genes are lost during the translocation event. This visual depiction represents a compilation of the information currently available about SatDNA relation to breakpoint location Adapted from: Choo et al. 1992; Earle et al. 1992; Gravholt et al. 1992; Kalitsis et al. 1993; Trowell et al. 1993; Han et al. 1994; Tagarro et al. 1994; Page et al. 1996; Lee et al. 1997; Bandyopadhyay et al. 2002; Jarmuz-Szymczak et al. 2014.

Knowingly, translocation events may result from DSBs, as their deviant repair can lead to chromosomal rearrangements (Richardson et al. 1998). A reasonable explanation can be the formation of DNA DSBs from a stalled or damaged replication fork (Constantinou et al. 2001). Satellite III, for example, being a repetitive sequence involved in rob(14;21) and other ROBs, might form uncommon DNA structures that give rise to replication fork arrest (Akgun et al. 1997), consequently forming DSBs (Bandyopadhyay et al. 2002).

While studying ROBs, disclosing breakpoint exact location is a problematic task, given the poor characterization of acrocentric centromeres and short arms (Jarmuz-Szymczak et al. 2014). Nevertheless, the named location seems to be consistent between ROBs involving the same chromosomes (Page et al. 1996).

## I.3. Genomic tackling of satellite DNA

Highly repetitive satellite DNA undoubtedly represents a major gap in current human genome assemblies, significantly contributing to the lack of high-resolution sequencing studies in the field of centromere genomics. Availability of computer software algorithms for sequence analysis has been limited to methods excluding repetitive sequences and disregarding their annotation (Li 2014; Miga 2015). This is also the case of ROB breakpoint detection, since the homogenous nature of the involved sequences results in huge gaps when trying to standardly assemble pericentromeric regions or the short arms of acrocentric chromosomes (Eichler et al. 2004; Rudd and Willard 2004).

In order to annotate sequencing reads, the obtained data should be compared to previously existing reference assemblies. Notwithstanding, reads with repetitive nature and, therefore, mapping to multiple locations are overlooked. Additionally, repetitive genomic regions are often longer than the obtained reads from different sequencing technologies (Nishibuchi and Déjardin 2017). NGS (Next-Generation Sequencing) technologies (such as Illumina sequencing) and preceding ones (like Sanger sequencing) are mechanistically associated with the attainment of short reads, which increases the difficulty of unravelling complex repetitive sequences (Cao et al. 2017). These problems cause misalignments and misassemblies (Ameur et al. 2018) with a high number of contigs, assessing untraceable genomic positions to the analyzed repeats (Figure I.4) (Cao et al. 2017). It is a known fact that satellite repeats are greatly represented in assembly pools, but the exact determination of their location in linear stretches within centromeric regions becomes impossible (McNulty and Sullivan 2018). Theoretically, the correct placement of centromeric repeats in a linear assembly requires the presence of unique sequencing information that cannot actually represent sequencing errors and is often absent from homogenized satellite arrays (Luce et al. 2006; Miller et al. 2010; Li et al. 2012). As a possible alternative to linear assembly, centromeric reads can be represented in a graphic mode that distinguishes satellite families as nodes in a circle, allowing for multiple exact copies to be mapped to the same location (multiple similar repeats function as a single representative element) (Novak et al. 2015; Miga 2017). Then, probabilistic reversal of circular graphs allows the prediction of a linear assembly (Figure I.4) (Miga et al. 2014). However, the named graphic representation is not ideal for organizing entire, often specific, HOR units in a chromosome, therefore not allowing a long-range organization of a satellite array (McNulty and Sullivan 2018).

The uprising of long read technologies (PacBio (Pacific Biosciences) or Oxford Nanopore Sequencing) allowed to surpass some limitations of short reads, namely the profiling of tandem repetitive sequences (Harris et al. 2018). Despite enabling highly accurate genotyping in non-repetitive genomic regions, technologies like short-read Illumina sequencing do not deliver *de novo* genome assemblies, limiting the reconstruction of repetitive sequences (Jain et al. 2018a). Accordingly, with the advance of read length, sequencing interrogation methods can more accurately evaluate the size of repeated monomers in satellite sequences (Cacheux et al. 2016).

### I.3.1. The deep analysis of repetitive DNA content using nanopore sequencing

Early sequencing projects paved the way for sequencing evolution, bringing progressively larger genome assemblies: first, the bacteriophage Phi-X174 (Sanger et al. 1977a), followed by a 1Mb-sized bacterial genome and the fruit fly genome (120 Mb) (Adams et al. 2000), culminating in the first draft of the human genome (3 Gb) (Consortium 2001; Venter et al. 2001). These events resulted from the use of dideoxy or enzymatic chain termination method, typically known as Sanger sequencing (Sanger et al. 1977b). The later development of NGS, such as Illumina sequencing (Bentley et al. 2008) brought about a sudden transition from Sanger sequencing to NGS methods, which demanded the rethinking of assembly methods. Short-read NGS generates a low-cost high-efficiency sequence analysis, expanding the number of sequenced genomes, yet significantly reducing contig and scaffold sizes. Today, the major scaffold lengths arise from the newer, advanced long-read technologies, like nanopore sequencing, as it becomes possible to produce reads orders of scale larger than short Illumina reads (Phillippy 2017). Additionally, short-read technologies include PCR (Polymerase Chain Reaction) amplification before sequencing, which can result in a biased sequence representation, namely in sequences with extreme GC content. By its turn, long-read sequencing methods are often PCR-free library preparation techniques (Carneiro et al. 2012; Buermans and Den Dunnen 2014; Ameur et al. 2018; Lower et al. 2018).

The mechanism behind nanopore sequencing makes use of nanopores embedded in a lipid membrane or in a solid-state film, across which a defined voltage is applied causing the formation of an ionic current, subsequently interrupted by the passage of single nucleotide bases. Different nucleotides produce different patterns of ionic flow and the current changes allow the precise designation of bases passing in real-time through the pore (Feng et al. 2015; Eisenstein 2017; Leggett and Clark 2017) (each base has a unique ionic current profile)

(Rosenstein 2014). The constant progression of nanopore chemistries has allowed to reduce the error rate and to increase accuracy and throughput rates (Brown and Clarke 2016). These new schematics rely, for example, on different pores or better software approaches and intend the achievement of the accuracy necessary for complex genome assemblies (de Lannoy et al. 2017).

Clearly, the golden card for nanopore sequencing technology is the nature of read length, as limits are only introduced by sample preparation and DNA quality and, therefore, the obtained reads are theoretically unlimited in size (the larger reported go up to 1 Mb) (Jain et al. 2018a; van Dijk et al. 2018). The obtained ultra-long reads allow the high-resolution analysis of long stretches of repetitive sequences, opening up the path for highly repeated sequences, such as satellite DNA, to be more intensively studied (Figure I.4) (van Dijk et al. 2018). In the past year, Jain et al. achieved success in sequencing and assembling centromeric satellite DNA regions with nanopore sequencing reads with enough length to approach the Y's centromere (Jain et al. 2018b).

By offering the necessary informative sites, the obtainment of high-quality long reads can then be linked to a straightforward and exact displacement of overlap assembly methods applied to repetitive sequences. So, the final linear prediction is supported by experimental corroboration (Miga 2017). Thereby, nanopore sequencing presents itself as a forthcoming high-resolution step when analysing chromosomal translocations with other techniques like FISH (Fluorescent *in situ* hybridization), Southern blotting or PCR. It is clear that this technology constitutes a valuable instrument for resolving translocation breakpoints, especially if the breakpoints are located in repetitive regions, as in ROBs (Hu et al. 2018).

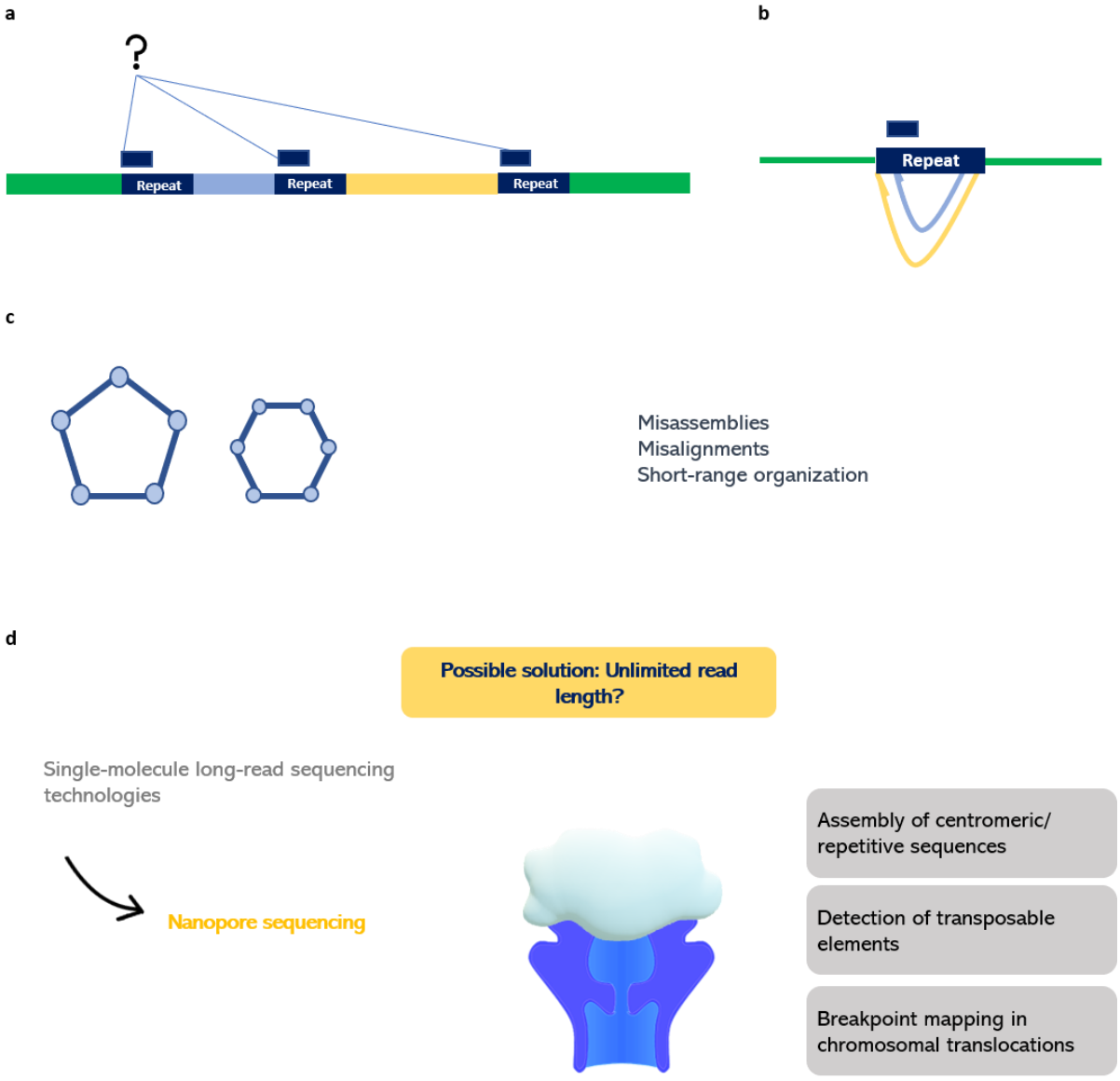



**Figure I.4. Comparison between short and long-read technologies, hereby represented by nanopore sequencing (alternative for the accurate assembly of repetitive sequences). (a) -** The graph represents the interrogation of three sequence repeats, in order to uncover the organization and presence of both the target sequences and possible adjacent genomic elements. **(b)** - With short-read alignments, the repeats represented in (a) are erroneously compiled into one uninterrupted repetitive region. **(c)** - The present circular graphs can be a possible approach for multi-mapping using short-read technologies. For example, when studying satellite families, each satellite HOR can be represented in a single graph element, while adjacent sequences are portraited in edges. Afterwards, circular graphs are converted in the most probable linear assembly. However, this method does not allow to obtain a long-range organization of satellite HOR arrays in a chromosomal level, as HOR specificity in the same satellite DNA family is not considered. Therefore, both linear and circular assemblies using short reads show difficulties, being inevitably associated with misassemblies, misalignments and a short-range limited organization. Adapted from: Miga 2015; Miga 2017. **(d)** - The unrestricted read length offered by single-molecule technologies, such as nanopore sequencing, can be a possible answer for unravelling repetitive sequences, as already demonstrated in several applications: assembly of centromeric/ repetitive sequences (Jain et al. 2018b), detection of transposable elements (Debladis et al. 2017) and breakpoint mapping in chromosomal translocations (Dutta et al. 2018; Hu et al. 2018).

## I.4. Work Aims

Throughout this work we try to address genomic and mechanistic issues related with the significance of repetitive DNA in ROBs (specifically rob(14;21)), essentially recurring to physical mapping and *in silico* approaches. Gathering the most information about SatDNA families present at the (peri)centromere/ acrocentric short-arms revealed to be necessary. So, the second chapter specifically tackles satellite I family (associated with information gaps), while the third chapter presents an integrative mapping methodology using satellite probes to physically map chromosomes 14, 21 and der(14;21) and to assess satellite representation in the currently available human reference genome. Therefore, the main work aims proceeded as followed:

- Isolate and clone human satellite I DNA sequences and assess clone features, like similarity and period size;
- Characterize SatI as a relevant co-player in ROB biological context, by mapping SatI clones probes onto chromosome preparations bearing rob(14;21), concomitantly using HSA14 and HSA21 painting probes;
- Apply Geneious as an *in silico* method to map SatI in all human chromosomes (available in the human reference genome GRCh38.p13) and assess its representativity comparing to available bibliographic data;
- Analyze SatI hits and SatI hit flanking regions using Dfam software;
- Compare SatI physical mapping with *in silico* results and infer about SatI localization and/or organization;
- Physically map SatI, SatII, SatIII, αSat and βSat clones onto chromosomes preparations bearing rob(14;21), followed by the use of HSA14 and HSA21 painting probes;
- *In silico* map SatDNA (peri)centromeric sequences in human chromosomes 14 and 21 using Geneious software;
- Compare *in silico* results with the previous physical mapping information;
- With the obtained physical maps for chromosomes 14, 21 and der(14;21), try to disclose possible alternatives for ROB mechanism of formation.
- Fundamentally, provide information about (peri)centromeric and short-arm sequences involved in ROBs, to facilitate assembly procedures in the future deployment of sequencing long-read methods in the analysis of repetitive sequences.

## References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF. 2000. The genome sequence of Drosophila melanogaster. *Science* **287**(5461): 2185-2195.

Adega F, Chaves R, Guedes-Pinto H, Heslop-Harrison J. 2006. Physical organization of the 1.709 satellite IV DNA family in Bovini and Tragelaphini tribes of the Bovidae: sequence and chromosomal evolution. *Cytogenetic and Genome Research* **114**(2): 140-146.

Adega F, Guedes-Pinto H, Chaves R. 2009. Satellite DNA in the karyotype evolution of domestic animals–clinical considerations. *Cytogenetic and Genome Research* **126**(1-2): 12-20.

AIexandrov I, Medvedev L, Mashkova T, Kisselev L, Romanova L, Yurov Y. 1993. Definition of a new alpha satellite suprachromosomal family characterized by monomeric organization. *Nucleic Acids Research* **21**(9): 2209-2215.

Akgun E, Zahn J, Baumes S, Brown G, Liang F, Romanienko PJ, Lewis S, Jasin M. 1997. Palindrome resolution and recombination in the mammalian germ line. *Molecular and Cellular Biology* **17**(9): 5559-5570.

Aldrup-MacDonald M, Sullivan B. 2014. The Past, Present, and Future of Human Centromere Genomics. *Genes* **5**(1): 33-50.

Altemose N, Miga KH, Maggioni M, Willard HF. 2014. Genomic characterization of large heterochromatic gaps in the human genome assembly. *PLoS Computational Biology* **10**(5): e1003628.

Ameur A, Kloosterman WP, Hestand MS. 2018. Single-molecule sequencing: Towards clinical applications. *Trends in Biotechnology* **37**(1): 72-85.

Antonarakis SE, Group* DSC. 1991. Parental origin of the extra chromosome in trisomy 21 as indicated by analysis of DNA polymorphisms. *New England Journal of Medicine* **324**(13): 872-876.

Bandyopadhyay R, Heller A, Knox-DuBois C, McCaskill C, Berend SA, Page SL, Shaffer LG. 2002. Parental origin and timing of de novo Robertsonian translocation formation. *American Journal of Human Genetics* **71**(6): 1456-1462.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**(7218): 53.

Britten RJ, Kohne DE. 1968. Repeated sequences in DNA. *Science* **161**(3841): 529-540.

Brown CG, Clarke J. 2016. Nanopore development at Oxford Nanopore. *Nature Biotechnology* **34**(8): 810.

Buermans H, Den Dunnen J. 2014. Next generation sequencing technology: advances and applications. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* **1842**(10): 1932-1941.

Buxton KE, Kennedy-Darling J, Shortreed MR, Zaidan NZ, Olivier M, Scalf M, Sridharan R, Smith LM. 2017. Elucidating Protein-DNA Interactions in Human Alphoid Chromatin via Hybridization Capture and Mass Spectrometry. *Journal of Proteome Research* **16**(9): 3433-3442.

Cacheux L, Ponger L, Gerbault-Seureau M, Richard FA, Escudé C. 2016. Diversity and distribution of alpha satellite DNA in the genome of an Old World monkey: Cercopithecus solatus. *BMC Genomics* **17**(1): 916.

Cao MD, Nguyen SH, Ganesamoorthy D, Elliott AG, Cooper MA, Coin LJ. 2017. Scaffolding and completing genome assemblies in real-time with nanopore sequencing. *Nature Communications* **8**: 14515.

Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo MA. 2012. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics* **13**(1): 375.

Chaves R, Adega F, Heslop-Harrison J, Guedes-Pinto H, 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.

Chaves R, Ferreira D, Mendes-da-Silva A, Meles S, Adega F. 2017. FA-SAT Is an Old Satellite DNA Frozen in Several Bilateria Genomes. *Genome Biology and Evolution* **9**(11): 3073-3087.

Chaves R, Heslop-Harrison J, Guedes-Pinto H. 2000. Centromeric heterochromatin in the cattle rob (1; 29) translocation: α-satellite I sequences, in-situ MspI digestion patterns, chromomycin staining and C-bands. *Chromosome Research* **8**(7): 621-626.

Chaves R, Santos S, Guedes-Pinto H. 2004. Comparative analysis (Hippotragini versus Caprini, Bovidae) of X-chromosome's constitutive heterochromatin by in situ restriction endonuclease digestion: X-chromosome constitutive heterochromatin evolution. *Genetica* **121**(3): 315-325.

Cheung S, Sun L, Featherstone T. 1990. Molecular cytogenetic evidence to characterize breakpoint regions in Robertsonian translocations. *Cytogenetic and Genome Research* **54**(3-4): 97-102.

Choo K, Earle E, Vissel B, Kalitsis P. 1992. A chromosome 14-specific human satellite III DNA subfamily that shows variable presence on different chromosomes 14. *American Journal of Human Genetics* **50**(4): 706.

Choo K, Vissel B, Brown R, Filby R, Earle E. 1988. Homologous alpha satellite sequences on human acrocentric chromosomes with selectivity for chromosomes 13, 14 and 21: implications for recombination between nonhomologues and Robertsonian translocations. *Nucleic Acids Research* **16**(4): 1273-1284.

Choo KA. 1997. *The centromere*, Vol 320. Oxford University Press Oxford.

Consortium IHGS. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860.

Constantinou A, Davies AA, West SC. 2001. Branch migration and Holliday junction resolution catalyzed by activities from mammalian cells. *Cell* **104**(2): 259-268.

Cooke HJ, Hindley J. 1979. Cloning of human satellite III DNA: different components are on different chromosomes. *Nucleic Acids Research* **6**(10): 3177-3198.

de Lannoy C, de Ridder D, Risse J. 2017. The long reads ahead: de novo genome assembly using the MinION. *F1000Research* **6**.

Debladis E, Llauro C, Carpentier M-C, Mirouze M, Panaud O. 2017. Detection of active transposable elements in Arabidopsis thaliana using Oxford Nanopore Sequencing technology. *BMC Genomics* **18**(1): 537.

Denison SR, Multani AS, Pathak S, Greenbaum IF. 2002. Fragility in the 14q21q translocation region. *Genetics and Molecular Biology* **25**(3): 271-276.

Devilee P, Cremer T, Slagboom P, Bakker E, Scholl HP, Hager H, Stevenson A, Cornelisse C, Pearson P. 1986. Two subsets of human alphoid repetitive DNA show distinct preferential localization in the pericentric regions of chromosomes 13, 18, and 21. *Cytogenetic and Genome Research* **41**(4): 193-201.

Dutta UR, Rao SN, Pidugu VK, Vineeth V, Bhattacherjee A, Bhowmik AD, Ramaswamy SK, Singh KG, Dalal A. 2018. Breakpoint mapping of a novel de novo translocation t (X; 20)(q11. 1; p13) by positional cloning and long read sequencing. *Genomics* **111**(5): 1108-1114.

Earle E, Shaffer L, Kalitsis P, McQuillan C, Dale S, Choo K. 1992. Identification of DNA sequences flanking the breakpoint of human t (14q21q) Robertsonian translocations. *American Journal of Human Genetics* **50**(4): 717.

Eichler EE, Clark RA, She X. 2004. An assessment of the sequence gaps: unfinished business in a finished human genome. *Nature Reviews Genetics* **5**(5): 345.

Eisenstein M. 2017. An ace in the hole for DNA sequencing. *Nature* **550**(7675): 285-288.

Enukashvily NI, Ponomartsev NV. 2013. Mammalian satellite DNA: a speaking dumb. *Advances in Protein Chemistry and Structural Biology* **90**: 31-65.

Escudeiro A, Ferreira D, Mendes-da-Silva A, Heslop-Harrison JS, Adega F, Chaves R. 2019. Bovine satellite DNAs – a history of the evolution of complexity and its impact in the Bovidae family. *The European Zoological Journal* **86**(1): 20-37.

Evans H, Buckland R, Pardue ML. 1974. Location of the genes coding for 18S and 28S ribosomal RNA in the human genome. *Chromosoma* **48**(4): 405-426.

Feng Y, Zhang Y, Ying C, Wang D, Du C. 2015. Nanopore-based fourth-generation DNA sequencing technology. *Genomics, Proteomics & Bioinformatics* **13**(1): 4-16.

Ferguson-Smith M. 1964. The sites of nucleolus formation in human pachytene chromosomes. *Cytogenetic and Genome Research* **3**(2-3): 124-134.

Ferguson-Smith M. 1967. Chromosomal satellite association. *The Lancet* **289**(7500): 1156-1157.

Ferguson-Smith M, Handmaker S, Hopkins AJ. 1961. Observations on the satellited human chromosomes. *The Lancet* **277**(7178): 638-640.

Ferguson-Smith M, Handmaker SD. 1963. The association of satellited chromosomes with specific chromosomal regions in cultured human somatic cells. *Annals of Human Genetics* **27**(2): 143-156.

Ferreira D, Escudeiro A, Adega F, Anjo SI, Manadas B, Chaves R. 2019. FA-SAT ncRNA interacts with PKM2 protein: depletion of this complex induces a switch from cell proliferation to apoptosis. *Cellular and Molecular Life Sciences*: 1-16.

Ferreira D, Meles S, Escudeiro A, Mendes-da-Silva A, Adega F, Chaves R. 2015. Satellite non-coding RNAs: the emerging players in cells, cellular pathways and cancer. *Chromosome Research* **23**(3): 479-493.

Foltz DR, Jansen LE, Bailey AO, Yates III JR, Bassett EA, Wood S, Black BE, Cleveland DW. 2009. Centromere-specific assembly of CENP-a nucleosomes is mediated by HJURP. *Cell* **137**(3): 472-484.

Garagna S, Marziliano N, Zuccotti M, Searle JB, Capanna E, Redi CA. 2001. Pericentromeric organization at the fusion point of mouse Robertsonian translocation chromosomes. *Proceedings of the National Academy of Sciences* **98**(1): 171-175.

Gosden J, Lawrie S, Gosden C. 1981. Satellite DNA sequences in the human acrocentric chromosomes: information from translocations and heteromorphisms. *American Journal of Human Genetics* **33**(2): 243.

Gosden J, Mitchell A, Buckland R, Clayton R, Evans H. 1975. The location of four human satellite DNAs on human chromosomes. *Experimental Cell Research* **92**(1): 148-158.

Gravholt CH, Friedrich U, Caprani M, Jørgensen AL. 1992. Breakpoints in Robertsonian translocations are localized to satellite III DNA by fluorescence in situ hybridization. *Genomics* **14**(4): 924-930.

Hall LL, Byron M, Carone DM, Whitfield TW, Pouliot GP, Fischer A, Jones P, Lawrence JB. 2017. Demethylated HSATII DNA and HSATII RNA foci sequester PRC1 and MeCP2 into cancer-specific nuclear bodies. *Cell Reports* **18**(12): 2943-2956.

Han J-Y, Choo K, Shaffer LG. 1994. Molecular cytogenetic characterization of 17 rob (13q14q) Robertsonian translocations by FISH, narrowing the region containing the breakpoints. *American Journal of Human Genetics* **55**(5): 960.

Harris RS, Cechova M, Makova K. 2018. Noise-Cancelling Repeat Finder: Uncovering tandem repeats in error-prone long-read sequencing data. *bioRxiv*: 475194.

Henderson A, Warburton D, Atwood K. 1972. Location of ribosomal DNA in the human chromosome complement. *Proceedings of the National Academy of Sciences* **69**(11): 3394-3398.

Henikoff S, Dalal Y. 2005. Centromeric chromatin: what makes it unique? *Current Opinion in Genetics & Development* **15**(2): 177-184.

Higgins AW, Gustashaw KM, Willard HF. 2005. Engineered human dicentric chromosomes show centromere plasticity. *Chromosome Research* **13**(8): 745-762.

Hu L, Liang F, Cheng D, Zhang Z, Yu G, Zha J, Wang Y, Wang F, Tan Y, Wang D. 2018. Localization of balanced chromosome translocation breakpoints by long-read sequencing on the Oxford Nanopore platform. *bioRxiv*: 419531.

Hurley JE, Pathak S. 1977. Elimination of nucleolus organizers in a case of 13/14 Robertsonian translocation. *Human Genetics* **35**(2): 169-173.

Iannuzzi L, King W, Di Berardino D. 2009. Chromosome evolution in domestic bovids as revealed by chromosome banding and FISH-mapping techniques. *Cytogenetic and Genome Research* **126**(1-2): 49-62.

Jagannathan M, Yamashita YM. 2017. Function of Junk: Pericentromeric Satellite DNA in Chromosome Maintenance. *Cold Spring Harbor Symposia on Quantitative Biology* **82**: 319-327.

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT. 2018a. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* **36**(4): 338.

Jain M, Olsen HE, Turner DJ, Stoddart D, Bulazel KV, Paten B, Haussler D, Willard HF, Akeson M, Miga KH. 2018b. Linear assembly of a human centromere on the Y chromosome. *Nature Biotechnology* **36**(4): 321-323.

Jarmuz-Szymczak M, Janiszewska J, Szyfter K, Shaffer LG. 2014. Narrowing the localization of the region breakpoint in most frequent Robertsonian translocations. *Chromosome Research* **22**(4): 517-532.

Jeanpierre M. 1994. Human satellites 2 and 3. *Annales de Genetique* **37**(4): 163-171.

Jin H, Ping L, Jie Q, Ying L, Yongjian C. 2010. Translocation chromosome karyotypes of the Robertsonian translocation carriers' embryos. *Fertility and Sterility* **93**(4): 1061-1065.

Jørgensen AL, Kølvraa S, Jones C, Bak AL. 1988. A subfamily of alphoid repetitive DNA shared by the NOR-bearing human chromosomes 14 and 22. *Genomics* **3**(2): 100-109.

Kaiser-Rogers K, Rao KW. 2013. Structural Chromosome Rearrangements. In *The Principles of Clinical Cytogenetics*, pp. 139-174. Springer, New York, NY.

Kalitsis P, Earle E, Vissel B, Shaffer LG, Choo KA. 1993. A chromosome 13-specific human satellite I DNA subfamily with minor presence on chromosome 21: further studies on Robertsonian translocations. *Genomics* **16**(1): 104-112.

Kim J-H, Ebersole T, Kouprina N, Noskov VN, Ohzeki J-I, Masumoto H, Mravinac B, Sullivan BA, Pavlicek A, Dovat S. 2009. Human gamma-satellite DNA maintains open chromatin structure and protects a transgene from epigenetic silencing. *Genome Research* **19**(4): 533-544.

Kolgeci S, Kolgeci J, Azemi M, Shala R, Daka A, Sopjani M. 2013. Reproductive Risk of the Silent Carrier of Robertsonian Translocation. *Medical Archives* **67**(1): 56.

Lee C, Wevrick R, Fisher RB, Ferguson-Smith MA, Lin CC. 1997. Human centromeric DNAs. *Human Genetics* **100**(3-4): 291-304.

Leggett RM, Clark MD. 2017. A world of opportunities with nanopore sequencing. *Journal of Experimental Botany* **68**(20): 5419-5429.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G. 2007. The diploid genome sequence of an individual human. *PLoS Biology* **5**(10): e254.

Li H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**(20): 2843-2851.

Li Z, Chen Y, Mu D, Yuan J, Shi Y, Zhang H, Gan J, Li N, Hu X, Liu B. 2012. Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Briefings in Functional Genomics* **11**(1): 25-37.

Lin C, Sasi R, Fan Y-S. 1993. Isolation and identification of a novel tandemly repeated DNA sequence in the centromeric region of human chromosome 8. *Chromosoma* **102**(5): 333-339.

López-Flores I, Garrido-Ramos M. 2012. The repetitive DNA content of eukaryotic genomes. In *Repetitive DNA*, Vol 7, pp. 1-28. Karger Publishers.

Lower SS, McGurk MP, Clark AG, Barbash DA. 2018. Satellite DNA evolution: old ideas, new approaches. *Current Opinion in Genetics & Development* **49**: 70-78.

Luce AC, Sharma A, Mollere OS, Wolfgruber TK, Nagaki K, Jiang J, Presting GG, Dawe RK. 2006. Precise centromere mapping using a combination of repeat junction markers and chromatin immunoprecipitation–polymerase chain reaction. *Genetics* **174**(2): 1057-1061.

Maio JJ. 1971. DNA strand reassociation and polyribonucleotide binding in the African green monkey, Cercopithecus aethiops. *Journal of Molecular Biology* **56**(3): 579-595.

Malik HS, Henikoff S. 2002. Conflict begets complexity: the evolution of centromeres. *Current Opinion in Genetics & Development* **12**(6): 711-718.

Masumoto H, Masukata H, Muro Y, Nozaki N, 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.

McKinley KL, Cheeseman IM. 2016. The molecular basis for centromere identity and function. *Nature Reviews Molecular Cell Biology* **17**(1): 16.

McNulty SM, Sullivan BA. 2017. Centromere silencing mechanisms. In *Centromeres and Kinetochores*, pp. 233-255. Springer.

McNulty SM, Sullivan BA. 2018. Alpha satellite DNA biology: finding function in the recesses of the genome. *Chromosome Research* **26**(3): 115-138.

McNulty SM, Sullivan LL, Sullivan BA. 2017. Human Centromeres Produce Chromosome-Specific and Array-Specific Alpha Satellite Transcripts that Are Complexed with CENP-A and CENP-C. *Developmental Cell* **42**(3): 226-240 e226.

Meyne J, Goodwin EH, Moyzis RK. 1994. Chromosome localization and orientation of the simple sequence repeat of human satellite I DNA. *Chromosoma* **103**(2): 99-103.

Miga KH. 2015. Completing the human genome: the progress and challenge of satellite DNA assembly. *Chromosome Research* **23**(3): 421-426.

Miga KH. 2017. The Promises and Challenges of Genomic Studies of Human Centromeres. *Progress in Molecular and Subcellular Biology* **56**: 285-304.

Miga KH, Newton Y, Jain M, Altemose N, Willard HF, Kent WJ. 2014. Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Research* **24**(4): 697-707.

Miller JR, Koren S, Sutton G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics* **95**(6): 315-327.

Morin SJ, Eccles J, Iturriaga A, Zimmerman RS. 2017. Translocations, inversions and other chromosome rearrangements. *Fertility and Sterility* **107**(1): 19-26.

Muro Y, Masumoto H, Yoda K, Nozaki N, Ohashi M, Okazaki T. 1992. Centromere protein B assembles human centromeric alpha-satellite DNA at the 17-bp sequence, CENP-B box. *The Journal of Cell Biology* **116**(3): 585-596.

Musacchio A, Desai A. 2017. A molecular view of kinetochore assembly and function. *Biology* **6**(1): 5.

Nakagawa T, Okita AK. 2019. Transcriptional silencing of centromere repeats by heterochromatin safeguards chromosome integrity. *Current Genetics*: 1-10.

Niebuhr E. 1972. Dicentric and monocentric Robertsonian translocations in man. *Humangenetik* **16**(3): 217-226.

Nielsen J, Wohlert M. 1991. Chromosome abnormalities found among 34,910 newborn children: results from a 13-year incidence study in Arhus, Denmark. *Human Genetics* **87**(1): 81-83.

Nishibuchi G, Déjardin J. 2017. The molecular basis of the organization of repetitive DNA-containing constitutive heterochromatin in mammals. *Chromosome Research* **25**(1): 77-87.

Novak AM, Rosen Y, Haussler D, Paten B. 2015. Canonical, stable, general mapping using context schemes. *Bioinformatics* **31**(22): 3569-3576.

Ohno S, Trujillo J, Kaplan W, Kinosita R, Stenius C. 1961. Nucleolus-organisers in the causation of chromosomal anomalies in man. *The Lancet* **278**(7194): 123-126.

Ohzeki J, Nakano M, Okada T, 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.

Paço A, Chaves R, Vieira-da-Silva A, Adega F. 2013. The involvement of repetitive sequences in the remodelling of karyotypes: the Phodopus genomes (Rodentia, Cricetidae). *Micron* **46**: 27-34.

Padilla CD, Cutiongco-de la Paz EM, Chiong MAD, Charcos GS, Cadag NS. 2009. Translocation Down Syndrome among Filipinos and Its Implications on Genetic Counseling. *Acta Medica Philippina* **43**(1): 12-15.

Page SL, Shaffer LG. 1998. Chromosome stability is maintained by short intercentromeric distance in functionally dicentric human Robertsonian translocations. *Chromosome Research* **6**(2): 115-122.

Page SL, Shin J-C, Han J-Y, Andy Choo K, Shaffer LG. 1996. Breakpoint diversity illustrates distinct mechanisms for Robertsonian translocation formation. *Human Molecular Genetics* **5**(9): 1279-1288.

Phillippy AM. 2017. New advances in sequence assembly. Cold Spring Harbor Lab.

Plohl M, Luchetti A, Mestrovic N, Mantovani B. 2008. Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. *Gene* **409**(1-2): 72-82.

Plohl M, Mestrovic N, Mravinac B. 2014. Centromere identity from the DNA point of view. *Chromosoma* **123**(4): 313-325.

Plohl M, Meštrović N, Mravinac B. 2012. Satellite DNA evolution. In *Repetitive DNA*, Vol 7, pp. 126-152. Karger Publishers.

Podgornaya OI, Ostromyshenskii DI, Enukashvily NI. 2018. Who Needs This Junk, or Genomic Dark Matter. *Biochemistry Biokhimiia* **83**(4): 450-466.

Prosser J, Frommer M, Paul C, Vincent PC. 1986. Sequence relationships of three human satellite DNAs. *Journal of Molecular Biology* **187**(2): 145-155.

Richardson C, Moynahan ME, Jasin M. 1998. Double-strand break repair by interchromosomal recombination: suppression of chromosomal translocations. *Genes & Development* **12**(24): 3831-3842.

Robertson WRB. 1916. Chromosome studies. I. Taxonomic relationships shown in the chromosomes of Tettigidae and Acrididae: V-shaped chromosomes and their significance in Acrididae, Locustidae, and Gryllidae: chromosomes and variation. *Journal of Morphology* **27**(2): 179-331.

Rosenstein J. 2014. The Promise of Nanopore Technology: Nanopore DNA sequencing represents a fundamental change in the way that genomic information is read, with potentially big savings. *Ieee Pulse* **5**(4): 52-54.

Rudd M, Schueler M, Willard H. 2003. Sequence organization and functional annotation of human centromeres. In *Cold Spring Harbor Symposia on Quantitative Biology*, Vol 68, pp. 141-150. Cold Spring Harbor Laboratory Press.

Rudd MK, Willard HF. 2004. Analysis of the centromeric regions of the human genome assembly. *Trends in Genetics* **20**(11): 529-533.

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, Hutchison C, Slocombe PM, Smith M. 1977a. Nucleotide sequence of bacteriophage φX174 DNA. *Nature* **265**(5596): 687.

Sanger F, Nicklen S, Coulson AR. 1977b. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* **74**(12): 5463-5467.

Schmickel RD, Knoller M. 1977. Characterization and localization of the human genes for ribosomal ribonucleic acid. *Pediatric Research* **11**(8): 929.

Schueler MG, Dunn JM, Bird CP, Ross MT, Viggiano L, Rocchi M, Willard HF, Green ED, Program NCS. 2005. Progressive proximal expansion of the primate X chromosome centromere. *Proceedings of the National Academy of Sciences* **102**(30): 10563-10568.

Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF. 2001. Genomic and genetic definition of a functional human centromere. *Science* **294**(5540): 109-115.

Schueler MG, Sullivan BA. 2006. Structural and functional dynamics of human centromeric chromatin. *Annual Review of Genomics and Human Genetics* **7**: 301-313.

Scriven P, Flinter F, Braude P, Ogilvie CM. 2001. Robertsonian translocations—reproductive risks and indications for preimplantation genetic diagnosis. *Human Reproduction* **16**(11): 2267-2273.

Sears ER, Camara A. 1952. A transmissible dicentric chromosome. *Genetics* **37**(2): 125.

Shepelev V, Uralsky L, Alexandrov A, Yurov Y, Rogaev EI, Alexandrov I. 2015. Annotation of suprachromosomal families reveals uncommon types of alpha satellite organization in pericentromeric regions of hg38 human genome assembly. *Genomics Data* **5**: 139-146.

Smurova K, De Wulf P. 2018. Centromere and pericentromere transcription: roles and regulation… in sickness and in health. *Frontiers in Genetics* **9**.

Stahl A, Luciani J, Hartung M, Devictor M, Bergé-Lefranc J, Guichaoua M. 1983. Structural basis for Robertsonian translocations in man: association of ribosomal genes in the nucleolar fibrillar center in meiotic spermatocytes and oocytes. *Proceedings of the National Academy of Sciences* **80**(19): 5946-5950.

Stimpson KM, Matheny JE, Sullivan BA. 2012. Dicentric chromosomes: unique models to study centromere function and inactivation. *Chromosome Research* **20**(5): 595-605.

Stimpson KM, Song IY, Jauch A, Holtgreve-Grez H, Hayden KE, Bridger JM, Sullivan BA. 2010. Telomere disruption results in non-random formation of de novo dicentric

chromosomes involving acrocentric human chromosomes. *PLoS Genetics* **6**(8): e1001061.

Sullivan BA, Jenkins LS, Karson EM, Leana-Cox J, Schwartz S. 1996. Evidence for structural heterogeneity from molecular cytogenetic analysis of dicentric Robertsonian translocations. *American Journal of Human Genetics* **59**(1): 167.

Sullivan BA, Schwartz S. 1995. Identification of centromeric antigens in dicentric Robertsonian translocations: CENP-C and CENP-E are necessary components of functional centromeres. *Human Molecular Genetics* **4**(12): 2189-2197.

Sullivan BA, Willard HF. 1998. Stable dicentric X chromosomes with two functional centromeres. *Nature Genetics* **20**(3): 227.

Sullivan BA, Wolff DJ, Schwartz S. 1994. Analysis of centromeric activity in Robertsonian translocations: implications for a functional acrocentric hierarchy. *Chromosoma* **103**(7): 459-467.

Sullivan LL, Boivin CD, Mravinac B, Song IY, Sullivan BA. 2011. Genomic size of CENP-A domain is proportional to total alpha satellite array size at human centromeres and expands in cancer cells. *Chromosome Research* **19**(4): 457.

Sullivan LL, Chew K, Sullivan BA. 2017. α satellite DNA variation and function of the human centromere. *Nucleus* **8**(4): 331-339.

Tagarro I, Wiegant J, Raap AK, González-Aguilera JJ, Fernández-Peralta AM. 1994. Assignment of human satellite 1 DNA as revealed by fluorescent in situ hybridization with oligonucleotides. *Human Genetics* **93**(2): 125-128.

Therman E, Susman B, Denniston C. 1989. The nonrandom participation of human acrocentric chromosomes in Robertsonian translocations. *Annals of Human Genetics* **53**(1): 49-65.

Therman E, Trunca C, Kuhn EM, Sarto GE. 1986. Dicentric chromosomes and the inactivation of the centromere. *Human Genetics* **72**(3): 191-195.

Trevisan P, Rosa RFM, Koshiyama DB, Zen TD, Paskulin GA, Zen PRG. 2014. Congenital heart disease and chromossomopathies detected by the karyotype. *Revista Paulista de Pediatria* **32**(2): 262-271.

Trowell HE, Nagy A, Vissel B, Choo KA. 1993. Long-range analyses of the centromeric regions of human chromosomes 13, 14 and 21: identification of a narrow domain containing two key centromeric DNA elements. *Human Molecular Genetics* **2**(10): 1639-1649.

Turpin R, Lejeune J, Lafourcade J, Gautier M. 1959. [Chromosome aberrations & human diseases; multiple spinal abnormalities with 45 chromosomes]. *Comptes rendus hebdomadaires des seances de l'Academie des Sciences* **248**(25): 3636-3638.

van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. 2018. The third revolution in sequencing technology. *Trends in Genetics* **34**(9): 666-681.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA. 2001. The sequence of the human genome. *Science* **291**(5507): 1304-1351.

Vieira-da-Silva A, Louzada S, Adega F, Chaves R. 2015. A high-resolution comparative chromosome map of Cricetus cricetus and Peromyscus eremicus reveals the involvement of constitutive heterochromatin in breakpoint regions. *Cytogenetic and Genome Research* **145**(1): 59-67.

Vissel B, Nagy A, Choo K. 1992. A satellite III sequence shared by human chromosomes 13, 14, and 21 that is contiguous with α satellite DNA. *Cytogenetic and Genome Research* **61**(2): 81-86.

Warburton PE, Hasson D, Guillem F, Lescale C, Jin X, Abrusan G. 2008. Analysis of the largest tandemly repeated DNA families in the human genome. *BMC Genomics* **9**: 533.

Waye JS, Willard HF. 1987. Nucleotide sequence heterogeneity of alpha satellite repetitive DNA: a survey of alphoid sequences from different human chromosomes. *Nucleic Acids Research* **15**(18): 7549-7569.

Waye JS, Willard HF. 1989. Human beta satellite DNA: genomic organization and sequence definition of a class of highly repetitive tandem DNA. *Proceedings of the National Academy of Sciences* **86**(16): 6250-6254.

Wilch ES, Morton CC. 2018. Historical and Clinical Perspectives on Chromosomal Translocations. *Advances in Experimental Medicine and Biology* **1044**: 1-14.

Willard HF. 1985. Chromosome-specific organization of human alpha satellite DNA. *American Journal of Human Genetics* **37**(3): 524.

Wolff DJ, Schwartz S. 1992. Characterization of Robertsonian translocations by using fluorescence in situ hybridization. *American Journal of Human Genetics* **50**(1): 174.

Yasmineh W, Yunis J. 1974. Localization of repeated DNA sequences in CsCl gradients by hybridization with complementary RNA. *Experimental Cell Research* **88**(2): 340-344.

# Chapter II – Human Satellite I as a co-player in Robertsonian Translocations: From classical to forgotten

*(In preparation)*


## Abstract

Satellite DNA (SatDNA) sequences constitute a vital element of (peri)centromeres, being perchance associated with sequence sharing between non-homologous acrocentric chromosomes, and thus potentially informative in the framework of Robertsonian translocations (ROBs). Between human classical satellites and in contrary to the vastly approached centromeric α satellite, satellite I (SatI) has been unnoticed in respect to its presence, organization, and overall significance. Thought to locate at the pericentromeric regions of chromosomes 3, 4 and acrocentric chromosomes, this satellite could be involved in the breakpoint of the most common ROBs (rob(13;14) and rob(14;21)). Given the noteworthy association of rob(14;21) with Down syndrome, we hereby stand for the need of more intensive studies on related satellites sequences in chromosomes 14 and 21. Physical and *in silico* mapping approaches are presented to address SatI in organizational terms, also providing a contemporary line of knowledge of this once classified as classical satellite.


## II.1. Introduction

Heterochromatic (peri)centromeric regions of human chromosomes are rich in distinct classes of tandemly repeated sequences, arranged in several copies of a given array-composing repeat unit which is the base item in this system of genome organization (Warburton et al. 2008), as in satellite DNA sequences. These sequences were first acknowledged in the form of classical satellites I, II and III, that are easily distinguishable from the remaining genomic DNA with cesium chloride density gradients (Vissel et al. 1992; Choo 1997; Lee et al. 1997; Enukashvily and Ponomartsev 2013). The predominant location of classical satellite families is thought to be in short arms of acrocentric chromosomes and the q12 region of chromosomes 1, 9, 16 and Y. The verifiable sequence sharing between non-homologous acrocentric chromosomes allows for a more frequent interaction among chromosomal subsets, resulting in specific chromosomal alterations. The former statement has justified the need for a more complete molecular studying approach of centromeric/centromere-adjacent sequences, often

etiologically involved in rearrangements, such as Robertsonian translocations (Kalitsis et al. 1993).

In opposition to the intensively studied α satellite centromeric satellite DNA, other tandem repeats, such as classical satellites, have long been associated with unsatisfactory knowledge and classified as "poorly covered" (Warburton et al. 2008), both in terms of size and organization (Shiels et al. 1997). Recent advances in the field of genomics and the acquaintance of more information have also left the human (peri)centromeric region and its satellite DNAs out of the analytical scope (Miga 2019). Particularly, satellite I, represented by a 42 bp-repeat unit (Prosser et al. 1986) and located at the pericentromeres of chromosomes 3, 4, 14, 15, 21 and 22 (Meyne et al. 1994; Tagarro et al. 1994; Therkelsen et al. 1997) has been related with a "experimental gap" for the last 20 years, which clearly does not exclude its potential significance in a variety of biological contexts related with centromere organization. This stagnant overlooking situation can be conceivably related with the fact that this satellite family is the least abundant classical satellite, being also the most AT-rich sequence (≈72.4%) of the human genome (difficult tackling) (Tagarro et al. 1994).

Satellite family I is composed of alternatively arranged tandem repeats of two sequence types: A (17 bp-long) and B (25 bp-long), combined to form 42 bp repeat units (Lee et al. 1997). This classical satellite was first described using a probe (pTRI-6) that hybridizes with all acrocentric chromosomes at low stringency and only with chromosomes 13 and 21 at high stringency (Kalitsis et al. 1993; Trowell et al. 1993).

The pericentromeric presence of satellite family I in acrocentric chromosomes seems to be deeply connected with Robertsonian translocations (ROBs). For example, the sequence composition of chromosome 13 corroborates its involvement in ROBs with knowingly superior statistical incidence: the large presence of satellite I could determine its interaction with similar repeats from other acrocentric chromosomes during prophase of meiosis, providing the conditions for ROB formation (Tagarro et al. 1994). Similarly, the satellite composition of the pericentromeric regions of chromosome 21 is clinically pertinent in the case of trisomy 21, possibly arising from robertsonian translocations between acrocentric chromosomes with related satellite sequences. Hence, homology differences between short arm sequences dictate the frequency of the translocation. On that note, rob(14;21) is the most recurrent trisomy 21-related ROB (60%), followed by rob(21;21) and the rarer ones (5%) (Dey 2011) between chromosome 21 and other acrocentric chromosomes (13, for example). This may be explained by the assumption that homologous sequences at 14p are inversely positioned in relation to

sequences present at 21p and 13p (Choo et al. 1988; Therman et al. 1989; Shaffer 2002). The importance of satellite I in ROBs arises from the fact that this sequence might be involved in the breakpoint in chromosomes 13 and 21, thought to locate at 13p11 or 21p11, between SatI family pTRI-6 and the rDNA genes (Kalitsis et al. 1993).

By isolating, analyzing and mapping satellite I sequences, we intended to contribute for a better understanding of their relevant nature in the framework of ROBs, increasing the centromere resolution in general but also providing new insights into the mechanistic dynamics of this chromosomal rearrangement.

## II.2. Material and Methods

### Cell culture, chromosome preparation and genomic DNA isolation

The present study presupposed the comparative use of two commercially available human cell lines:

-GM03417, a mosaic holding the rob(14;21) (46,XX,der(14;21),+21/45,XX,der(14;21));

-GM12878, karyotypically normal and previously used as reference in the Human Genome Project. Both cell lines were maintained in DMEM medium supplemented with: 13% AmnioMax C-100 Basal Medium, 2% AmnioMax C-100 supplement, 15% FBS (Fetal Bovine Serum), 1% Glutamine and 1% of antibiotic mixture Penicillin (100 U/mL) / Streptomycin (100 µg/mL). All the reagents mentioned above are commercialized by Gibco, Thermo Fisher Scientific. Chromosome harvesting and chromosomal preparations were achieved recurring to routine procedures. Genomic DNA extraction with the commercial kit QuickGene DNA Tissue Kit S (Fujifilm Life Science) was achieved according to the manufacturer's instructions.

### Human Satellite I DNA isolation and cloning

Satellite I (SatI) was amplified by PCR (Polymerase Chain Reaction) of human genomic DNA with four set of specific designed primers. Primers were designed using the web-based interface Primer 3 (Rozen and Skaletsky 2000) and are described in Supplementary Table II.S1. PCR program was as following: initial denaturing step at 94ºC for 10 min; 30 cycles of 94ºC for 1 min (denaturation), 54ºC for 45 s (annealing) and 72ºC for 45 s (extension); final extension at 72ºC for 10 min. The annealing temperature was optimized for each set of primers. For LHSatI a similar PCR program was repeated with the annealing temperature of 57ºC. PCR

products were run in an agarose gel and the bands obtained from the amplification with JxHSatI and LHSatI primers were extracted. The first PCR program corresponded to 200 bp, 900 bp and 550 bp PCR bands (JxHSatI and LHSatI). The second PCR program corresponded to the equal obtainment and extraction of a 550 bp PCR band (LHSatI). Bands were purified using the QIAquick PCR Purification Kit (Qiagen). SatI PCR amplicons were then cloned using the vector pUC19DNA/SmaI, which requires the use of the Fast DNA End Repair (Thermo Fisher Scientific) to blunt and phosphorylate sequence ends for ligation to occur (sequences are ligated to SmaI site on pUC19 with T4 DNA ligase). Transformation was performed with DH5α competent bacterial cells (Invitrogen, Thermo Fisher Scientific). Colonies were selected with blue-white screening (β-galactosidase blue-white α complementation) and positives were confirmed by PCR. Positive clones were sequenced in StabVida by Sanger methodology in order to deeply analyze the isolated sequences and to assess clone similarity.

## Sequence analysis of SatI DNA clones

Multiple sequence alignments were obtained with ClustalW matrix Geneious R9 version 9.1.5 (Biomatters); parameters were set to default values. Sequence analysis was performed with BLAST (Basic Local Alignment Search Tool) from NCBI (National Center for Biotechnology Information) databases. Human chromosome (HSA) sequences (GRCh38.p13; assembly accession: GCA_000001405.28) were collected in FASTA format from NCBI. Satellite I repeat sequences (one illustrative of the AB 42 bp sequence and two representative obtained clones) were searched in human chromosomes using BLAST, with the following parameters: max_target_seqs was set to 10000 and word size to 11. BLAST hits were filtrated for scores ≥90 and e-values $\leq 10^{-16}$ (in the case of the SatI AB 42 bp, only alignments with e-values $>10^{-4}$ were discarded). Filtrated BLAST hits were mapped to human chromosomes (*Homo sapiens* reference genome GRCh38.p13) using Geneious software. Sequences of SatI DNA clones were also analyzed with Tandem Repeats Finder software (Benson 1999) and scanned for the presence of other repetitive elements in Repbase using the Censor software. Flanking regions of satellite I hits were screened for the presence of repetitive elements using the Dfam database software (Hubley et al. 2015). The (peri)centromeric region of HSA14 was analyzed in Geneious and Dfam to show a representation of satellite I organization. A dotplot of SatI was obtained with Geneious R9 based on the EMBOSS 6.5.7 dotmatcher software set to the following parameters: 10 for window size and 23 for threshold.

## DNA-Fluorescent *in situ* hybridization (DNA-FISH)

FISH was standardly performed (Heslop-Harrison and Schwarzacher 2011; Chaves et al. 2017), in order to physically map SatI clones onto human chromosomes. Human metaphases were sequentially hybridized with cloned sequences and painting probes for human chromosomes 14 and 21, obtained by chromosome sorting. In between hybridization protocols, slides were treated to eliminate previous hybridization signals. Clone probes were PCR labelled and painting probes were labelled by DOP-PCR, with digoxigenin-11-dUTP or biotin-16-dUTP (both from Roche Applied Science). DOP-PCR was performed with degenerated primer 6MW (Supplementary Table II.S1). Hybridization was performed over-night for clone probes and during approximately one week for painting probes. In both cases, post-hybridization washes were guaranteed with temperature (37ºC) and 50% formamide/2xSSC. Digoxigenin-labelled probes were detected with antidigoxigenin-5'-TAMRA (Sigma-Aldrich) and biotin-labelled probes were detected with FITC-conjugated avidin (Vector Laboratories). Preparations were mounted using Vectashield containing 4'-6-diamidino-2-phenylindole (DAPI) (Vector Laboratories) to counterstain chromosomes.

## Image capture and processing

FISH images were observed using a Zeiss ImagerZ microscope coupled to an Axiocam digital camera using AxioVision software (version Rel. 4.5, Zeiss). Digitized photos were prepared for printing in Adobe Photoshop (version 7.0).

## II.3. Results

### Satellite I isolation and analysis

In this work, human satellite I family was isolated by PCR, cloned and sequenced. Sequences of three sizes (≈ 200, 550 and 900 bp) were obtained with JxHSatI and LHSatI primers (Supplementary Table II.S1), resulting in a total of 83 clones with a high degree of similarity, essentially between 200 bp clones and between 900 bp clones (Figure II.1, Supplementary Figure II.S1, Supplementary Figure II.S2).

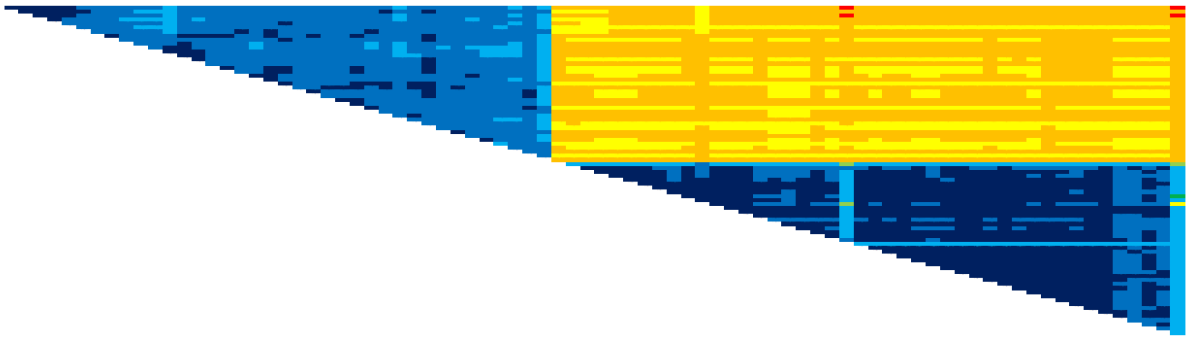


**Figure II.1. Distance matrix of the pairwise alignment of all SatI clones**. The distances were calculated using the alignment algorithm CLUSTALW, and the matrix was generated by Geneious R9 version 9.1.2 (Biomatters) under default settings. Cells showing nucleotide identities of : 98–100% (dark blue), 94–97.9% (medium blue), 90–93.9% (light blue), 88–89.9% (dark green), 84.1–87.9% (light green), 80–84% (yellow), 70–80% (orange), and<70% (red).

All satellite I clones were then analyzed using Tandem Repeats Finder. Consistently, all clones showed a 42 bp period size repeated to all sequence extent. Some 550/900 bp-sized clones also showed period sizes consisting in approximate multiples of the 42 bp repeat. All examined sequences demonstrated the high, satellite I-characteristic, AT content (in this case a medium of ≈ 77%) (Supplementary Table II.S2). Consensus sequences for satellite I clones are presented (Supplementary Table II.S2), showing elevated resemblance.

### Satellite I physical and *in silico* mapping

Isolated and cloned SatI sequences (representative 200, 550 and 900 bp clones) were physically mapped (by *in situ* hybridization) to human metaphases bearing rob(14;21) (Figure II.2, Supplementary Figure II.S3). All 3 clones showed to colocalize in a pericentromeric location (Supplementary Figure II.S3). A strong hybridization signal, corresponding to a large block of satellite I, is observed in a specific pair of acrocentric chromosomes (identified by reserve-DAPI to represent HSA13 chromosome pair). Human painting probes corresponding to chromosomes 14 and 21 (Figure II.2c) allowed to identify hybridization signals in chromosomes 14, 21 and rob(14;21) (Figure II.2).

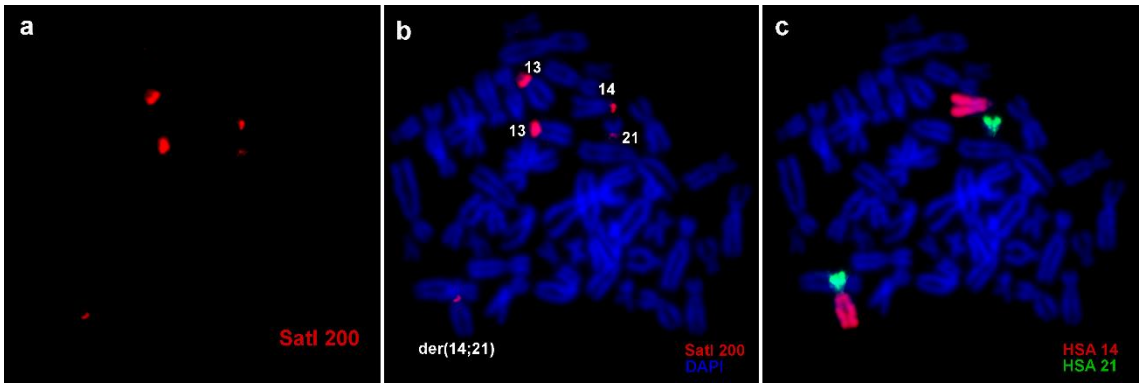



**Figure II.2. Physical mapping of a representative 200 bp satellite I clone in human metaphases bearing the rob(14;21). (a-b)** - A strong hybridization signal is present on chromosome pair 13 (pericentromeric region). Chromosomes 14, 21 and rob(14;21) also show hybridization signals in the (peri)centromere. **(c)** - The use of painting probes for human chromosomes 14 and 21 allows to identify both chromosomes and the derivative robertsonian chromosome. The name and color of each probe are indicated within each section. Chromosomes are counterstained with DAPI (blue). Digoxigenin-labelled probes (SatI 200 clone probe and HSA14 painting probe) were detected with antidigoxigenin-5'-TAMRA (red). Biotin-labelled HSA21 painting probe was detected with FITC-conjugated avidin (green).

A previously reported SatI repeat unit (AB, 42 bp) (Kalitsis et al. 1993; Lee et al. 1997) and two obtained SatI clones (200 bp and 900 bp, representative of both JX174276.1 and L01057.1 accessions) were submitted to NCBI BLAST tools against the current human genome assembly (*Homo sapiens* reference genome GRCh38.p13) and the obtained filtrated hits were mapped to human chromosomes using Geneious R9 software, placing a graphical representation of the current knowledge about satellite I location. *In silico* mapping showed a pericentromeric location of SatI BLAST hits in particular human chromosomes, namely HSA3, 4, 8, 14 and 22) (Supplementary Table II.S3, Figure II.3, II.4, II.5, II.6, II.7). Regions flanking SatI BLAST hits were submitted to Dfam database to scan the presence of repetitive elements. Only satellite I repeats from HSA8 showed to be flanked by transposable elements, namely Non-LTR-retrotransposons (Figure II.5). The analysis of the HSA14 (peri)centromeric region (with Geneious and Dfam) allowed to examine satellite I organization recurring to a dotplot representation and the graphical representation of current genome annotations for this family of satellite DNA (Figure II.8). Comparing to the Geneious analysis (placing SatI in chromosomes 3, 4, 8, 14 and 22), annotations from the Dfam database place this SatDNA in human chromosomes 4, 8, 14, 15 and 22 (Figure II.8).

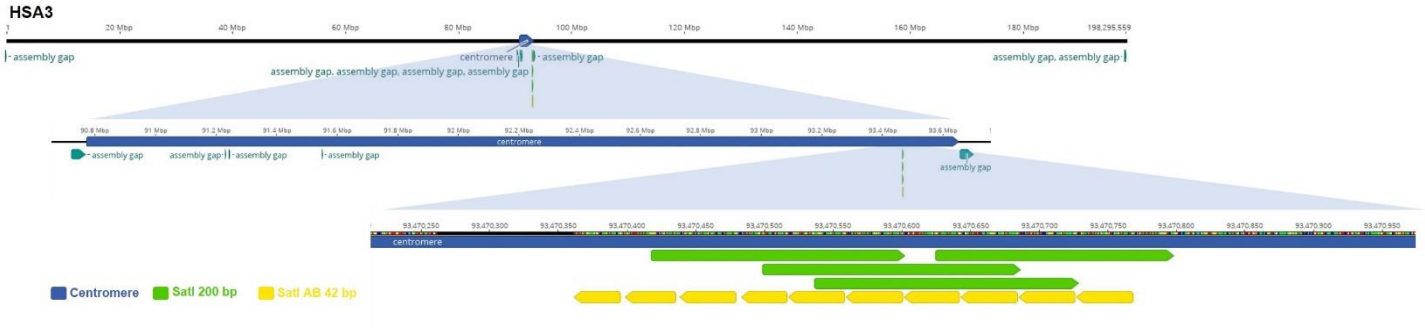



**Figure II.3. *In silico* mapping of SatI BLAST hits onto HSA3.** The whole length of human chromosome 3 is shown and zoom in to the q-arm-adjacent (peri)centromeric region. Assembly gaps in the current genome assembly (GRCh38.p13) are also depicted. Centromere is represented in deep blue. Annotations of SatI 200 bp and SatI AB 42 bp hits are shown in the bottom line (green and yellow, respectively).

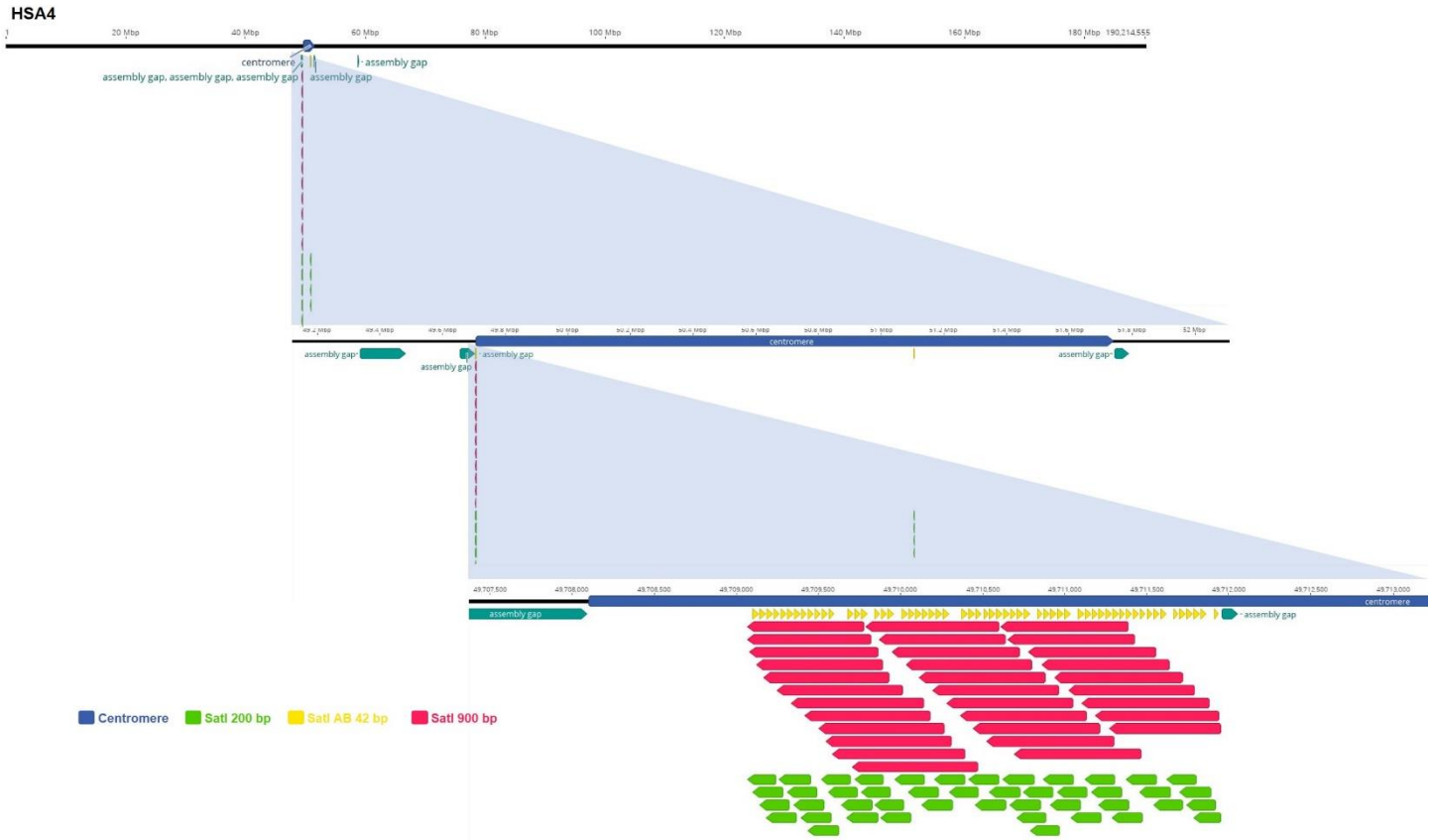



**Figure II.4. *In silico* mapping of SatI BLAST hits onto HSA4.** The whole length of human chromosome 4 is shown and zoom in to the p-arm-adjacent (peri)centromeric region. Assembly gaps in the current genome assembly (GRCh38.p13) are also depicted. Centromere is represented in deep blue. Annotations of SatI 200 bp, SatI 900 bp and SatI AB 42 bp hits are shown in the bottom line (green, pink and yellow, respectively).

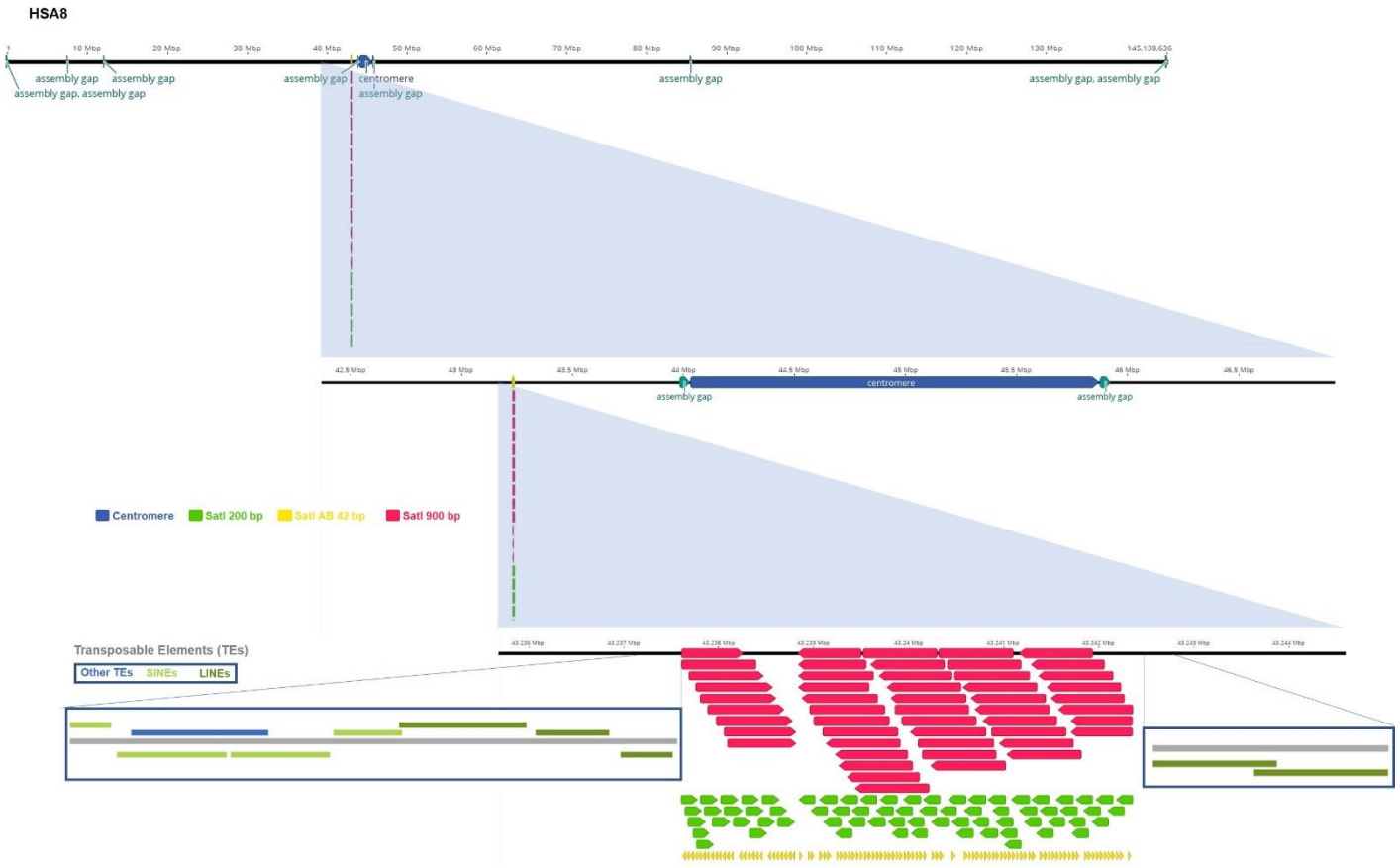



**Figure II.5. *In silico* mapping of SatI BLAST hits onto HSA8 and analysis of flanking regions.** The whole length of human chromosome 8 is shown and zoom in to the p-arm-adjacent (peri)centromeric region. Assembly gaps in the current genome assembly (GRCh38.p13) are also depicted. Centromere is represented in deep blue. Annotations of SatI 200 bp, SatI 900 bp and SatI AB 42 bp hits are shown in the bottom line (green, pink and yellow, respectively). SatI annotations are flanked by transposable elements (TEs) (in the bottom, Geneious and Dfam representation). Shown TEs are classified as non-LTR-retrotransposons, namely SINEs (light green) and LINEs (deep green). The TE in blue, classified as "Other", represents a composite retroelement (SVA: SINE+VNTR+Alu).

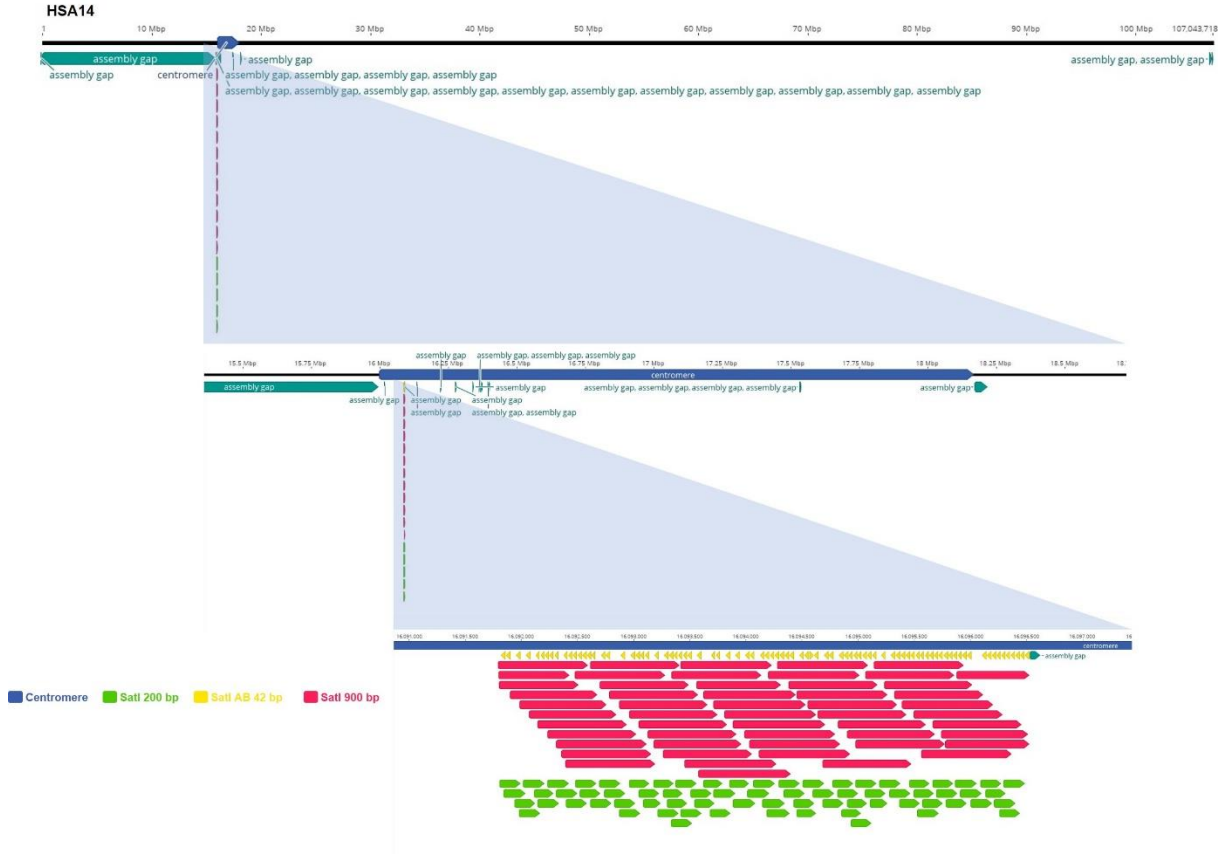



**Figure II.6. *In silico* mapping of SatI BLAST hits onto HSA14.** The whole length of human chromosome 14 is shown and zoom in to the p-arm-adjacent (peri)centromeric region. Assembly gaps in the current genome assembly (GRCh38.p13) are also depicted. Centromere is represented in deep blue. Annotations of SatI 200 bp, SatI 900 bp and SatI AB 42 bp hits are shown in the bottom line (green, pink and yellow, respectively).

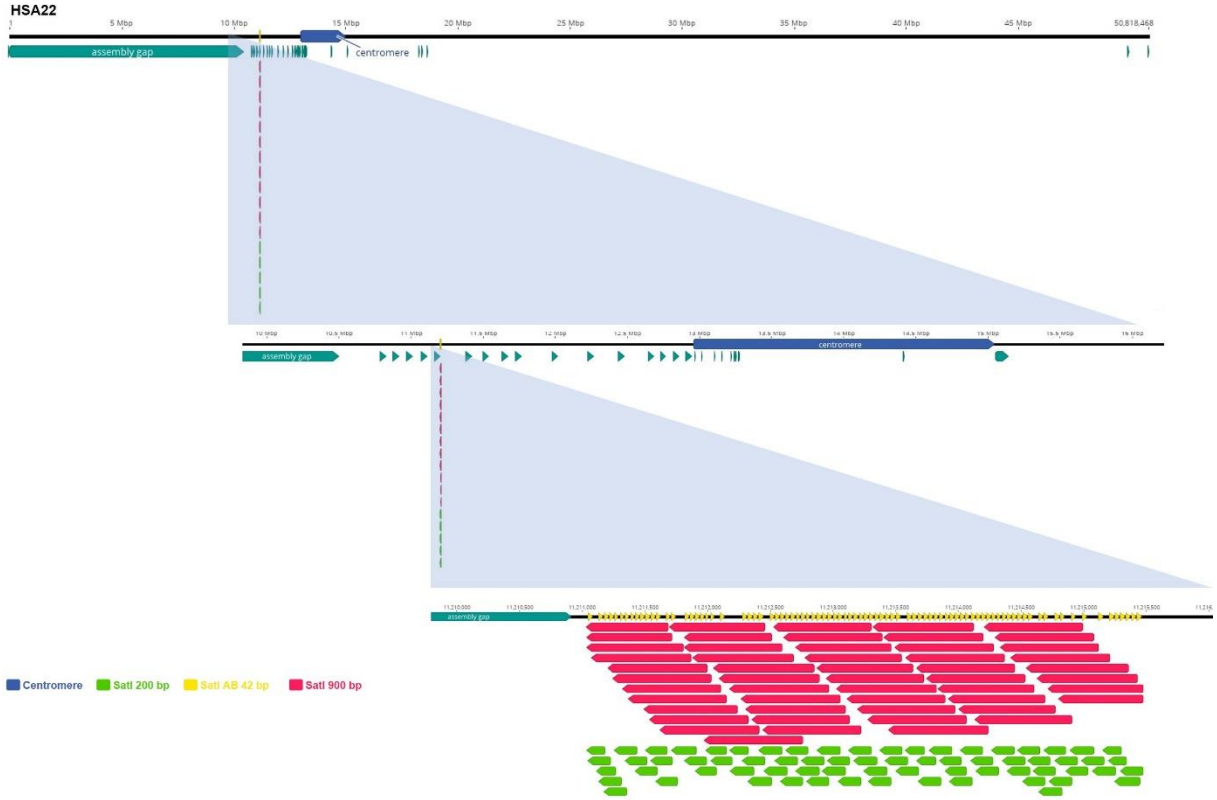



**Figure II.7. *In silico* mapping of SatI BLAST hits onto HSA22.** The whole length of human chromosome 22 is shown and zoom in to the p-arm (peri)centromeric region. Assembly gaps in the current genome assembly (GRCh38.p13) are also depicted. Centromere is represented in deep blue. Annotations of SatI 200 bp, SatI 900 bp and SatI AB 42 bp hits are shown in the bottom line (green, pink and yellow, respectively).

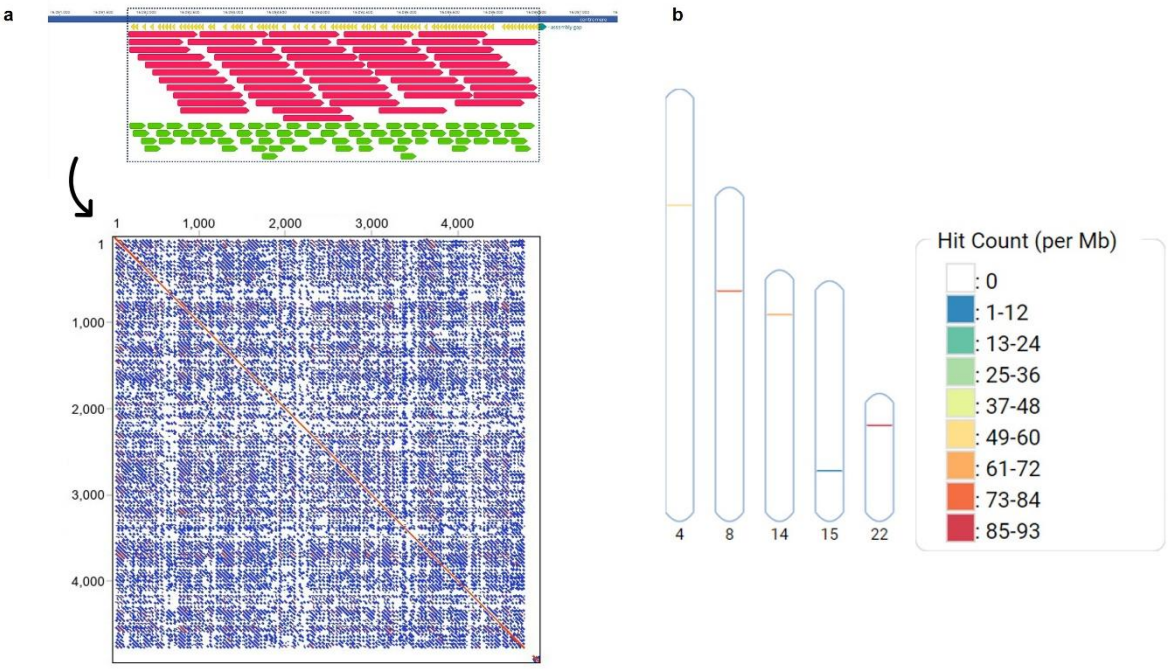



**Figure II.8. Illustration of HSA14 (peri)centromeric region (representative) and current whole-genome annotations. (a)** - Satellite I organization analysis (Geneious) depicted in a dotplot of the pericentromere region rich in SatI hits (self-to-self comparison). **(b)** - Graphical representation of current human genome annotations for satellite I (Dfam), locating this SatDNA family solely in chromosomes 4, 8, 14, 15 and 22.

## II.4. Discussion

The present study of classical satellite I allowed to molecularly describe and map this SatDNA family in the particular rob(14;21) chromosomal context. Surely belonging to the same satellite family, SatI clones presented a high identity degree, nonetheless with some sequence variability, what may result from the isolation of different satellite variants, subfamilies or even divergence within the same subfamily, as already proposed for this satellite (Kalitsis et al. 1993). All SatI clones showed a persistent repeat size of 42 bp in the Tandem Repeats Finder analysis, occasionally displaying some redundancy with repeat periods sized as 42 bp-multiples, probably resulting from repeat unit amplification. Both results are concordant with previous work on isolating pTRI-6, reported as a SatI subfamily: 42 bp monomeric repeats with a conservation of approximately 85%, establishing a HOR-like structure of 2,97 Kb characterized by a 77% AT content (same result as reported above) (Kalitsis et al. 1993). The consistence of the 42 bp repeat, found back in 1986 (Prosser et al. 1986), poses a question: can the repeat be classified as a satellite monomer? Despite the usual rule for hundreds of bp-long monomer units, SatDNA sequences can display much smaller monomer sizes, as is the case with human classical satellites. So, in a wider interpretation, repeat units as small as in micro- and minisatellites can be considered monomers if one considers their organized tandem repetition in constitutive heterochromatin as a characteristic feature of satellite DNA, alongside others (like the known capacity to form heterochromatic regions) (Plohl et al. 2012; Garrido-Ramos 2015; Garrido-Ramos 2017).

Physical mapping of SatI sequences revealed a strong hybridization signal in the pericentromere of chromosome 13 and less intense signals on chromosomes 14, 21 and der(14;21), what is compatible with the former knowledge about SatI retention on the Robertsonian chromosome, at the breakpoint region (Kalitsis et al. 1993). Some authors have long stated that SatI hybridization at lower stringencies produces heteromorphic pericentromeric signals in chromosomes 3, 4 and all acrocentric chromosomes (Tagarro et al. 1994; Therkelsen et al. 1997).

With current genomic data accession, the expected *in silico* analysis would be for SatI detection to correspond to chromosomes 3, 4, 13, 14, 15, 21 and 22 or, at least, chromosomes 13 and 21, the former perceptively matching to a longer SatI array (compatible with present high-stringency physical mapping results with intense hybridization FISH signals). This, however, is unverifiable. *In silico* mapping of 200 bp and 900 bp representative clones revealed

a pericentromeric location on chromosomes 3, 4, 8, 14 and 22. The remaining chromosomes did not display any mappable BLAST hits when SatI sequences were compared against the currently available human reference genome GRCh38.p13, corroborating that the sequence of the reference genome remains extensively uncharacterized, fragmentary, and essentially unassembled, mostly in satellite-composed genomic regions (Miga et al. 2014; Biscotti et al. 2015; Miga 2015; Alvarez-Cubero et al. 2018). In fact, the annotation-poor assembly gap coinciding with the (peri)centromere has been called 'golden path gap' since it is related with the most limitedly analyzed genomic region (due to the lack of accurate assembly algorithms, tandem repetitive sequences are overlooked) (Podgornaya et al. 2018). Even in the chromosomes with mappable hits, it is possible to perceive the current assembly gaps (Figure II.3, II.4, II.5, II.6 and II.7), with a substantial representation.

Unlike the considered parameters for 200 bp and 900 bp SatI sequences, mapping of the reported SatI repeat unit (AB, 42 bp) was performed considering hits with higher e-values ($\leq 10^{-4}$ *versus* $\leq 10^{-16}$), only to show its possible location. The e-value parameter evaluates the number of expected by-chance hits (Kinser 2010) (smaller e-values, more reliable results). Still, sequence length must be taken into account, as short alignments have higher e-values (shorter sequences have a higher probability of occurring by chance in a given genome) (Balding et al. 2008). So, when considering an *in silico* analysis, and for more consistent results, SatI should perhaps be analyzed with longer sequences than the 42 bp repeat unit.

In this work, the flanking regions of SatI hits were also investigated, and the obtained results likewise support the unfinished character of the current genome assembly. TEs have a tendency to accumulate in pericentromeric regions, being closely associated with SatDNAs. Human pericentromeres are indeed frequently interrupted by LINE elements (Schueler et al. 2001; Plohl et al. 2014). Still, this scenario was only demonstrated on chromosome 8, where p-arm hit-rich region showed to be flanked by various TEs.

Figure II.8b illustrates current SatI genome annotations, also limiting its presence, this time to chromosomes 4, 8, 14, 15 and 22. The pericentromeric region of chromosome 14 (representative, with SatI hits) was analyzed through a dotplot matrix with a self-to-self comparison (Figure II.8a), where visible horizontal and vertical lines represent the repetitive nature of this chromosomal section (Huang and Zhang 2004) and diagonal lines correspond to sequence homology (Sonnhammer and Durbin 1995). Interpretation of this dotplot could mean that SatI is organized in HOR-like structures, with arrays arranged in a head-to-tail fashion, as previously postulated (Meyne et al. 1994).

The present study clearly highlights the need for a close relationship between cytogenetic and *in silico* approaches to obtain a chromosomal map with higher resolution. *In silico* analysis can provide unknown information, such as the identifiable high presence of SatI in chromosome 8, unreported until today. Though, for instance, while physical mapping allowed to identify SatI (significant for the context of ROB formation) in chromosomes 13, 14, 21 and der(14;21), *in silico* mapping still did not considered its existence in chromosomes 13 and 21 (clear signal by FISH). Clarifying information about repetitive sequences in these specific chromosomes is essential for understanding the mechanism of the most common ROBs in the human population and, in this case, of rob(14;21). In this line of thought, satellite I should not be skipped when addressing such issues, since the associated information gap is unavoidable in the attempt to narrowly understand ROB formation.

Research in the context of satellite DNA and ROBs should take advantage of a strict collaboration between new technologies like long-read high-throughput sequencing and cytogenetic mapping techniques, both of them closely supported by *in silico* and bioinformatic approaches.

## References

Alvarez-Cubero MJ, Santiago O, Martinez-Labarga C, Martinez-Garcia B, Marrero-Diaz R, Rubio-Roldan A, Perez-Gutierrez AM, Carmona-Saez P, Lorente JA, Martinez-Gonzalez LJ. 2018. Methodology for Y Chromosome Capture: A complete genome sequence of Y chromosome using flow cytometry, laser microdissection and magnetic streptavidin-beads. *Scientific Reports* **8**(1): 9436.

Balding DJ, Bishop M, Cannings C. 2008. *Handbook of Statistical Genetics*. John Wiley & Sons.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**(2): 573-580.

Biscotti MA, Olmo E, Heslop-Harrison JS. 2015. Repetitive DNA in eukaryotic genomes. *Chromosome Research* **23**(3): 415-420.

Chaves R, Ferreira D, Mendes-da-Silva A, Meles S, Adega F. 2017. FA-SAT Is an Old Satellite DNA Frozen in Several Bilateria Genomes. *Genome Biology and Evolution* **9**(11): 3073-3087.

Choo K, Vissel B, Brown R, Filby R, Earle E. 1988. Homologous alpha satellite sequences on human acrocentric chromosomes with selectivity for chromosomes 13, 14 and 21: implications for recombination between nonhomologues and Robertsonian translocations. *Nucleic Acids Research* **16**(4): 1273-1284.

Choo KA. 1997. The centromere, Vol 320. Oxford University Press Oxford.

Dey S. 2011. *Genetics and etiology of Down Syndrome*. BoD–Books on Demand.

Enukashvily NI, Ponomartsev NV. 2013. Mammalian satellite DNA: a speaking dumb. *Advances in Protein Chemistry and Structural Biology* **90**: 31-65.

Garrido-Ramos MA. 2015. Satellite DNA in Plants: More than Just Rubbish. *Cytogenetic and Genome Research* **146**(2): 153-170.

Heslop-Harrison J, Schwarzacher T. 2011. Organisation of the plant genome in chromosomes. *The Plant Journal* **66**(1): 18-33.

Huang Y, Zhang L. 2004. Rapid and sensitive dot-matrix methods for genome analysis. *Bioinformatics* **20**(4): 460-466.

Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AF, Wheeler TJ. 2015. The Dfam database of repetitive DNA families. *Nucleic Acids Research* **44**(D1): D81-D89.

Kalitsis P, Earle E, Vissel B, Shaffer LG, Choo KA. 1993. A chromosome 13-specific human satellite I DNA subfamily with minor presence on chromosome 21: further studies on Robertsonian translocations. *Genomics* **16**(1): 104-112.

Kinser J. 2010. *Python for bioinformatics*. Jones & Bartlett Publishers.

Lee C, Wevrick R, Fisher RB, Ferguson-Smith MA, Lin CC. 1997. Human centromeric DNAs. *Human Genetics* **100**(3-4): 291-304.

Meyne J, Goodwin EH, Moyzis RK. 1994. Chromosome localization and orientation of the simple sequence repeat of human satellite I DNA. *Chromosoma* **103**(2): 99-103.

Miga KH, Newton Y, Jain M, Altemose N, Willard HF, Kent WJ. 2014. Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Research* **24**(4): 697-707.

Miga KH. 2015. Completing the human genome: the progress and challenge of satellite DNA assembly. *Chromosome Research* **23**(3): 421-426.

Miga KH. 2017. Satellite DNA: An Evolving Topic. *Genes* **8**(9).

Miga KH. 2019. Centromeric Satellite DNAs: Hidden Sequence Variation in the Human Population. *Genes* **10**(5): 352.

Plohl M, Meštrović N, Mravinac B. 2012. Satellite DNA evolution. In *Repetitive DNA*, Vol 7, pp. 126-152. Karger Publishers.

Plohl M, Mestrovic N, Mravinac B. 2014. Centromere identity from the DNA point of view. *Chromosoma* **123**(4): 313-325.

Podgornaya OI, Ostromyshenskii DI, Enukashvily NI. 2018. Who Needs This Junk, or Genomic Dark Matter. *Biochemistry Biokhimiia* **83**(4): 450-466.

Prosser J, Frommer M, Paul C, Vincent PC. 1986. Sequence relationships of three human satellite DNAs. *Journal of Molecular Biology* **187**(2): 145-155.

Rozen S, Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. In *Bioinformatics methods and protocols*, pp. 365-386. Springer.

Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF. 2001. Genomic and genetic definition of a functional human centromere. *Science* **294**(5540): 109-115.

Shaffer LG. 2002. Robertsonian translocations. *Wiley Encyclopedia of Molecular Medicine*.

Shiels C, Coutelle C, Huxley C. 1997. Contiguous arrays of satellites 1, 3, and β form a 1.5-Mb domain on chromosome 22p. *Genomics* **44**(1): 35-44.

Sonnhammer EL, Durbin R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167**(1-2): GC1-GC10.

Tagarro I, Wiegant J, Raap AK, González-Aguilera JJ, Fernández-Peralta AM. 1994. Assignment of human satellite 1 DNA as revealed by fluorescent in situ hybridization with oligonucleotides. *Human Genetics* **93**(2): 125-128.

Therkelsen A, Nielsen A, Kølvraa S. 1997. Localisation of the classical DNA satellites on human chromosomes as determined by primed in situ labelling (PRINS). *Human Genetics* **100**(3-4): 322-326.

Therman E, Susman B, Denniston C. 1989. The nonrandom participation of human acrocentric chromosomes in Robertsonian translocations. *Annals of Human Genetics* **53**(1): 49-65.

Trowell HE, Nagy A, Vissel B, Choo KA. 1993. Long-range analyses of the centromeric regions of human chromosomes 13, 14 and 21: identification of a narrow domain containing two key centromeric DNA elements. *Human Molecular Genetics* **2**(10): 1639-1649.

Vissel B, Nagy A, Choo K. 1992. A satellite III sequence shared by human chromosomes 13, 14, and 21 that is contiguous with α satellite DNA. *Cytogenetic and Genome Research* **61**(2): 81-86.

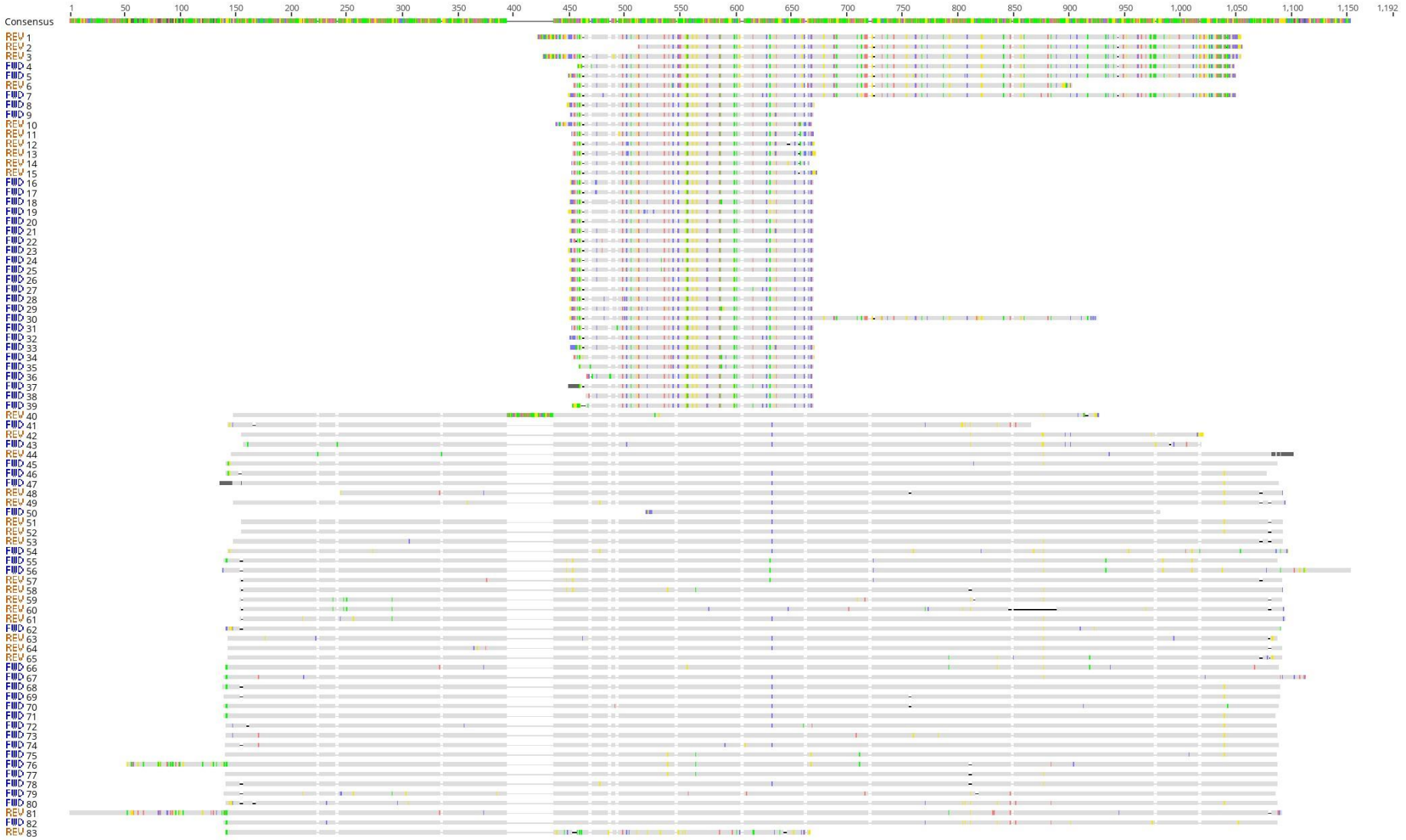
Warburton PE, Hasson D, Guillem F, Lescale C, Jin X, Abrusan G. 2008. Analysis of the largest tandemly repeated DNA families in the human genome. *BMC Genomics* **9**: 533.

## Supplementary Information

**Supplementary Table II.S1. Sequences of the four sets of primers utilized for SatI isolation and chromosome painting probes amplification**. Primers for SatI were designed from the sequences corresponding to the Accession Numbers NONHSAT216991.1, X00470.1, JX174276.1 and L01057.1, respectively. Both forward and reverse sequences are presented.

| | |
|---|---|
| **NonHSatI_Fw** | 5'-TTCGTCTAGTTTGATATTTTGG-3' |
| **NonHSatI_Rev** | 5'-CATATTACATATGTGCATAAAA-3' |
| **XHSatI_Fw** | 5'-GTCTTTCAAAGGTCAGAAGA-3' |
| **XHSatI_Rev** | 5'-CATAACCGATGAAACCTACT-3' |
| **JxHSatI_Fw** | 5'-ATGTGCGGTACATAAGAT-3' |
| **JxHSatI_Rev** | 5'-AAATATGGTTGGGTACTT-3' |
| **LHSatI_Fw** | 5'-TGTGCAGCATGTAATATGAA-3' |
| **LHSatI_Rev** | 5'-ACGTTGCATAAACTATCAAA-3' |
| **6MW** | CCGACTCGAGNNNNNNATGTGG |

**Supplementary Figure II.S1 (additional PDF). Distance matrix of the pairwise alignment of all SatI clones presented in Figure II.1 with more detailed information.** Nucleotide identities are shown by value and the same color code of Figure II.1. To particularly access distances in the matrix check the additional provided PDF file.




**Supplementary Figure II.S2. Multiple sequence alignment (CLUSTALW matrix) of all 83 SatI clones.** Consensus sequence and clone identity are graphically represented.

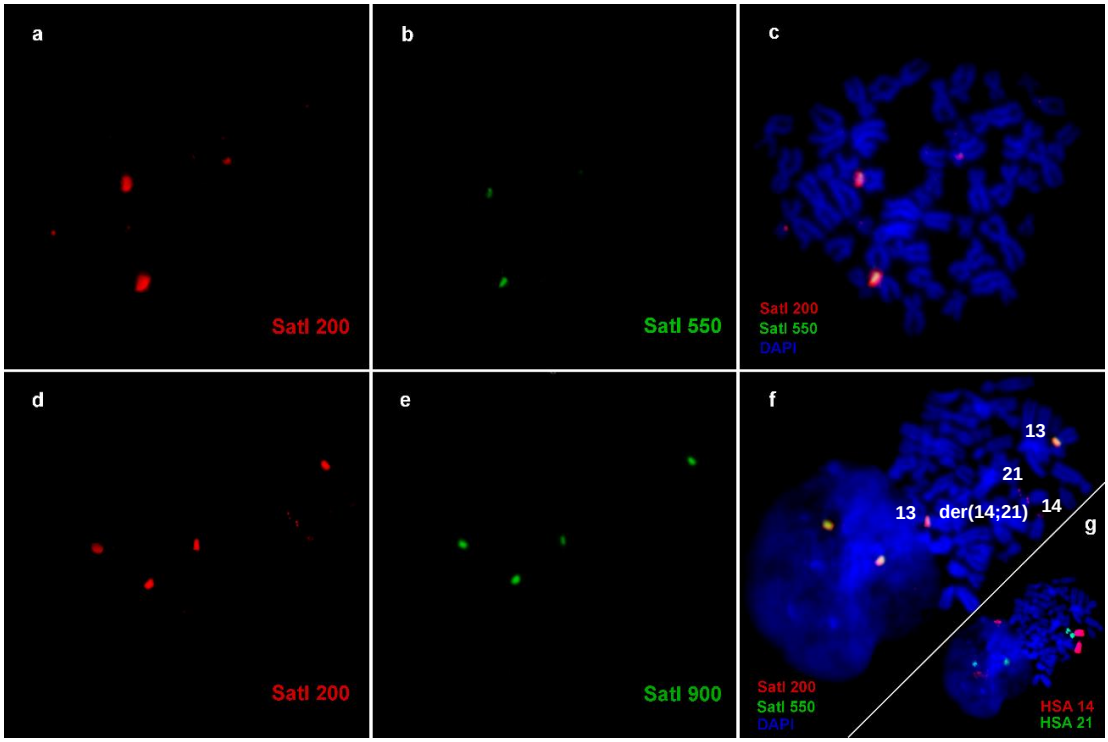
**Supplementary Table II.S2. Summary of SatI clone's analysis in Tandem Repeats Finder.** A number of different parameters are shown, namely the extent of the repeat region, the percentage of matches, the obtained period size and its copy number, the percentage of GC and the consensus pattern for each sequence. Period sizes are 42 bp-long or display approximate sizes corresponding to multiples of 42. ≈200 bp clones are shown in pink, ≈550 bp are in green and the blue cells correspond to ≈900 bp clones.

| | Indices | % matches | Period size | Copy number | %GC | Consensus pattern |
|---|---|---|---|---|---|---|
| *Cl1* | 14-205 | 82 | 42 | 4.6 | 27 | TTGGGTACTTTGATATTTTATGTACAGTATATAATACATATT |
| *Cl2* | 12-203 | 83 | 42 | 4.6 | 23 | TTGGGTACTTTGATATTTTATGTACAGTATATAATACATATT |
| *Cl3* | 16-198 | 82 | 42 | 4.2 | 23 | ATATCAAAGTACCCAAAATATATATTATATACTGTACATAAA |
| *Cl4* | 7-194 | 82 | 42 | 4.5 | 22 | ATAAAATATCAAAGTACCCAAAATATATATTATATACTGTAC |
| *Cl5* | 11-202 | 81 | 42 | 4.6 | 24 | TTGGGTACTTTGATATTTTATGTACAGTATATAATACATATT |
| *Cl6* | 11-201 | 83 | 42 | 4.6 | 23 | GTACATAAAATATCAAAGTACCCAAAATATATATTATATACT |
| *Cl7* | 11-202 | 81 | 42 | 4.6 | 24 | TTGGGTACTTTGATATTTTATGTACAGTATATAATACATATT |
| *Cl8* | 11-202 | 83 | 42 | 4.6 | 23 | TTGGGTACTTTGATATTTTATGTACAGTATATAATACATATT |
| *Cl9* | 13-204 | 83 | 42 | 4.6 | 24 | TTGGGTACTTTGATATTTTATGTACAGTATATAATACATATT |
| *Cl10* | 10-202 | 83 | 42 | 4.6 | 23 | TTGGGTACTTTGATATTTTATGTACAGTATATAATACATATT |
| *Cl11* | 12-203 | 82 | 42 | 4.6 | 23 | TTGGGTACTTTGATATTTTATGTACAGTATATAATACATATT |
| *Cl12* | 12-203 | 83 | 42 | 4.6 | 24 | TTGGGTACTTTGATATTTTATGTACAGTATATAATACATATT |
| *Cl13* | 41-198 | 83 | 42 | 3.8 | 23 | TATATACTGTACATAAAATATCAAATACCCAAAATATGTAT |
| *Cl14* | 14-201 | 81 | 42 | 4.5 | 24 | ATAAAATATCAAAGTACCCAAAATATATATTATATACTGTAC |
| *Cl15* | 12-203 | 83 | 42 | 4.6 | 23 | TTGGGTACTTTGATATTTTATGTACAGTATATAATACATATT |
| *Cl16* | 11-202 | 83 | 42 | 4.6 | 23 | TTGGGTACTTTGATATTTTATGTACAGTATATAATACATATT |
| *Cl17* | 11-202 | 81 | 42 | 4.6 | 24 | TTGGGACTTTGATATTTTATGTACAGTATATAATACATATT |
| *Cl18* | 12-199 | 81 | 42 | 4.5 | 23 | ATAAAATATCAAAGTACCCAAAATATATATTATATACTGTAC |
| *Cl19* | 9-200 | 80 | 42 | 4.6 | 24 | TTGGGTACTTTGATATTTTATGTACAGTATATAATACATATT |
| *Cl20* | 12-203 | 79 | 42 | 4.6 | 24 | TTGGGTACTTTGATATTTTATGTACAGTATATAATACATATT |
| *Cl21* | 12-203 | 77 | 42 | 4.6 | 25 | TTGGGTACTTTGATATTTTATGTACAGTATATAATACATATT |
| *Cl22* | 12-203 | 80 | 42 | 4.6 | 24 | TTGGGTACTTTGATATTTTATGTACAGTATATAATACATATT |
| *Cl23* | 11-202 | 81 | 42 | 4.6 | 24 | TTGGGTACTTTGATATTTTATGTACAGTATATAATACATATT |
| *Cl24* | 13-204 | 81 | 42 | 4.6 | 24 | TTGGGTATTTGATATTTTATGTACAGTATATAATACATATT |
| *Cl25* | 4-197 | 83 | 42 | 4.6 | 24 | TTTGGGTACTTTGATATTTTATGTACAGTATATAATACATAT |
| *Cl26* | 4-197 | 80 | 42 | 4.6 | 23 | TTTGGGTACTTTGATATTTTATGTACAGTATATAATACATAT |
| *Cl27* | 5-190 | 82 | 42 | 4.5 | 24 | ACTTTGATATTTTATGTACAGTATATAATACATATTTTGGGT |
| *Cl28* | 11-202 | 83 | 42 | 4.6 | 23 | TTGGGTACTTTGATATTTTATGTACAGTATATAATACATATT |
| *Cl29* | 13-204 | 83 | 42 | 4.6 | 23 | TTGGGTACTTTGATATTTTATGTACAGTATATAATACATATT |
| *Cl30* | 12-203 | 83 | 42 | 4.6 | 24 | TTGGGTACTTTGATATTTTATGTACAGTATATAATACATATT |
| *Cl31* | 9-189 | 81 | 42 | 4.3 | 23 | ATATTTTATGTACAGTATATAATACATATTTTGGGTACTTTG |
| *Cl1* | 8-433 | 81 | 42 | 10.2 | 23 | CATAAAATATCAAAGTACCCAAAATATATATTATATACTGTA |
| *Cl2* | 14-527 | 79 | 42 | 12.3 | 23 | ATAAAATATCAAAATACCCAAAATATATATTATATACTGTAC |
| | 14-529 | 81 | 167 | 3.1 | 23 | |
| | 14-519 | 81 | 209 | 2.4 | 23 | |
| | 9-529 | 79 | 125 | 4.2 | 24 | |
| *Cl3* | 13-580 | 82 | 42 | 13.6 | 23 | TTGGGTACTTTGATATTTTATGTACAGTATATAATATATATT |
| | 13-580 | 82 | 125 | 4.5 | 23 | |
| | 13-580 | 81 | 167 | 3.4 | 23 | |
| *Cl4* | 4-736 | 79 | 42 | 17.6 | 23 | TTTGGGTACTTTGATATTTTATGTACAGTATATAATATATAT |
| | 3-736 | 79 | 83 | 8.8 | 24 | |
| | 4-736 | 80 | 125 | 5.9 | 23 | |
| | 4-740 | 84 | 165 | 4.4 | 23 | |

*(follows in the next 2 pages)*

| | | | | | | |
|---|---|---|---|---|---|---|
| ***Cl5*** | 13-580 | 80 | 42 | 13.6 | 24 | TTGGGTACTTTGATATTTTATGTACAGTATATAATATATATT |
| | 13-580 | 80 | 83 | 6.8 | 24 | |
| | 13-585 | 80 | 125 | 4.6 | 24 | |
| | 13-580 | 80 | 167 | 3.4 | 24 | |
| ***Cl6*** | 12-577 | 81 | 42 | 13.5 | 23 | ACATAAAATATCAAAGTACCCAAAATATATATTATATACTGT |
| | 12-577 | 80 | 167 | 3.4 | 23 | |
| | 12-577 | 81 | 125 | 4.5 | 23 | |
| Cl7 | 9-426 | 79 | 42 | 10.0 | 22 | AATATATATTATATACTGTACATAAAATATCAAAGTACCCAA |
| | 9-426 | 78 | 125 | 3.3 | 22 | |
| | 9-426 | 79 | 167 | 2.4 | 23 | |
| **Cl8** | 13-576 | 81 | 42 | 13.5 | 23 | ATAAAATATCAAAGTACCCAAAATATATATTATATACTGTAC |
| | 13-576 | 81 | 167 | 3.4 | 23 | |
| | 8-576 | 81 | 125 | 4.5 | 24 | |
| **Cl9** | 7-432 | 86 | 42 | 10.2 | 23 | TTATGTACAGTATATAATATATATTTTGGGTACTTTGATATT |
| **Cl10** | 12-454 | 81 | 42 | 10.6 | 24 | TTGGGTACTTTGATATTTTATGTACAGTATATAATATATATT |
| | 12-460 | 79 | 125 | 3.6 | 24 | |
| **Cl1** | 41-889 | 82 | 42 | 20.2 | 23 | ATATTTTATGTACAGTATATAATATATATTTGGGTACTTTG |
| **Cl2** | 21-869 | 83 | 42 | 20.2 | 22 | TAAAATATCAAAGTACCCAAAATATATATTATATACTGTACA |
| **Cl3** | 36-883 | 81 | 42 | 20.2 | 22 | ATATTTTATGTACAGTATATAATATATATTTGGGTACTTTG |
| | 36-883 | 83 | 167 | 5.1 | 22 | |
| | 36-883 | 82 | 125 | 6.7 | 22 | |
| **Cl4** | 36-884 | 81 | 42 | 20.2 | 22 | ATATTTTATGTACAGTATATAATATATATTTGGGTACTTTG |
| **Cl5** | 37-885 | 81 | 42 | 20.2 | 22 | ATATTTTATGTACAGTATATAATATATATTTGGGTACTTTG |
| **Cl6** | 37-883 | 81 | 42 | 20.2 | 23 | ATATTTTATGTACAGTATATAATATATATTTGGGTACTTTG |
| | 29-883 | 81 | 125 | 6.8 | 23 | |
| | 37-883 | 81 | 208 | 4.0 | 23 | |
| **Cl7** | 14-854 | 83 | 42 | 20.1 | 22 | TATCAAAGTACCCAAAATATATATTATATACTGTACATAAAA |
| | 14-854 | 83 | 167 | 5.0 | 22 | |
| | 14-862 | 81 | 125 | 6.8 | 22 | |
| **Cl8** | 10-850 | 82 | 42 | 20.0 | 23 | TATCAAAGTACCCAAAATATATATTATATACTGTACATAAAA |
| **Cl9** | 35-883 | 82 | 42 | 20.2 | 23 | ATATTTTATGTACAGTATATAATATATATTTTGGGTACTTTG |
| **Cl10** | 6-785 | 81 | 42 | 18.6 | 22 | TATTATATACTGTACATAAAATATCAAAGTACCCAAATATA |
| **Cl11** | 8-855 | 83 | 42 | 20.2 | 22 | CATAAAATATCAAAGTACCCAAAATATATATTATATACTGTA |
| **Cl12** | 36-884 | 82 | 42 | 20.2 | 23 | ATATTTTATGTACAGTATATAATATATATTTGGGTACTTTG |
| **Cl13** | 15-856 | 83 | 42 | 20.1 | 22 | TATCAAAGTACCCAAAATATATATTATATACTGTACATAAAA |
| **Cl14** | 36-884 | 81 | 42 | 20.2 | 22 | ATATTTTATGTACAGTATATAATATATATTTGGGTACTTTG |
| **Cl15** | 9-856 | 82 | 42 | 20.2 | 22 | CATAAAATATCAAAGTACCCAAAATATATATTATATACTGTA |
| | 9-856 | 82 | 125 | 6.7 | 22 | |
| | 9-856 | 82 | 167 | 5.1 | 22 | |
| **Cl16** | 35-883 | 82 | 42 | 20.2 | 22 | ATATTTTATGTACAGTATATAATATATATTTGGGTACTTTG |
| **Cl17** | 34-875 | 83 | 42 | 20.1 | 22 | ATATTTTATGTACAGTATATAATATATATTTTGGGTACTTTG |
| **Cl18** | 20-888 | 81 | 42 | 18.6 | 23 | ATATTTTATGTACAGTATATAATATATATTTGGGTACTTTG |
| **Cl19** | 34-882 | 82 | 42 | 20.2 | 22 | ATATTTTATGTACAGTATATAATATATATTTGGGTACTTTG |
| **Cl20** | 34-883 | 80 | 42 | 20.2 | 23 | ATATTTTATGTACAGTATATAATATATATTTGGGTACTTTG |
| **Cl21** | 34-881 | 82 | 42 | 20.2 | 22 | ATATTTTATGTACAGTATATAATATATATTTGGGTACTTTG |
| | 34-881 | 83 | 167 | 5.1 | 22 | |
| | 34-881 | 82 | 124 | 6.7 | 22 | |
| **Cl22** | 15-856 | 83 | 42 | 20.1 | 22 | TATCAAAGTACCCAAAATATATATTATATACTGTACATAAAA |
| **Cl23** | 37-885 | 82 | 42 | 20.2 | 22 | ATATTTTATGTACAGTATATAATATATATTTGGGTACTTTG |
| **Cl24** | 24-854 | 83 | 42 | 19.8 | 22 | CAAAATATATATTATATACTGTACATAAAATATCAAAGTACC |
| **Cl25** | 36-883 | 82 | 42 | 20.2 | 22 | ATATTTTATGTACAGTATATAATATATATTTGGGTACTTTG |
| | 36-883 | 83 | 167 | 5.1 | 22 | |
| | 36-883 | 82 | 125 | 6.7 | 22 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| Cl26 | 37-885 | 82 | 42 | 20.2 | 22 | ATATTTTATGTACAGTATATAATATATATTTGGGTACTTTG |
| Cl27 | 14-855 | 20.1 | 42 | 82 | 22 | TATCAAAGTACCCAAAATATATATTATATACTGTACATAAAA |
| | 745-1030 | 1.9 | 154 | 96 | 24 | |
| Cl28 | 37-885 | 82 | 42 | 20.2 | 23 | ATATTTTATGTACAGTATATAATATATATTTTGGGTACTTTG |
| Cl29 | 125-972 | 82 | 42 | 20.2 | 22 | ATATTTTATGTACAGTATATAATATATATTTGGGTACTTTG |
| | 125-972 | 82 | 167 | 5.1 | 22 | |
| | 125-972 | 80 | 125 | 6.8 | 22 | |
| Cl30 | 25-855 | 83 | 42 | 19.8 | 23 | CAAAATATATATTATATACTGTACATAAAATATCAAAGTACC |
| Cl31 | 36-882 | 82 | 42 | 20.2 | 22 | ATATTTTATGTACAGTATATAATATATATTTGGGTACTTTG |
| | 28-882 | 81 | 125 | 6.8 | 23 | |
| | 36-882 | 82 | 208 | 4.0 | 22 | |
| Cl32 | 27-857 | 83 | 42 | 19.8 | 23 | CAAAATATATATTATATACTGTACATAAAATATCAAAGTACC |
| Cl33 | 11-664 | 80 | 42 | 15.6 | 23 | TAATATATATTTGGGTACTTTGATATTTTATGTACAGTATA |
| Cl34 | 9-779 | 83 | 42 | 18.4 | 23 | CATAAAATATCAAAGTACCCAAAATATATATTATATACTGTA |
| | 9-779 | 83 | 125 | 6.1 | 23 | |
| Cl35 | 4-733 | 84 | 42 | 17.4 | 22 | TACTGTACATAAAATATCAAAGTACCCAAAATATATATTATA |
| Cl36 | 37-885 | 82 | 42 | 20.2 | 23 | ATATTTTATGTACAGTATATAATATATATTTGGGTACTTTG |
| Cl37 | 11-859 | 83 | 42 | 20.2 | 22 | CATAAAATATCAAAGTACCCAAAATATATATTATATACTGTA |
| Cl38 | 35-883 | 82 | 42 | 20.2 | 22 | ATATTTTATGTACAGTATATAATATATATTTGGGTACTTTG |
| Cl39 | 19-881 | 80 | 42 | 20.6 | 23 | ATTTGGGTACTTTGATATTTTATGTACAGTATATAATATAT |
| Cl40 | 14-855 | 82 | 42 | 20.1 | 22 | TATCAAAGTACCCAAAATATATATTATATACTGTACATAAAA |
| Cl41 | 16-815 | 82 | 42 | 19.1 | 22 | TATCAAAGTACCCAAAATATATATTATATACTGTACATAAAA |
| Cl42 | 34-882 | 82 | 42 | 20.2 | 22 | ATATTTTATGTACAGTATATAATATATATTTGGGTACTTTG |




**Supplementary Figure II.S3. Physical mapping of representative 200, 550 and 900 bp SatI clones in human metaphases bearing the rob(14;21). (a-f)** - All three represented clones co-hybridize in the same pericentromeric locations. **(g)** - The use of painting probes for human chromosomes 14 and 21 allows to identify both chromosomes and the derivative robertsonian chromosome. The name and color of each probe are indicated within each section. Chromosomes are counterstained with DAPI (blue). Digoxigenin-labelled probes (SatI 200 and HSA14) were detected with antidigoxigenin-5'-TAMRA (red). Biotin-labelled probes (SatI 550, SatI 900 and HSA21) were detected with FITC-conjugated avidin (green).

**Supplementary Table II.S3**. **Number of mapped hits of SatI 200 bp, SatI 900 bp and Sat I AB 42 bp in each human chromosome. ***Satellite I AB 42 bp (in grey) hits are shown despite being considered with higher e-values (BLAST hits were filtrated for e-values $\leq 10^{-16}$ but in the case of the SatI AB 42 bp only alignments with e-values $>10^{-4}$ were discarded), aiming to address the localization of the reported SatI monomer (42 bp).

| Chromosome | Satellite I 200 bp | Satellite I 900 bp | Satellite I AB 42 bp* |
|---|---|---|---|
| HSA 1 (NC_000001.11) | 0 | 0 | 0 |
| HSA 2 (NC_000002.12) | 0 | 0 | 0 |
| **HSA 3 (NC_000003.12)** | **4** | **0** | **10** |
| **HSA 4 (NC_000004.12)** | **51** | **32** | **68** |
| HSA 5 (NC_000005.10) | 0 | 0 | 0 |
| HSA 6 (NC_000006.12) | 0 | 0 | 0 |
| HSA 7 (NC_000007.14) | 0 | 0 | 0 |
| **HSA 8 (NC_000008.11)** | **70** | **51** | **98** |
| HSA 9 (NC_000009.12) | 0 | 0 | 0 |
| HSA 10 (NC_000010.11) | 0 | 0 | 0 |
| HSA 11 (NC_000011.10) | 0 | 0 | 0 |
| HSA 12 (NC_000012.12) | 0 | 0 | 0 |
| HSA 13 (NC_000013.11) | 0 | 0 | 0 |
| **HSA 14 (NC_000014.9)** | **71** | **54** | **86** |
| HSA 15 (NC_000015.10) | 0 | 0 | 0 |
| HSA 16 (NC_000016.10) | 0 | 0 | 0 |
| HSA 17 (NC_000017.11) | 0 | 0 | 0 |
| HSA 18 (NC_000018.10) | 0 | 0 | 0 |
| HSA 19 (NC_000019.10) | 0 | 0 | 0 |
| HSA 20 (NC_000020.11) | 0 | 0 | 0 |
| HSA 21 (NC_000021.9) | 0 | 0 | 0 |
| **HSA 22 (NC_000022.11)** | **68** | **52** | **88** |
| HSA X (NC_000023.11) | 0 | 0 | 0 |
| HSA Y (NC_000024.10) | 0 | 0 | 0 |

# Chapter III – Mapping the human (peri)centromeres involved in rob(14;21) by physical and *in silico* approaches

*(In preparation)*


## Abstract

(Peri)centromeric repetitive sequences, and more precisely satellite DNA (SatDNA) sequences are a major heterochromatic genomic element, deeply involved in Robertsonian translocations (ROBs). ROBs occur nonrandomly between acrocentric chromosomes (human chromosomes 13, 14, 15, 21 and 22). Currently, even in the era of genomic innovations, ROB mechanism and breakpoint location remain not entirely understood, essentially due to the difficulties of assemble (peri)centromeric sequences. Thus, the study of Down-associated rob(14;21) significantly lacks mapping information for centromere-adjacent 14p and 21p sequences. By comparing the hybridization pattern of (peri)centromeric SatDNAs and the still-contemporary assembly gaps of the human reference genome, it is possible to recognize the preserved utility of physically mapping satellite probes. The goal of the present work is precisely to shed more light into satellite arrangement (order and composition), in chromosomes 14, 21 and der(14;21), while comparing current genome information and physical cytogenetic mapping.


## III.1. Introduction

Clearly an example of structural uniqueness, acrocentric chromosomes constitute a study subject when discussing chromosomal alterations, and more precisely Robertsonian Translocations (ROBs). ROBs arise from the sequence permutation between short arms of acrocentric chromosomes (in humans, chromosomes 13, 14, 15, 21 and 22) (Jarmuz-Szymczak et al. 2014). Although all acrocentric chromosomes can undergo ROBs, the dissimilar frequency of occurrence in the human population demonstrates a nonrandom acrocentric participation (Han et al. 1994). The most common ROBs in the human population are rob(13;14) (70%) and rob(14;21) (10%) (Therman et al. 1989). The clinical significance of rob(14;21) justifies its deep understanding since it can be responsible for the extra chromosome 21 material related with Down syndrome (the translocation is present in 3-4% of Down karyotypes) (Wilch and Morton 2018).

Robertsonian chromosomes display a monocentric appearance, however often having a dicentric nature (Earle et al. 1992; Gravholt et al. 1992; Bandyopadhyay et al. 2002), probably as a result of short arm fusion. Mitotic stability of the dicentric Robertsonian chromosome is achieved by the suppression of one of the centromeres, nonrandomly (Sullivan et al. 1996). However, small inter-centromeric distances may promote the upkeep of both active centromeres (by reducing chances for illegitimate chromosome segregation) (Stimpson et al. 2012) and thus allowing the occurrence of functional dicentric chromosomes (Sullivan and Schwartz 1995; Page and Shaffer 1998; Sullivan and Willard 1998; Higgins et al. 2005; Stimpson et al. 2010).

In the most frequent ROBs, p-arm fusion arises from two recurrent breakpoint regions. Nevertheless, the exact location of the breakpoints remains unknown, in spite of seeming connected with the type of ROB or the mechanism for ROB formation (still not understood) (Kaiser-Rogers and Rao 2013). One of the proposed sequence of events involves the recombination of satellite III DNA sequences and other short-arm satellite sequences (Han et al. 1994). Additionally, the preferential manifestation of some ROBs over others was hypothesized to be due to the 14p inverse sequence orientation in relation to 21p and 13p (homologous satellite sequences may be inversely positioned) (Choo et al. 1988; Therman et al. 1989; Shaffer 2002).

(Peri)centromeric sequences of human chromosomes are largely composed of repetitive elements, from which the organized tandem arrays of satellite DNA (SatDNA) sequences stand out for being the most representative heterochromatic element. SatDNA families can be quite variable, namely in abundance, repeat unit length, sequence composition or chromosomal location (Garrido-Ramos 2017). The human centromere is formed by tandem head-to-tail-organized 171 bp-long satellite repeats, named arrays of α satellite, that can be adjacently placed in the shape of Higher Order Repeats (HORs). As for the pericentromere, the tandemly repeated sequences are shorter, mainly establishing the presence of the three classical satellite families (SatI, II and III), often associated with acrocentric chromosomes (Lee et al. 1997; Barra and Fachinetti 2018).

SatI and SatIII families are thought to be of crucial importance for determining ROB breakpoints, and more precisely rob(14;21). In rob(14;21) formation, breakpoint at 14p seems to occur between two SatIII subfamilies: pTRS-47 (more close to the centromere; maintained in the derivative chromosome) and pTRS-63 (more distal, lost during the translocation event) (Earle et al. 1992); breakpoint in 21p is reported between the SatI subfamily pTRI-6 and the r-

DNA (SatI repeats remain in the translocated chromosome) (Kalitsis et al. 1993). β satellite DNA family, present on the p13 of acrocentric chromosomes, is postulated to be lost in ROBs together with rDNA genes (Waye and Willard 1989; Han et al. 1994).

Today, in the era of genomic studies, ROB mechanism and breakpoint location continue to be not fully comprehended, essentially because of the difficulties related with linearly assemble and validate tandem repeated sequences, spread across the longitude of the (peri)centromere (Miga 2019). The deficiency of mapping information for 14p and 21p, acknowledged by some authors while narrowing rob(14;21) breakpoint (Jarmuz-Szymczak et al. 2014), remains a contemporary issue. The present work aims to shed more light into satellite arrangement (order and composition), in chromosomes 14, 21 and der(14;21), while bringing together current genomic data and physical cytogenetic mapping.

## III.2. Material and Methods

### Cell culture, chromosome preparation and genomic DNA isolation

The present study presupposed the comparative use of two commercially available human cell lines:

-GM03417, a mosaic holding the rob(14;21) (46,XX,der(14;21),+21/45,XX,der(14;21));

-GM12878, karyotypically normal and previously used as reference in the Human Genome Project (for comparative purposes). Cells were cultured in DMEM medium supplemented with: 13% AmnioMax C-100 Basal Medium, 2% AmnioMax C-100 supplement, 15% FBS (Fetal Bovine Serum), 1% Glutamine and 1% of antibiotic mixture Penicillin (100 U/mL) / Streptomycin (100 µg/mL). All the reagents mentioned above are commercialized by Gibco, Thermo Fisher Scientific. Chromosome harvesting and chromosomal preparations were achieved recurring to routine procedures. Genomic DNA isolation was performed with the commercial kit QuickGene DNA Tissue Kit S (Fujifilm Life Science), according to the manufacturer's instructions.

### Human SatDNAs isolation and cloning

SatDNAs amplification was performed by PCR (Polymerase Chain Reaction) of human genomic DNA with specific designed primers (Supplementary Table III.S1). Primers were designed using the web-based interface Primer 3 (Rozen and Skaletsky 2000). PCR

amplification followed the subsequent steps: initial denaturing step at 94ºC for 10 min; 30 cycles of 94ºC for 1 min (denaturation), 54ºC for 45 s (annealing) and 72ºC for 45 s (extension); final extension at 72ºC for 10 min. The annealing temperature was optimized for each set of primers. Obtained isolated bands were purified using the QIAquick PCR Purification Kit (Qiagen). PCR amplicon cloning required the use of the Fast DNA End Repair (Thermo Scientific) to blunt and phosphorylate sequence ends for ligation to occur (sequences are ligated to SmaI site on pUC19 with T4 DNA ligase). Transformation was performed with DH5α competent bacterial cells (Invitrogen, Thermo Fisher Scientific). Colonies were selected with blue-white screening (β-galactosidase blue-white α complementation) and positives were confirmed by PCR. Positive clones were sequenced in StabVida by Sanger methodology in order to deeply analyze the isolated sequences and to assess clone similarity.

## DNA-Fluorescent *in situ* hybridization (DNA-FISH)

FISH was standardly performed (Heslop-Harrison and Schwarzacher 2011; Chaves et al. 2017), in order to physically map satellite clones (SatI, SatII, SatIII, αSat and βSat) onto human chromosomes. Human metaphases were sequentially hybridized with two-by-two combinations of satellite cloned sequences and painting probes for human chromosomes 14 and 21, the latter obtained by chromosome sorting. In between hybridization protocols, slides were treated to eliminate previous hybridization signals. Clone probes were PCR labelled and painting probes were labelled by DOP-PCR, with digoxigenin-11-dUTP or biotin-16-dUTP (both from Roche Applied Science). DOP-PCR was performed with degenerated primer 6MW (Supplementary Table III.S1). Hybridization was performed over-night for clone probes and during approximately one week for painting probes. In both cases, post-hybridization washes were guaranteed with temperature (37ºC) and 50% formamide/2xSSC. Digoxigenin-labelled probes were detected with antidigoxigenin-5'-TAMRA (Sigma-Aldrich) and biotin-labelled probes were detected with FITC-conjugated avidin (Vector Laboratories). Preparations were mounted using Vectashield containing 4'-6-diamidino-2-phenylindole (DAPI) (Vector Laboratories) to counterstain chromosomes.

### *In silico* mapping of SatDNAs

Sequence analysis was performed with BLAST (Basic Local Alignment Search Tool) from NCBI (National Center for Biotechnology Information) databases. HSA chromosome sequences (GRCh38.p13; assembly accession: GCA_000001405.28) were collected in FASTA format from NCBI. Satellite I (one representative 200 bp-sized clone; sequence in Supplementary Table III.S2), II (GenBank accession number: X06199.1), III (subfamily pTRS-63, GenBank accession number: S90110.1; subfamily pTRS-47, GenBank accession number: X54108.1) , α (171 bp monomer; sequence in Supplementary Table III.S2) and β (GenBank accession number: M81228.1). DNA sequences were searched in human chromosomes using BLAST, with the following parameters: max_target_seqs was set to 10000 and word size to 11. BLAST hits were filtrated for scores ≥90 and e-values $\leq 10^{-16}$. Filtrated BLAST hits were mapped to human chromosomes (*Homo sapiens* reference genome GRCh38.p13) using Geneious software.

### Image capture and processing

FISH images were observed using a Zeiss ImagerZ microscope coupled to an Axiocam digital camera using AxioVision software (version Rel. 4.5, Zeiss). Digitized photos were prepared for printing in Adobe Photoshop (version 7.0).

## III.3. Results

### SatDNAs physical mapping

Isolated and cloned sequences of human satellites I, II, III, α and β were hybridized in two-by-two combinations onto human chromosomal preparations with rob(14;21) in order to analyze their presence and organization, thus determining their order relatively to each other in chromosomes 14, 21 and der(14;21). Chromosomes were identified by painting probes for chromosomes 14 and 21 (Figure III.1b, d, f; Figure III.2b). SatII probe hybridization did not allow to obtain mappable results, as no signal was detected in the chromosomes of interest. SatI, III, α and β demonstrated to hybridize in the (peri)centromeric region. SatI, III, α hybridization signals were then examined and compared to determine the order of satellite arrangement (Figure III.1, Figure III.3). Observed p-arm (centromere-distal to proximal) SatDNAs disposition was: SatI - αSat - SatIII for HSA14; SatIII - SatI - αSat for HSA21; SatI

- αSat - SatIII for der(14;21). βSat showed to be present in acrocentric chromosomes, more specifically HSA14 and HSA21 and absent in der(14;21) (Figure III.2a).

### *In silico* mapping of SatDNAs in chromosomes 14 and 21

Representative sequences for human satellites (α, β, I, II and two III subfamilies) were submitted to BLAST tools against human chromosomes 14 and 21 (Supplementary Table III.S2). Filtrated BLAST hits (e-value $\leq 10^{-16}$ and bit-score $\geq 90$) were mapped in HSA14 and HSA21 using Geneious software. Mappable hits are presented in Figures III.4 and III.5. In HSA14, only SatI, II, SatIII subfamily pTRS-63 and αSat displayed BLAST hits according to set parameters (even so in greatly discording numbers). In HSA21, no SatI or βSat hits were of mappable character, so the presented hits are for SatII, SatIII subfamilies pTRS-63 and pTRS-47 and αSat. Figures III.4 and III.5 also show the high amount of assembly gaps of the current reference genome (GRCh38.p13), essentially in the (peri)centromeric region.

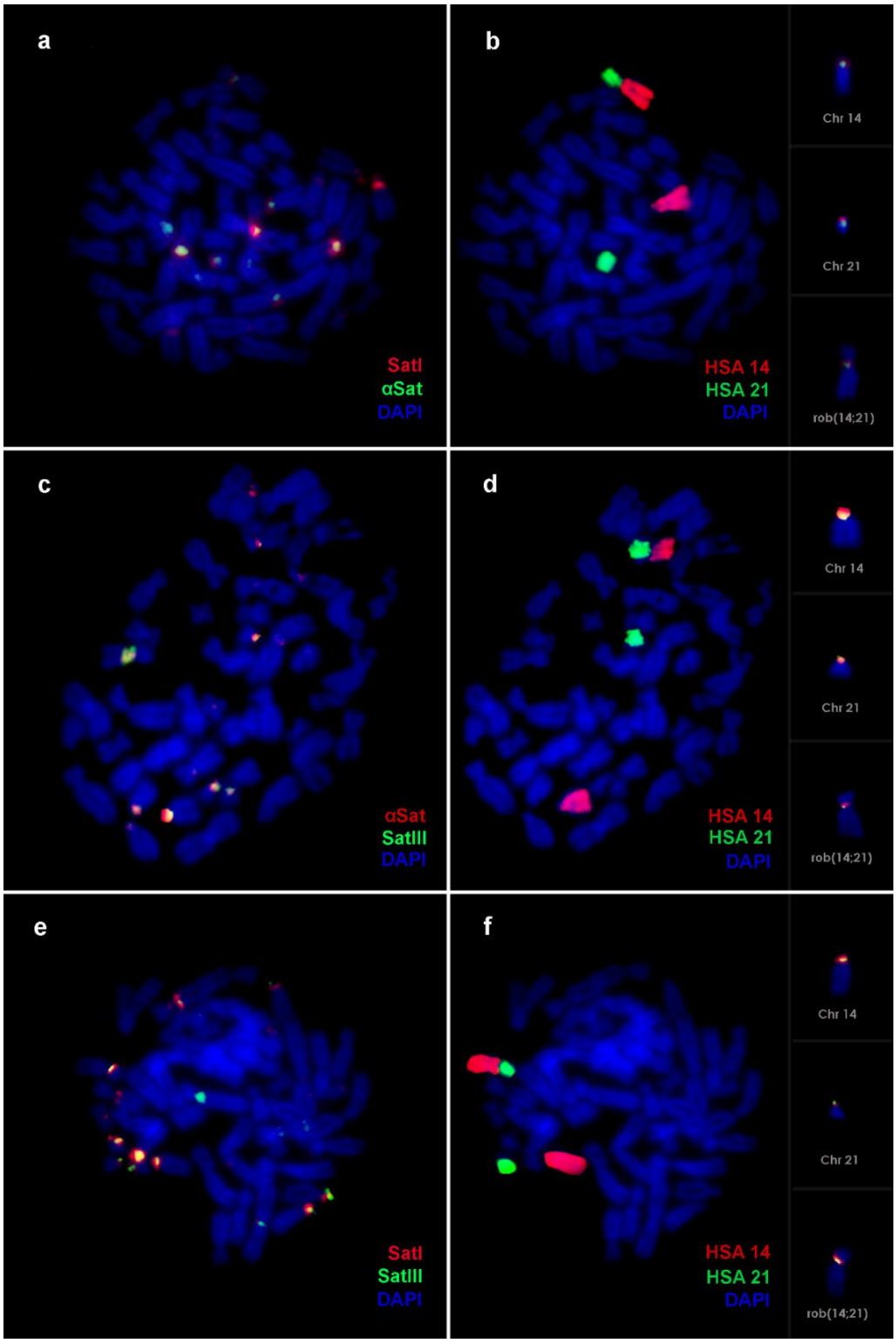



**Figure III.1. Physical mapping of SatI, SatIII and αSat clones in human metaphases bearing the rob(14;21).** **(a, c, e)** - Two-by-two combinations of satellite clones. Hybridization signals can be analyzed in terms of order. **(b, d, f)** - The use of painting probes for human chromosomes 14 and 21 allows to identify both chromosomes and the derivative robertsonian chromosome. Chromosomes of interest are zoomed in (right). The name and color of each probe are indicated within each section. Chromosomes are counterstained with DAPI (blue). Digoxigenin-labelled clone probes (SatI and αSat) and HSA14 painting probe were detected with antidigoxigenin-5'-TAMRA (red). Biotin-labelled clone probes (αSat and SatIII) and HSA21 painting probe were detected with FITC-conjugated avidin (green).

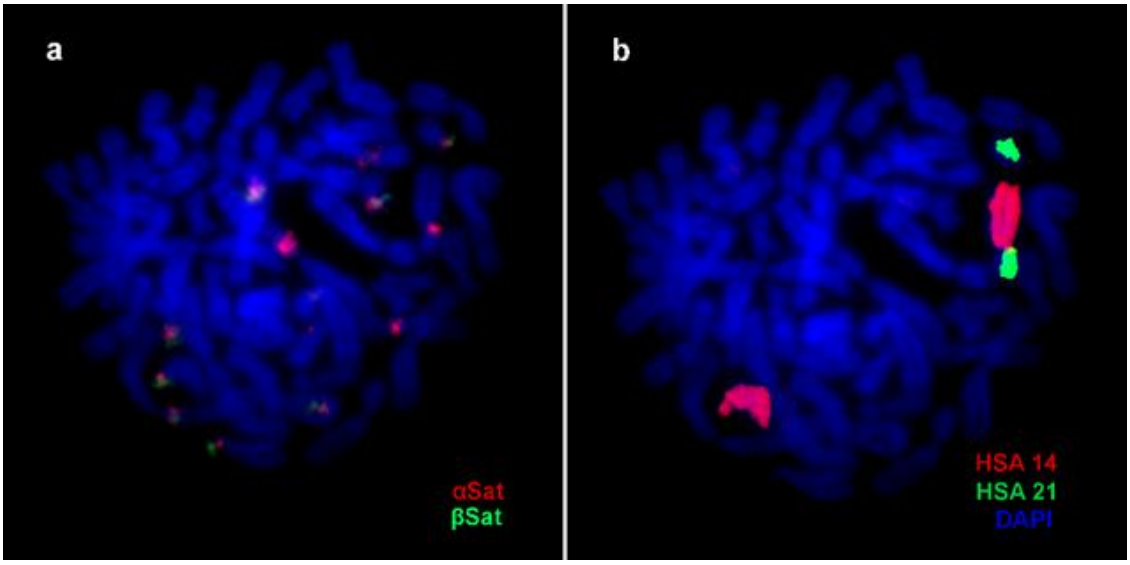



**Figure III.2. Physical mapping representative βSat and αSat clones in human metaphases bearing the rob(14;21). (a)** - Hybridization pattern of the two analyzed satellites. βSat shows to hybridize in acrocentric chromosomes and localize distally to αSat**. (b)** - The use of painting probes for human chromosomes 14 and 21 allows to identify both chromosomes and the derivative robertsonian chromosome. The name and color of each probe are indicated within each section. Chromosomes are counterstained with DAPI (blue). Digoxigenin-labelled αSat clone probe and HSA14 painting probe were detected with antidigoxigenin-5'-TAMRA (red). Biotin-labelled βSat clone probe and HSA21 painting probe were detected with FITC-conjugated avidin (green).

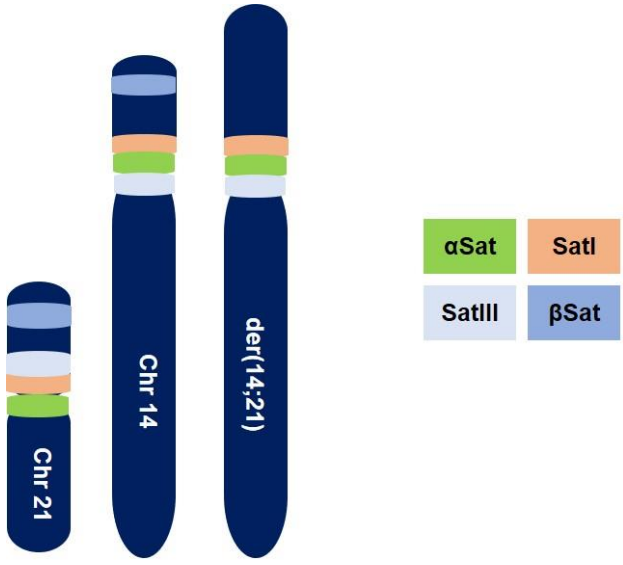



**Figure III.3. Schematic representation of satellite organization observed while physically mapping SatI, SatIII, βSat and αSat clones.** Observed SatDNAs organization was: p-arm - SatI - αSat - SatIII - q-arm for HSA14; p-arm - SatIII - SatI - αSat - q-arm for HSA21; p-arm - SatI - αSat - SatIII - q-arm for der(14;21).


**Figure III.4.** ***In silico*** **mapping of SatI, SatII, SatIII pTRS-63 and αSat in human chromosome 14.** Graphical representation (Geneious) of the chromosome is progressively zoomed in to show hit location and organization. Centromere (in blue) is presented in three subsequent sections. SatI, SatII, SatIII pTRS-63 appear to locate in the pericentromeric p-arm region, while αSat seems to spread across the length of the centromere. Current assembly gaps (human reference genome GRCh38.p13) are also presented.

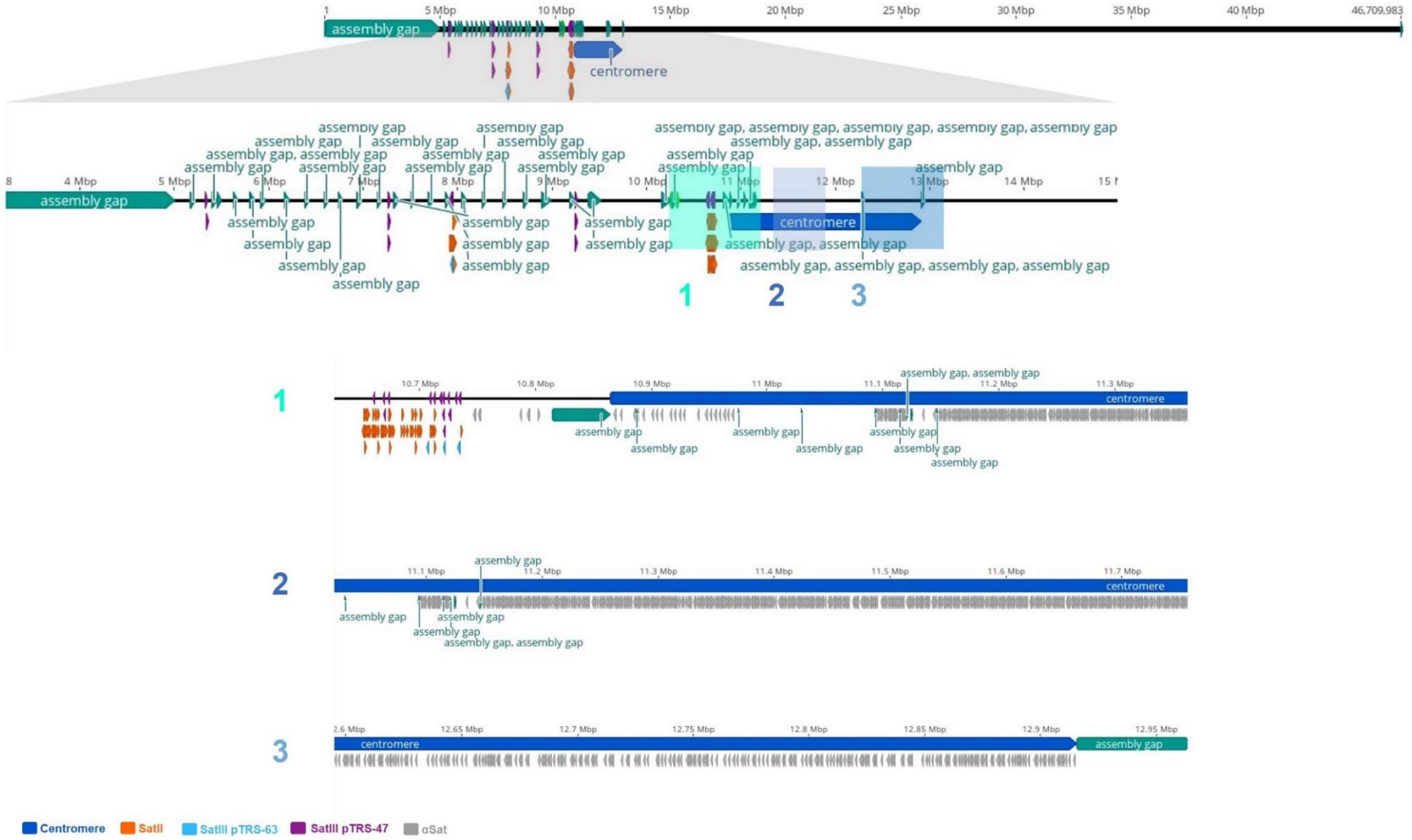



**Figure III.5. *In silico* mapping of SatII, SatIII pTRS-63, SatIII pTRS-47 and αSat in human chromosome 21.** Graphical representation (Geneious) of the chromosome is progressively zoomed in to show hit location and organization. Centromere (in blue) is presented in three subsequent sections. SatII, SatIII pTRS-63, SatIII pTRS-47 appear to locate in the pericentromeric p-arm region, while αSat arrays spread across the length of the centromere. Current assembly gaps (human reference genome GRCh38.p13) are also presented.

## III.4. Discussion

Although being so undeniably important in centromere and overall chromosome structure, repetitive satellite DNA is still seen and understood as the "most obscure genome component" (Satovic et al. 2016). Etiologically, ROB breakpoints seem to be related with satellite sequence organization. In this way, the present study enriched the physical map of the most frequent Down-associated ROB (rob(14;21)) by analyzing the hybridization pattern of (peri)centromeric satellites (I, II, III, α and β).

Short arm of acrocentric chromosomes (p11-p13) are known to contain classical satellites I, II and III and β satellites repeats (Choo et al. 1988; Waye and Willard 1989; Choo et al. 1992; Gravholt et al. 1992). SatII was not analyzable, as no hybridization signals were detected in the chromosomes of interest, possibly because of high stringency matters, FISH resolution or even low satellite copy-number in the respective chromosomes. βSat was observed to localize in the p-arm of acrocentric chromosomes and to be lost in the robertsonian chromosome (Figure III.2, Figure III.3), according to previous bibliography (Waye and Willard 1989; Han et al. 1994). SatI, SatIII and αSat were analyzed in two-by-two combinations to determine their relative organization.

Chromosome 14 displayed the following placement: SatI - αSat - SatIII (p-arm to q-arm direction) (Figure III.1, Figure III.3). SatI location in HSA14 p-arm pericentromeric region is consistent with early reports (Jones et al. 1974; Kalitsis et al. 1993; Lee et al. 1997). However, and given the reported breakpoint localization between SatIII subfamilies pTRS-47 and pTRS-63 (Earle et al. 1992), SatIII signal would be also expected in the p-arm pericentromere (distally localized to αSat). Yet, the obtained results are still compatible with other findings that a specific SatIII subfamily (named pTR9-H2) (Vissel et al. 1992) is located in the pericentromere of HSA14 between two other families of αSat (pTRA-2 and pTRA-7; different from the functionally active centromeric α array). All these subfamilies would then be distal to centromere-proximally organized as following: SatIII pTRS-63, SatIII pTRS-47, αSat pTRA-2, SatIII pTR9-H2 and αSat pTRA-7 (Trowell et al. 1993). So, the fact that SatIII repeats were found after αSat (p-arm to q-arm direction) does not exclude the possibility of the breakpoint between pTRS-63 and pTRS-47.

Chromosome 21 hybridization pattern was: SatIII - SatI - αSat (p-arm to q-arm direction) (Figure III.1, Figure III.3). This result can be consistent with previous reported SatI locations, both in pericentric p11 and distal p13 regions, as well as in pericentromeric 21q (Trowell et al.

1993) (not detected here). Breakpoint in HSA21 was reported to locate between SatI subfamily pTRI-6 and the rDNA genes (Kalitsis et al. 1993).

In the translocated chromosome, satellite probes hybridized as following: SatI - αSat - SatIII (p-arm to q-arm direction) (Figure III.1, Figure III.3), consistently with the previously mentioned breakpoint locations. If the breakpoint in chromosome 21 occurs between SatI and rDNA genes, SatI repeats would be retained. In the case of chromosome 14, where the breakpoint supposedly occurs between SatIII pTRS-47 and pTRS-63 subfamilies, all material distally localized to pTRS-47 would be present in the Robertsonian chromosome, maintaining SatIII repeats. The named translocated chromosome can then be perceived as containing an intercentromeric region with p-arm material from both original chromosomes and plausibly missing βSat sequences and rDNA genes (Hurley and Pathak 1977; Cheung et al. 1990; Gravholt et al. 1992; Wolff and Schwartz 1992; Page et al. 1996; Sullivan et al. 1996; Denison et al. 2002).

The *in silico* analysis was performed by mapping satellite sequences (I, II, III, α and β) to human chromosomes 14 and 21 (current genome assembly GRCh38.p13) (Supplementary Table III.S2). Chromosome 14 only showed mappable hits for SatI, SatII, SatIII pTRS-63 (one hit only) and αSat sequences (Figure III.4). By its turn, chromosome 21 displayed mappable hits for SatII, SatIII pTRS-63, SatIII pTRS-47 and αSat (Figure III.5). To this date, all human genome assemblies continue to be unsatisfactory for the study of satellite sequences. For example, SatII and SatIII display an approximate representation of 0.01% in the GRCh38 genome assembly (Miga 2019) (close to 210 times less than the real genome representativity). Likewise, the absence of βSat hits for HSA14 and HSA21 and SatI hits for HSA21 reveal the existence of handicaps in the current human reference genome, seeing that cytogenetic physical mapping techniques allow to detect the named satellites in these specific chromosomes (Figure III.1, Figure III.2). As for SatIII, the number of hits in HSA14 is evidently poor (Figure III.4), since both pTRS-63 and pTRS-47 are known to be represented in this chromosome. In fact, pTRS-63 is described as chromosome 14-specific (Choo et al. 1992) and pTRS-47 as specific for chromosomes 14 and 22 (Choo et al. 1990). Thus, the hit representation of both subfamilies in HSA21 (Figure III.5) in contrast to the one found in HSA14 is illogical (at least in the light of current knowledge). The present results also demonstrate the uneven representation of satellite families in general comparatively to αSat (greatly represented in HSA14 and HSA21).

The (peri)centromeric physical map presented in Figure III.3 allows us to infer about the etiological and mechanistic origin of ROBs. In Figure III.6, the possible chain of events leading

to this rearrangement is introduced with two alternatives, the first of which contains two previously reported 21p breakpoints (in SatI or SatIII). These alternative scenarios point out to a two-step mechanism for ROB formation involving the loss and reorganization of satellite repeats, as already suggested (Chaves et al. 2003). However, the named reorganization can be perceived as a consequent or causative factor. Figure III.6b suggests a previous reorganization of satellite families in chromosome 21, perhaps leading to its predisposition to ROB formation (reported as a possibility in this translocation (Earle et al. 1992)). In both cases an inversion appears to be necessary to obtain the observable der(14;21) organization. Notwithstanding, the need for possible alternatives throughout this work highlights the urgent requirement for more mapping information and breakpoint confirmation.

The obtained results point out that the study of ROB mechanism and breakpoints still cannot solely rely on genomic technologies. Physical mapping should continue an intervening player in the attempt to achieve accurate maps for pericentromeric/short-arm regions of acrocentric chromosomes.

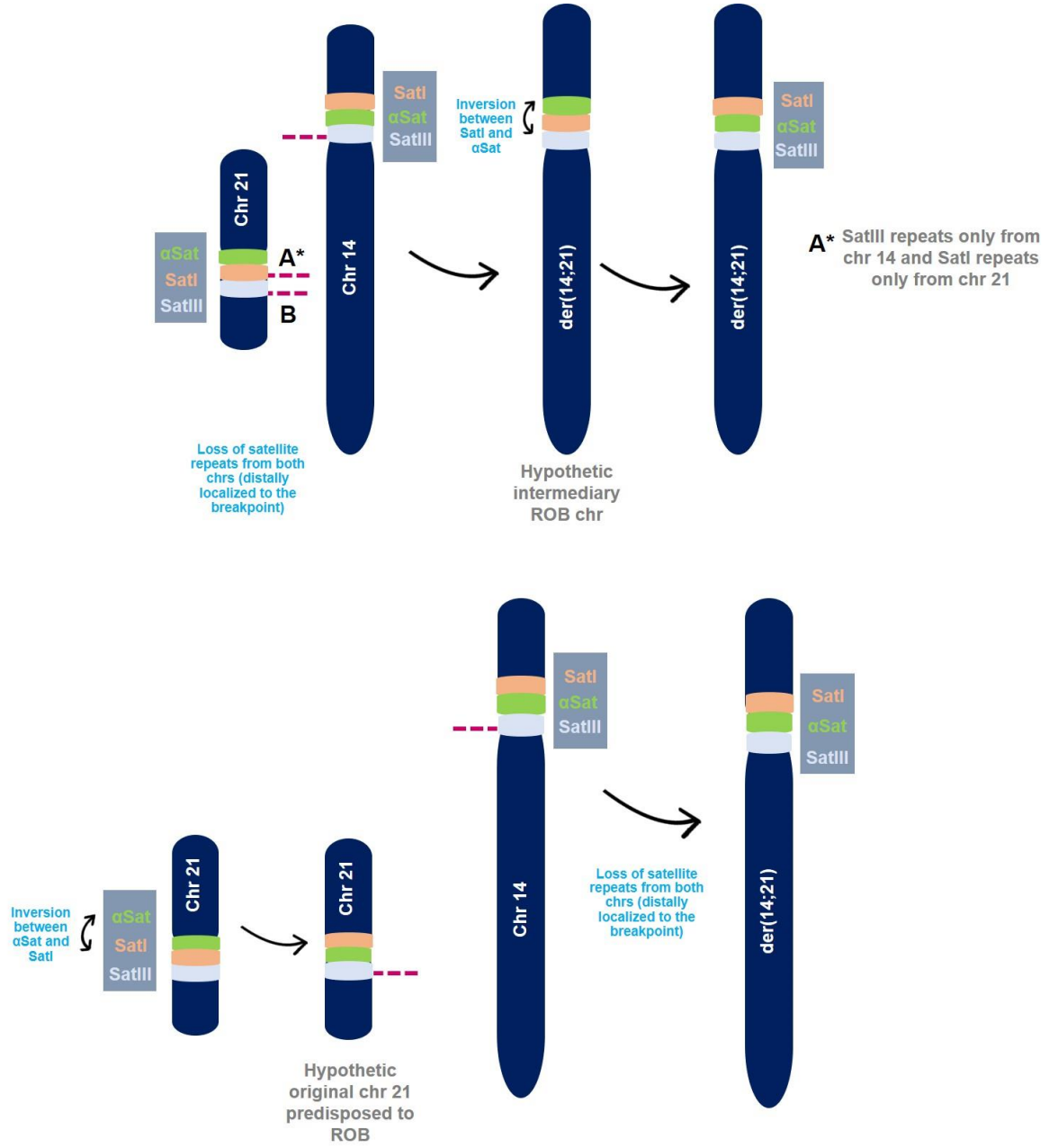



**Figure III.6. Two alternative scenarios for rob(14;21) mechanistic formation. (a)** - The breakpoint at HSA21 (p-arm - SatIII - SatI - αSat - q-arm) has been reported to possibly locate at SatI repeats (Han et al. 1994; Page et al. 1996) (**A***) or SatIII repeats (Gosden et al. 1981; Gravholt et al. 1992) (**B**). The breakpoint at HSA14 (p-arm – SatI - αSat - SatIII - q-arm) has been reported at SatIII (Gravholt et al. 1992; Han et al. 1994; Page et al. 1996; Bandyopadhyay et al. 2002). The expected outcome for der(14;21) organization did not correspond to the verifiable one, so one hypothetic intermediary chromosome is proposed (p-arm - αSat - SatI - SatIII - q-arm). After the formation of this intermediary chromosome (with the loss of satellite repeats distally localized to the breakpoints), SatDNA families are reorganized, possibly leading to the stabilization of the translocated chromosome. Reorganization occurs by an inversion between αSat and SatI and allows the obtainment of the following organization: p-arm - SatI- αSat - SatIII - q-arm. (*) Breakpoint at A causes for SatI and SatIII to have exclusive origins (HSA21 and HSA14, respectively). **(b)** - This alternative suggests the occurrence of an inversion between αSat and SatI in the original HSA21 (p-arm - SatIII – SatI - αSat - q-arm). The hypothetical previous reorganization of SatDNA families in HSA21 could lead to a predisposition for ROB formation, previously suggested as a possibility for this rearrangement (Earle et al. 1992). In this scenario breakpoints would both occur in SatIII and the loss of satellite repeats distally localized to the breakpoints would result in the observable der(14;21) organization: p-arm - SatI- αSat - SatIII - q-arm.

## References

Bandyopadhyay R, Heller A, Knox-DuBois C, McCaskill C, Berend SA, Page SL, Shaffer LG. 2002. Parental origin and timing of de novo Robertsonian translocation formation. *American Journal of Human Genetics* **71**(6): 1456-1462.

Barra V, Fachinetti D. 2018. The dark side of centromeres: types, causes and consequences of structural abnormalities implicating centromeric DNA. *Nature Communications* **9**(1): 4340.

Cheung S, Sun L, Featherstone T. 1990. Molecular cytogenetic evidence to characterize breakpoint regions in Robertsonian translocations. *Cytogenetic and Genome Research* **54**(3-4): 97-102.

Choo K, Earle E, McQuillan C. 1990. A homologous subfamily of satellite III DNA on human chromosomes 14 and 22. *Nucleic Acids Research* **18**(19): 5641-5648.

Choo K, Earle E, Vissel B, Kalitsis P. 1992. A chromosome 14-specific human satellite III DNA subfamily that shows variable presence on different chromosomes 14. *American Journal of Human Genetics* **50**(4): 706.

Choo K, Vissel B, Brown R, Filby R, Earle E. 1988. Homologous alpha satellite sequences on human acrocentric chromosomes with selectivity for chromosomes 13, 14 and 21: implications for recombination between nonhomologues and Robertsonian translocations. *Nucleic Acids Research* **16**(4): 1273-1284.

Denison SR, Multani AS, Pathak S, Greenbaum IF. 2002. Fragility in the 14q21q translocation region. *Genetics and Molecular Biology* 25(3): 271-276.

Earle E, Shaffer L, Kalitsis P, McQuillan C, Dale S, Choo K. 1992. Identification of DNA sequences flanking the breakpoint of human t (14q21q) Robertsonian translocations. *American Journal of Human Genetics* **50**(4): 717.

Garrido-Ramos MA. 2017. Satellite DNA: An Evolving Topic. *Genes* **8**(9).

Gosden J, Lawrie S, Gosden C. 1981. Satellite DNA sequences in the human acrocentric chromosomes: information from translocations and heteromorphisms. *American Journal of Human Genetics* **33**(2): 243.

Gravholt CH, Friedrich U, Caprani M, Jørgensen AL. 1992. Breakpoints in Robertsonian translocations are localized to satellite III DNA by fluorescence in situ hybridization. *Genomics* **14**(4): 924-930.

Han J-Y, Choo K, Shaffer LG. 1994. Molecular cytogenetic characterization of 17 rob (13q14q) Robertsonian translocations by FISH, narrowing the region containing the breakpoints. *American Journal of Human Genetics* **55**(5): 960.

Higgins AW, Gustashaw KM, Willard HF. 2005. Engineered human dicentric chromosomes show centromere plasticity. *Chromosome Research* **13**(8): 745-762.

Hurley JE, Pathak S. 1977. Elimination of nucleolus organizers in a case of 13/14 Robertsonian translocation. *Human Genetics* **35**(2): 169-173.

Jarmuz-Szymczak M, Janiszewska J, Szyfter K, Shaffer LG. 2014. Narrowing the localization of the region breakpoint in most frequent Robertsonian translocations. *Chromosome Research* **22**(4): 517-532.

Jones K, Purdom I, Prosser J, Corneo G. 1974. The chromosomal localisation of human satellite DNA I. *Chromosoma* **49**(2): 161-171.

Kaiser-Rogers K, Rao KW. 2013. Structural Chromosome Rearrangements. In *The Principles of Clinical Cytogenetics*, pp. 139-174. Springer, New York, NY.

Kalitsis P, Earle E, Vissel B, Shaffer LG, Choo KA. 1993. A chromosome 13-specific human satellite I DNA subfamily with minor presence on chromosome 21: further studies on Robertsonian translocations. *Genomics* **16**(1): 104-112.

Lee C, Wevrick R, Fisher RB, Ferguson-Smith MA, Lin CC. 1997. Human centromeric DNAs. *Human Genetics* **100**(3-4): 291-304.

Miga KH. 2019. Centromeric Satellite DNAs: Hidden Sequence Variation in the Human Population. *Genes* **10**(5): 352.

Page SL, Shaffer LG. 1998. Chromosome stability is maintained by short intercentromeric distance in functionally dicentric human Robertsonian translocations. *Chromosome Research* **6**(2): 115-122.

Page SL, Shin J-C, Han J-Y, Andy Choo K, Shaffer LG. 1996. Breakpoint diversity illustrates distinct mechanisms for Robertsonian translocation formation. *Human Molecular Genetics* **5**(9): 1279-1288.

Rozen S, Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. In *Bioinformatics Methods and Protocols*, pp. 365-386. Springer.

Satovic E, Vojvoda Zeljko T, Luchetti A, Mantovani B, Plohl M. 2016. Adjacent sequences disclose potential for intra-genomic dispersal of satellite DNA repeats and suggest a complex network with transposable elements. *BMC Genomics* **17**(1): 997.

Schwarzacher T, Heslop-Harrison P. 2000. *Practical in situ hybridization*. BIOS Scientific Publishers Ltd.

Shaffer LG. 2002. Robertsonian translocations. *Wiley Encyclopedia of Molecular Medicine*.

Stimpson KM, Matheny JE, Sullivan BA. 2012. Dicentric chromosomes: unique models to study centromere function and inactivation. *Chromosome Research* **20**(5): 595-605.

Stimpson KM, Song IY, Jauch A, Holtgreve-Grez H, Hayden KE, Bridger JM, Sullivan BA. 2010. Telomere disruption results in non-random formation of de novo dicentric chromosomes involving acrocentric human chromosomes. *PLoS Genetics* **6**(8): e1001061.

Sullivan BA, Jenkins LS, Karson EM, Leana-Cox J, Schwartz S. 1996. Evidence for structural heterogeneity from molecular cytogenetic analysis of dicentric Robertsonian translocations. *American Journal of Human Genetics* **59**(1): 167.

Sullivan BA, Schwartz S. 1995. Identification of centromeric antigens in dicentric Robertsonian translocations: CENP-C and CENP-E are necessary components of functional centromeres. *Human Molecular Genetics* **4**(12): 2189-2197.

Sullivan BA, Willard HF. 1998. Stable dicentric X chromosomes with two functional centromeres. *Nature Genetics* **20**(3): 227.

Therman E, Susman B, Denniston C. 1989. The nonrandom participation of human acrocentric chromosomes in Robertsonian translocations. *Annals of Human Genetics* **53**(1): 49-65.

Trowell HE, Nagy A, Vissel B, Choo KA. 1993. Long-range analyses of the centromeric regions of human chromosomes 13, 14 and 21: identification of a narrow domain containing two key centromeric DNA elements. *Human Molecular Genetics* **2**(10): 1639-1649.

Vissel B, Nagy A, Choo K. 1992. A satellite III sequence shared by human chromosomes 13, 14, and 21 that is contiguous with α satellite DNA. *Cytogenetic and Genome Research* **61**(2): 81-86.

Waye JS, Willard HF. 1989. Human beta satellite DNA: genomic organization and sequence definition of a class of highly repetitive tandem DNA. *Proceedings of the National Academy of Sciences* **86**(16): 6250-6254.

Wilch ES, Morton CC. 2018. Historical and Clinical Perspectives on Chromosomal Translocations. *Advances in Experimental Medicine and Biology* **1044**: 1-14.

Wolff DJ, Schwartz S. 1992. Characterization of Robertsonian translocations by using fluorescence in situ hybridization. *American Journal of Human Genetics* **50**(1): 174.

## Supplementary Information

**Supplementary Table III.S1. Primer sequences used for SatDNAs genomic isolation and chromosome painting probes amplification.**

| | Forward | Reverse |
|---|---|---|
| **αSat** | TGCAAGTGGATATTTGGACCT | CAAAAAGAGTGTTTCAAACCTGAAC |
| **βSat** | CCTAGAGGCACATTGGGACA | AATGCCCCGTGTAAGCAG |
| **SatI** | ATGTGCGGTACATAAGAT | AAATATGGTTGGGTACTT |
| **SatII** | CCATTCGATTCTTTGCGATG | TCGAATGGAATCATTGAACG |
| **SatIII** | ATCGGAGTGCAGTGGAAAGC | CACTCGATTCCCACTTGCACT |
| **6MW** | CCGACTCGAGNNNNNNATGTGG | |

**Supplementary Table III.S2**. **Number of mapped hits of SatDNAs in each human chromosome(SatI; SatII, Accession Number: X06199.1; SatIII, Accession Number: S90110.1 and X54108.1; αSat; βSat, Accession Number: M81229.1).** Two SatIII subfamilies (pTRS-47 and pTRS-63) were mapped. For SatI one representative obtained clone (200 bp) was mapped.

| Chromosome | Satellite I 200 bp*1 | α Satellite*2 | β Satellite | Satellite II | Satellite III (pTRS-47) | Satellite III (pTRS-63) |
|---|---|---|---|---|---|---|
| **HSA 14 (NC_000014.9)** | 71 | 1392 | 0 | 4 | 0 | 1 |
| **HSA 21 (NC_000021.9)** | 0 | 1427 | 0 | 77 | 28 | 5 |

***1**CCCCTAATGGTTGGGTACTTTCATATTTTATGTACAGTATATAATACATATTTTGGGAACTTTGCTATTTTATGTACAGTATATAATACATACTTTGTGTATTTGATAGTTTATGTAACGTATATAATATATATTTTGGGTATTTTGATATTTTAAGTACAGTATATACTATATAGCATGGGTACTTTGATATCTTATGTACCGCACATG

***2**CTTCTGTCTAGTTTTTATATGAAGATATTCCCGTTTCCAACCAAGGCCTCAAACGGTCCAAATATCCACAAGCTGATTCTACAAAAAGAGTGTTTCAAAACTGCTCTATGAAAAGGAAGGTTCAACTCTGTGAGTTGAATGTATACATCACAAAGAAGTTTCTGAGA

# Chapter IV - General Discussion

The overwhelming structure of the human centromere has been silenced since the Human Genome Project (HGP) in 2003 (Collins et al. 2003) and the apparent completion of the human genome, that in reality left out close to 10% of genomic elements, more specifically large portions of repetitive centromeric sequences. From the centromere to the peripheric pericentromeric sequences, homogeneous α HORs are gradually replaced by monomeric α repeats interspersed with highly heterogenous satellite families and TEs. This complex construction causes for reference genome annotations to become scarcer and more fractional (Black and Giunta 2018).

Satellite repeats arrangement is particularly interesting in acrocentric chromosomes, often lacking specificity in sequence organization but rather sharing a pattern of homology between satellite families (Bandyopadhyay et al. 2001). This pattern could be the answer for the question of unequal frequency of ROBs. Perhaps the recurrent association of chromosome 14 with chromosome 13 and 21 can be explained by sequence homology, as already hypothesized (Choo et al. 1988; Therman et al. 1989; Shaffer 2002). If this is the case, αSat HOR homology can be reasonably extended for other satellite families: αSat HOR arrays are chromosome-specific with the exception of chromosomes 14/22 and 13/21. So, at least chromosomes 13 and 21 seem to have high homology in different satellite families (Kalitsis et al. 1993; McNulty and Sullivan 2018), possibly indicating similar behavior in the context of ROBs. However, despite the presence of an analogous situation with chromosomes 14 and 22 (also sharing homology) (Choo et al. 1990; McNulty and Sullivan 2018), the low statistic incidence of ROBs involving chromosome 22 remains inexplicable with this kind of reasoning. Hence, sequence composition, and even sequencing organization, might not be sole factors for understanding ROBs.

Following physical mapping, the logical question is: what mechanistic steps were required to achieve a stable translocated chromosome? The loss and reorganization of particular SatDNA repeats seem to be determining factors. The clear understanding of the involved sequential steps remains unreachable, since our reasoning still depends on the most probable or parsimonious point of view. This situation proves that cytogenetic mapping approaches are vastly useful in providing information about ROBs and breakpoint uncovering, subjects associated with controversial and gapped information.

Knowing all aspects concerning this rearrangement proves to be a fine instrument for understanding the centromere itself. This chromosomal structure seems to be a core 'hotspot' location for chromosomal rearrangements involving its (peri)centromeric sequences (like ROBs) (Gravholt et al. 1992; Chaves et al. 2004; Adega et al. 2006; Adega et al. 2009; Vieira-da-Silva et al. 2015; Escudeiro et al. 2019). Thus, ROBs can maybe indicate that the centromere is often an encounter between disease and evolution.

In order to achieve better centromere contiguity and overall knowledge, it is essential to follow sequential steps, first consisting in obtaining an accurate map of the centromeric region. In this work, physical mapping allowed to obtain a first draft of satellite order in the context of rob(14;21). Though considered of low-resolution, this approach provided a basis for posterior high-resolution sequencing procedures.

The previously presented SatI study was performed because of the associated information gap. In spite of the assumption that this SatDNA family is the least abundant classical satellite (Tagarro et al. 1994), its pertinence in a variety of circumstances cannot be surpassed, namely in contexts related with (peri)centromeric/ acrocentric short-arm sequences (e.g. ROBs). The former sequence types (SatI included) constitute the greater challenge for the accurate and complete assembly of the human genome (Eichler et al. 2004; Jain et al. 2018a).

Alignment, analysis and annotation of sequence data presupposes the use of a reference genome assembly (Church et al. 2015). As the substrate for the annotation of any sort of new biological information (Schneider et al. 2017), an improved contiguity of the human reference genome is of ultimate importance in the case of repetitive satellite sequences. Regardless of the assumed comprehension and high-quality representation of the GRCh38 (Schneider et al. 2017), the presented *in silico* analysis allowed us to evaluate the presence of satellite families in this reference genome and to infer about their underrepresentation. Thus, nowadays, cytogenetic techniques are still providing helpful and complementary information (both in satellite I presence and in physical satellite mapping in rob(14;21)). If we consider the 'jigsaw puzzle' analogy, where repetitive sequences are "the blue sky in a landscape" for the assembly process (Sedlazeck et al. 2018), maybe the previous physical mapping approaches can be a primary adjuvant for distinguish the sky from the ocean.

The issue with assembly software development is that most tools are designed to ignore repetitive sequences, filtering obtained reads for the absence of annotations in repetitive genomic locations (Li 2014; Miga 2015; Miga 2017). The release of GRCh38 was a significant improvement in the field: the current human genome reference contains a representative

sequence for each centromere (Miga 2017). Yet, this representation was limited to the insertion of millions of bases of α satellite repeats (Miga 2019). From GRCh37 to GRCh38, the gaps representing centromeric sequences were filled by centromere models, only capable of identifying αSat sequences from sequencing reads (Guo et al. 2017), which clearly undermines the proper and proportional representation of repetitive sequences, biasing data interpretation. This situation was clear throughout this work, namely in the analysis of SatI flanking regions and in the *in silico* mapping of satellite families. Between the chromosomes with mappable SatI hits, only HSA8 hits showed to be flanked by TEs. Given the statement that pericentromeres are highly rich in TEs (namely active SINEs and LINEs) (Mills et al. 2007; McKinley and Cheeseman 2016; Black and Giunta 2018), it was expectable to find a more dominant representation. Likewise, the mapping of satellite sequences revealed a larger obtainment of α hits (in comparison to other satellites) instead of a representation compatible with FISH results.

The unfruitful attempt to map βSat in the current human reference genome also points out for its limitations in the complex analysis of Mb-spanning repetitive sequences. Indeed, human chromosome contigs contain rare (or none) βSat annotations (Yang et al. 2019) (also proved true for other SatDNA families).

*De novo* genome assembly would idyllically rely on simply merging maximal overlapping reads. However, with the presence of complex repetitive regions, accurate *de novo* genome assembly must rely on read length and bioinformatic algorithms as well. While assembling a genome, specific satellite-associated gaps arise due to the organization in HOR tandem arrays, causing reads to build up and accumulate in a unresolvable manner (Chaisson et al. 2015). With the uprising of long-read sequencing (PacBio and Oxford Nanopore), whole sections of repetitive DNA analysis became more accessible: HOR structure, TE interruption (Sevim et al. 2016) or the accurate determination of monomer size (Cacheux et al. 2016). The overall centromeric sequence composition is finally within reach in a variety of scenarios. The attainment of an improved reference genome can lead to more recurrent annotations and a comprehensive analysis in the setting of clinical causal variation (Chaisson et al. 2015). In the case of ROBs, breakpoint and derivative sequences examination at base pair resolution can have a great impact.

Nanopore sequencing was already used to assemble repetitive sequences (Jain et al. 2018a; Jain et al. 2018b) or determine breakpoints in chromosomal translocations (Dutta et al. 2018; Hu et al. 2018). Full harnessing of this technology (namely in what concerns data interpretation) should be allied to a large collection of mapping information, for example to direct read analysis

to specific genomic locations (Dutta et al. 2018) (physical and *in silico* mapping). Approaching repetitive sequences with nanopore sequencing can be a decisive alternative to overcome short-read-associated misassemblies, misalignments and, more precisely, short-range organization.

## References

Adega F, Chaves R, Guedes-Pinto H, Heslop-Harrison J. 2006. Physical organization of the 1.709 satellite IV DNA family in Bovini and Tragelaphini tribes of the Bovidae: sequence and chromosomal evolution. *Cytogenetic and Genome Research* **114**(2): 140-146.

Adega F, Guedes-Pinto H, Chaves R. 2009. Satellite DNA in the karyotype evolution of domestic animals–clinical considerations. *Cytogenetic and Genome Research* **126**(1-2): 12-20.

Bandyopadhyay R, McQuillan C, Page S, Choo K, Shaffer L. 2001. Identification and characterization of satellite III subfamilies to the acrocentric chromosomes. *Chromosome Research* **9**(3): 223-233.

Black EM, Giunta S. 2018. Repetitive Fragile Sites: Centromere Satellite DNA As a Source of Genome Instability in Human Diseases. *Genes* **9**(12).

Cacheux L, Ponger L, Gerbault-Seureau M, Richard FA, Escudé C. 2016. Diversity and distribution of alpha satellite DNA in the genome of an Old World monkey: Cercopithecus solatus. *BMC Genomics* **17**(1): 916.

Chaisson MJ, Wilson RK, Eichler EE. 2015. Genetic variation and the de novo assembly of human genomes. *Nature Reviews Genetics* **16**(11): 627.

Chaves R, Santos S, Guedes-Pinto H. 2004. Comparative analysis (Hippotragini versus Caprini, Bovidae) of X-chromosome's constitutive heterochromatin by in situ restriction endonuclease digestion: X-chromosome constitutive heterochromatin evolution. *Genetica* **121**(3): 315-325.

Choo K, Earle E, McQuillan C. 1990. A homologous subfamily of satellite III DNA on human chromosomes 14 and 22. *Nucleic Acids Research* **18**(19): 5641-5648.

Choo K, Vissel B, Brown R, Filby R, Earle E. 1988. Homologous alpha satellite sequences on human acrocentric chromosomes with selectivity for chromosomes 13, 14 and 21: implications for recombination between nonhomologues and Robertsonian translocations. *Nucleic Acids Research* **16**(4): 1273-1284.

Church DM, Schneider VA, Steinberg KM, Schatz MC, Quinlan AR, Chin C-S, Kitts PA, Aken B, Marth GT, Hoffman MM. 2015. Extending reference assembly models. *Genome Biology* **16**(1): 13.

Collins FS, Green ED, Guttmacher AE, Guyer MS. 2003. A vision for the future of genomics research. *Nature* **422**(6934): 835.

Dutta UR, Rao SN, Pidugu VK, Vineeth V, Bhattacherjee A, Bhowmik AD, Ramaswamy SK, Singh KG, Dalal A. 2018. Breakpoint mapping of a novel de novo translocation t (X; 20)(q11. 1; p13) by positional cloning and long read sequencing. *Genomics*.

Eichler EE, Clark RA, She X. 2004. An assessment of the sequence gaps: unfinished business in a finished human genome. *Nature Reviews Genetics* **5**(5): 345.

Escudeiro A, Adega F, Robinson TJ, Heslop-Harrison JS, Chaves R. 2019. Conservation, divergence and functions of centromeric satellite DNA families in the Bovidae. *Genome Biology and Evolution*.

Gravholt CH, Friedrich U, Caprani M, Jørgensen AL. 1992. Breakpoints in Robertsonian translocations are localized to satellite III DNA by fluorescence in situ hybridization. *Genomics* **14**(4): 924-930.

Guo Y, Dai Y, Yu H, Zhao S, Samuels DC, Shyr Y. 2017. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics* **109**(2): 83-90.

Hu L, Liang F, Cheng D, Zhang Z, Yu G, Zha J, Wang Y, Wang F, Tan Y, Wang D. 2018. Localization of balanced chromosome translocation breakpoints by long-read sequencing on the Oxford Nanopore platform. *bioRxiv*: 419531.

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT. 2018a. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* **36**(4): 338.

Jain M, Olsen HE, Turner DJ, Stoddart D, Bulazel KV, Paten B, Haussler D, Willard HF, Akeson M, Miga KH. 2018b. Linear assembly of a human centromere on the Y chromosome. *Nature Biotechnology* **36**(4): 321-323.

Kalitsis P, Earle E, Vissel B, Shaffer LG, Choo KA. 1993. A chromosome 13-specific human satellite I DNA subfamily with minor presence on chromosome 21: further studies on Robertsonian translocations. *Genomics* **16**(1): 104-112.

Li H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**(20): 2843-2851.

McKinley KL, Cheeseman IM. 2016. The molecular basis for centromere identity and function. *Nature Reviews Molecular Cell Biology* **17**(1): 16.

McNulty SM, Sullivan BA. 2018. Alpha satellite DNA biology: finding function in the recesses of the genome. *Chromosome Research* **26**(3): 115-138.

Miga KH. 2015. Completing the human genome: the progress and challenge of satellite DNA assembly. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* **23**(3): 421-426.

Miga KH. 2017. The Promises and Challenges of Genomic Studies of Human Centromeres. *Progress in Molecular and Subcellular Biology* **56**: 285-304.

Miga KH. 2019. Centromeric Satellite DNAs: Hidden Sequence Variation in the Human Population. *Genes* **10**(5): 352.

Mills RE, Bennett EA, Iskow RC, Devine SE. 2007. Which transposable elements are active in the human genome? *Trends in Genetics* **23**(4): 183-191.

Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D. 2017. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research* **27**(5): 849-864.

Sedlazeck FJ, Lee H, Darby CA, Schatz MC. 2018. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics* **19**(6): 329.

Sevim V, Bashir A, Chin C-S, Miga KH. 2016. Alpha-CENTAURI: assessing novel centromeric repeat sequence variation with long read sequencing. *Bioinformatics* **32**(13): 1921-1924.

Shaffer LG. 2002. Robertsonian translocations. *Wiley Encyclopedia of Molecular Medicine*.

Tagarro I, Wiegant J, Raap AK, González-Aguilera JJ, Fernández-Peralta AM. 1994. Assignment of human satellite 1 DNA as revealed by fluorescent in situ hybridization with oligonucleotides. *Human Genetics* **93**(2): 125-128.

Therman E, Susman B, Denniston C. 1989. The nonrandom participation of human acrocentric chromosomes in Robertsonian translocations. *Annals of Human Genetics* **53**(1): 49-65.

Vieira-da-Silva A, Louzada S, Adega F, Chaves R. 2015. A high-resolution comparative chromosome map of Cricetus cricetus and Peromyscus eremicus reveals the involvement of constitutive heterochromatin in breakpoint regions. *Cytogenetic and genome Research* **145**(1): 59-67.

Yang J, Zhou Y, Ma G, Zhang X, Guo Y. 2019. Parasites Acquired Beta Satellite DNAs from Hominid Hosts via Horizontal Gene Transfer. *bioRxiv*: 58953

# Chapter V - Conclusion and Future Perspectives

Acrocentric short-arm and (peri)centromeric sequences have long been a challenging and complex task to approach, especially using single strategies relying on short-read sequencing technologies that lead to a deprived or equivocal mapping of repetitive sequences (Sedlazeck et al. 2018). In order to fully address these genomic regions with recent long-read technologies, sequential steps must be tactically followed, from lower resolution techniques to higher ones, until the accomplishment of the striking base pair resolution. This work trails between the first base-providing mapping strategies.

SatDNA families should be more intensively studied, namely the ones with limited information (like SatI). Physical mapping by FISH techniques can be a possibility to analyze these tandemly repeated arrays of pericentromeric heterochromatin, as presented above. However informative, hybridization signals can be puzzling to determine in the case of close satellite arrays. FISH experiments with extended DNA fibers (Shiels et al. 1997) or with specific pericentromeric BAC clones (Jarmuz-Szymczak et al. 2014) constitute possible alternatives to consider hereafter towards the enrichment of satellite physical mapping. Optical mapping techniques from BioNano Genomics can also be used in the future to fingerprint megabase-long genomic regions and for Structural Variants (SVs) detection with higher resolution (Sedlazeck et al. 2018). Multiple alignments of diverse genome maps (obtained by optical mapping) can likewise be suitable for the disclosure of new information (absent in GRCh38 human reference genome) about acrocentric short-arm sequences (Levy-Sakin et al. 2019).

The future should guarantee that short-read-based assemblies are progressively complemented with long-read unfragmented sequencing data, namely from nanopore sequencing. However, improvements are on the move and still in order. Base-calling algorithms and long-read mapping/alignment software tools must be subject to constant development (Bowden et al. 2019). The obtainment of a better human reference genome has to come from the application of technologies like nanopore sequencing, as the currently available assembly does not allow for an accurate *in silico* mapping. Unquestionably, the authentic representation of all sequences and sequence types is crucial for annotating and understanding biological knowledge. One thing we know is that repetitive sequences are holding future research promises: the process of filling the assembly gaps, the deep understanding of their relation with clinical contexts like ROBs or the ultimate achievement of telemore-to-telemore sequence.

## References

Bowden R, Davies RW, Heger A, Pagnamenta AT, de Cesare M, Oikkonen LE, Parkes D, Freeman C, Dhalla F, Patel SY. 2019. Sequencing of human genomes with nanopore technology. *Nature Communications* **10**(1): 1869.

Jarmuz-Szymczak M, Janiszewska J, Szyfter K, Shaffer LG. 2014. Narrowing the localization of the region breakpoint in most frequent Robertsonian translocations. *Chromosome Research* **22**(4): 517-532.

Levy-Sakin M, Pastor S, Mostovoy Y, Li L, Leung AK, McCaffrey J, Young E, Lam ET, Hastie AR, Wong KH. 2019. Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nature Communications* **10**(1): 1025.

Sedlazeck FJ, Lee H, Darby CA, Schatz MC. 2018. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics* **19**(6): 329.

Shiels C, Coutelle C, Huxley C. 1997. Analysis of ribosomal and alphoid repetitive DNA by fiber-FISH. *Cytogenetic and Genome Research* **76**(1-2): 20-22.