Investigating DNA sequences of *Drosophila melanogaster* centromeres

Bianca Mocanu

Exit document for Masters of Science in Genetics and Genomics

Abstract

The even distribution of genetic material during cell division is strongly dependent upon the correct assembly of a special chromosomal region known as the centromere. The centromeres are epigenetically determined by a special nucleosome containing a histone H3 variant (CenH3) which is both necessary and sufficient for centromere function. Despite its conserved role in accurate chromosome segregation, the centromere and its components are subjected to rapid evolution in many organisms, possibly driven by genetic conflicts during the asymmetric process of ovogenesis. The repetitive nature of the underlying DNA of most centromeres renders them difficult to sequence and conceals valuable information regarding their structure and function. The current document iterates the main hypotheses about centromeres and findings that support them, with a focus on *Drosophila melanogaster* . Latest efforts to sequence the fruit fly centromeres and to allocate them to the respective X,2,3,4 and Y chromosomes reveal several candidate contigs displaying a core of CenH3 binding, relatively complex DNA immersed in simple satellite sequences of yet unknown size.

Major advisor: Prof. Dr. Barbara Mellone

Associate advisor: Prof. Dr. Rachel O'Neill

Associate advisor: Prof. Dr. Leighton Core

April 15, 2018

STATUTORY DECLARATION

With my signature, I certify that this document has been written by me using only the indicated resources and materials. I further confirm that this document has not been submitted, either in part or as a whole, for any other academic degree at this or another institution.

Bianca Mocanu Storrs, Connecticut, USA

April 15, 2018

1 INTRODUCTION

Centromeres are essential chromosomal structures that act as a platform for kinetochore assembly and microtubule attachment, ensuring accurate chromosome segregation [1]. In almost all studied eukaryotes, centromeres are epigenetically defined by a particular centromeric nucleosome which contains a histone H3 variant (CenH3). Despite the crucial and invariable role the centromere conducts in the cell division process, there is great variation between the centromeric DNA sequences of closely related organisms and, often, variations between the centromeric sequences of different chromosomes within the same species. Across Eukarya, centromeric DNA size varies from as little as 125 base pairs in *S. cerevisiae*'s 'point' centromere to megabases in humans [2].

Centromere research is particularly challenging since any interference with the chromosome segregation process results in severe genomic defects and general genomic instability [3]. Furthermore, the lack of sequence conservation is topped by the highly repetitive nature of the underlying DNA for most (but not all) centromeres of organisms with monokinetic chromosomes (i.e. possessing a single centromere per chromosome). Although there are big technical challenges in obtaining a reliable linear genetic map, some insight has been gained into the arrangement of the repetitive DNA of various centromeres. Currently there is no general pattern observed among different centromeres, as each organism appears to have a different arrangement. This observation adds even more complexity to centromere organization and calls for efforts to resolve linear genetic maps for each centromere of each organism separately [4].

Due to their condensed nature, it has been long assumed that centromeres as a whole were transcriptionally inactive and displayed a heterochromatin state. However, in order for centromeres to properly function, a finely tuned level of transcription seems to be required. In budding yeast, altering the transcription levels at the centromere in either direction results in chromosome segregation defects [5] . Likewise, in synthetic human artificial chromosomes (HACs), silencing or strong upregulation of transcription both impair centromere function by preventing CenH3 deposition or by causing its rapid removal, respectively [6]. All things considered, one hypothesis is that low transcription at centromeric sites could provide the ideal chromatin environment to either load the centromeric histone de novo during centromere establishment or to replenish the centromeric histone following dilution after S phase / cell division.

Until recently, it has been difficult to prove a causation between transcription and deposition of centromeric histones, for several reasons. First, CenH3 replenishment happens in the context of it already being present at the endogenous centromeres whose centromeric nucleosomes are maintained just like any epigenetic mark. Secondly, from a bioinformatics perspective, the lack of centromeric sequence assemblies prompts algorithms to discard unaligned transcripts, when the latest genome reference is used. Work on *Drosophila melanogaster* in the Mellone lab [7] bypassed some of these issues through the use of an artificial, inducible, ectopic centromere assembled on an external array of LacO repeats and, therefore, of known sequence. By fusing a CenH3 chaperone to LacI, they were able to bring CenH3 in physical proximity to the LacO array and correlate fluctuations in transcription with deposited centromeric histones at the array and sites bordering the LacO array [7].

This is, of course, only one of countless cases in which *Drosophila melanogaster* proved to be an excellent model organism. Thanks to its genetic similarity to humans and thanks to an army of genetic tools developed over the past century, *Drosophila* is a great model in the context of centromere research as well. Having only 4 monokinetic chromosomes and a relatively small genome size, it is a good candidate in obtaining complete genetic assemblies and maps of all its centromeres. Without a proper assembly of centromeric sequences, it is only possible to examine e.g. transcription in a mapping-free approach, as there is no reliable reference to align this data to. This also makes it difficult to investigate and identify new centromeric features such as epialleles or rare variants of satellites and transposons and to accurately estimate their abundance. Metastable epialleles have been recently described in humans, yet it is unknown whether this is a common feature of repetitive centromeres [8]. Moreover, increasing evidence suggests that centromeric sequences such as the dodeca satellite in *Drosophila melanogaster* may have been selected not by their primary sequence, but by their ability to form noncanonical secondary structures which ultimately orchestrate the interaction of the DNA with its protein partners [9]. Having assembled centromeres and the order in which satellites and transposons occur might help speculate about their functions in the underlying DNA. Despite the recent advances in genomics and ultra long read sequencing, contiguous sequences for such highly repetitive structures are yet to be elucidated.

2 EPIGENETIC ASPECTS OF CENTROMERES

Due to the dissimilarity of the centromeres, defining boundaries for them is not trivial. One epigenetic mark for the centromere, valid across all centromeres – regardless of size or localization – is the centromeric histone variant. Centromeric histones, denoted CenH3s, are variants of the Histone H3 family of proteins. These variants are necessary and suficient for centromeric activity and essential for accurate chromosome segregation. In all the eukaryotic organisms examined so far, there is a single gene encoding CenH3 but, unlike canonical histones, they have a high degree of sequence variability even among closely related species [10].

For example, in the case of closely related *Drosophila* species, the centromeric histones appear to be positively selected. Within the CenH3 protein itself, the selection is localized to the N-terminal tail and to Loop 1 of the histone fold domain – both with significant functional importance in targeting of CenH3 to centromeric DNA. Such sequence variation among CenH3s suggests that there is likely a DNA-binding specificity factor that drives the positive selection for this gene [3]. Outside *Drosophila* genus, similar selection patterns are observed for the well studied plant model *Arabidopsis thaliana* [11].

The positive selection and rapid evolution is not only observed for CenH3 proteins. In cases where CenH3 positive selection is not as accentuated, such as in humans and in other mammals, the CenH3 binding, inner kinetochore protein CENP-C is reported to be more rapidly evolving [11]. In flies, in addition to the positive selection imposed on CenH3, its assembly factor CAL1 has been proved to be co-evolving with CenH3 [12]. A series of transfection experiments performed by the Mellone Lab reveal that while 'foreign', exogenous CenH3 of distant *Drosophila* species are unable to be targeted to the centromeric DNA of *Drosophila melanogaster* alone, they are able to be rescued if their corresponding 'foreign' CAL1 is also co-expressed [12]. These observations of rapid evolution in centromeric DNA and centromeric proteins are indicative of a constant intragenomic competition that drives the changes in these centromeric components.

2.1 THE MEIOTIC DRIVE HYPOTHESIS

The variability of centromeric DNA and of centromeric histones seems paradoxical. Due to the crucial and universal role it serves in the process of cell division, one would expect the centromere

as a whole to be under a strong evolutionary constraint. One hypothesis named "meiotic drive" aims to elegantly explain this paradox by highlighting one genetic conflict the centromeres are found in during the asymmetric meiosis events of oogenesis.

The reduction division of meiosis I involves the separation of homologous chromosomes of a diploid cell and yields, in theory, two haploid secondary oocytes. However, the asymmetric spindle of meiosis I and subsequent uneven distribution of cytoplasm during cytokinesis most often lead to the formation of only one secondary oocyte and one smaller product that does not develop into an ovum, called 'polar body'. Due to this, one hypothesis is that chromosomes try to favor their own transmission to the egg rather than being lost in polar bodies and they do so by developing 'stronger' centromeres [13]. Currently, several experiments and observations in mice and *Drosophila* validate the meiotic drive hypothesis and further describe 'strong' centromeres as those that bind more centromeric and kinetochore proteins and that detach from microtubules more easily, most often in regions of the cytoplasm that maximize the likelihood of them being retained in the egg [11]. Such an example is depicted in Figure 1.

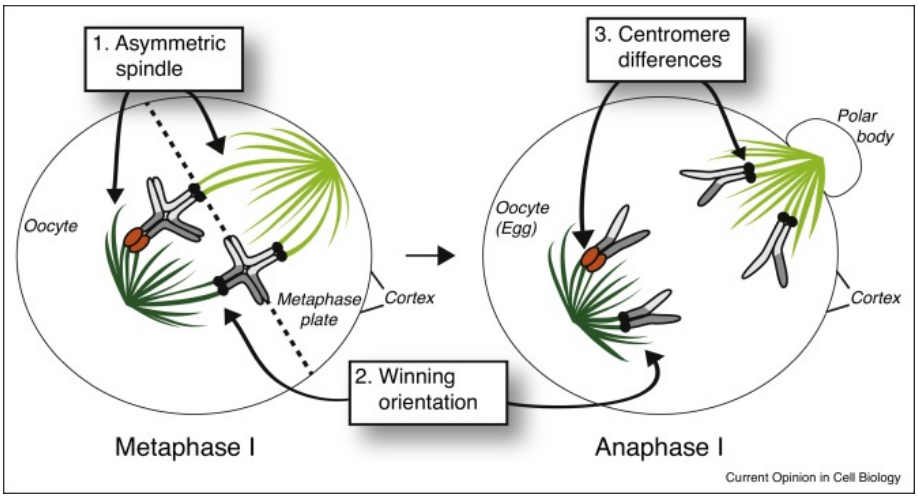

Figure 1. Model for meiotic drive in mice [14]. Two pairs of acrocentric homologous chromosomes are depicted segregating during an asymmetric meiosis 1 event. The asymmetry of the spinde is highlighted by different microtubule shades of green. Chromosome orientation towards the center of the cell is deemed as 'winning', as is has the maximum probability of being distributed to the oocyte during cytokinesis. The upper homolog pair exhibits centromeric heterozygosity in the sense that one homolog has 'stronger' centromeres which favor its transmission to the egg.

Mechanistically, one way to achieve this directly involves centromeric DNA and is backed by comparisons of mice strains with different levels of microtubule-binding kinetochore proteins such as HEC1 at centromeres. The 'strength' of the centromeres, when defined by the HEC1 protein levels, is directly correlated with the enrichment of the minor satellite (the most abundant DNA component of mouse centromeres) in CenH3 native chromatin immunoprecipitation experiments [14]. If the presence of more minor satellite results in binding of more CenH3 and subsequent strengthening of the centromere, this highlights a potential role for the underlying DNA and also speculates that the centromeric histones could, in fact, mutate in response, in an effort to cancel out this advantage.

2.2 EPIGENETIC DEFINITION SHORTCOMINGS

The meiotic drive hypothesis challenges the opinion that centromeres of most organisms are solely epigenetically defined. Despite being robust, the epigenetic definition of the centromere fails to explain the contribution of the underlying DNA sequences. In addition to the rapid evolution of the centromeric proteins that suggests a (yet unidentified) preference for the bound DNA, there are more aspects that further challenge this current opinion.

Apart from the sequence-defined point centromere of the budding yeast, there are two more genetic ways in which the centromeres could be specified: through proteins that bind sequence-specific DNA and through other sequence-dictated features such as DNA secondary structures – stem loops, hairpins, triplexes, commonly referred to as non-B-form DNA. [15]

So far, the only identified protein that binds centromeres in a sequence-specific manner is the CENP-B protein which is highly conserved in mammals. Paradoxically, the CENP-B DNA binding motifs, called CENP-B boxes, are not present in all mammalian genomes and, as is the case for the Y chromosome in humans, some chromosomes within a genome with CENP-B boxes actually lack them [15]. Transfection studies from as far back as 1997 using yeast artificial chromosome (YAC) systems with cloned human centromeric satellite arrays aimed to explain this paradox and confirmed that arrays with CENP-B boxes can form de novo HACs upon transfection of YACs into human cells, while those with mutated CENP-B boxes cannot. The presence of CENP-B boxes in humans seemed required for de novo centromere assembly, despite the fact that the Y chromosome alone is capable of assembling a functional centromere in vivo [16] [17].

One study examining vertebrates, fission yeast, and plants investigated the correlation between the abundance of DNA dyad symmetries – energetically favored to adopt non-B-form DNA conformations – and the prevalence of CENP-B boxes within centromeric satellite sequences. Starting with publicly available Illumina whole genome sequencing data, they retrieved sequences poorly represented in the genome assemblies and showed that great apes and mouse centromeres were depleted of dyads, as opposed to old world monkeys, yeast and chicken. By scanning the centromeres for CENP-B binding motifs, they were able to show that there is a negative correlation between CENP-B binding motifs and dyad symmetries. This observation holds true not only among various organisms but also for the case of human centromere Y, which, as opposed to the other human centromeres, lacks CENP-B motifs and is predicted to form non-B DNA structures more readily. In their discussion, the authors bring into picture HJURP, the human centromeric histone assembly factor initially named because of its cruciform DNA binding preferences and speculate how, in dyad-depleted organisms, CENP-B can facilitate the structural transition necessary for cruciform formation, HJURP binding and ultimately centromeric histone deposition. They also point out that in organisms lacking CENP-B motifs, the centromeric transcripts originating from the non-B form DNA conformations might adopt secondary structures themselves without external aid. The study, however, leaves out *Drosophila melanogaster* centromeres [15].

3 DROSOPHILA SEQUENCING AND ASSEMBLY EFFORTS

Having a complete DNA sequence assembly for the centromeres of *Drosophila* would be of great use, in order to understand their structure, characterize them and contribute to the validation of some hypotheses that currently concern centromeres. Recent technologies such as Pacific Biosciences and Oxford Nanopore allow for the generation of ultra long reads originating from single molecules, with overall base calling accuracy as high as 85%. These long reads can then serve as a scaffold for and be complemented by numerous, more accurate short reads generated for example by Illumina from whole genome sequences [18].

Over time, the centromeric sequence puzzle has been gradually addressed as multiple sequencing and assembly methods have been attempted.

3.1 DECIPHERING HETEROCHROMATIN

Emergence of shotgun sequenced genomes called for heterochromatin sequencing projects in order to understand the essential functions and genes residing in what makes up, for example in *Drosophila*, a good third of the genome. Before this, most progress had been made through sequencing individual cloned transposons, genes and satellite DNA repeats and determining their location with very low resolution through classical methods such as in situ hybridization of mitotic chromosomes [19].

In the 1990s and early 2000s, in fruit flies, one approach made use of the 1.3 Mb *Dp1187* minichromosome and several of its deletion derivatives generated through irradiation, as a way to gain access to difficult parts of heterochromatin. *Dp1187* heterochromatin was found to contain several complex DNA 'islands' immersed in simple satellite repeats. Analysis of its deletion derivatives identified the minimum centric DNA necessary for this minichromosome to be maintained through meiosis and mitosis and estimated it to be about 420 kb in size [19] (Figure 2). Sequencing and assembly efforts directed towards this minimal 420 kb sequence proved to be somewhat successful through restriction mapping analyses and generation of a 'centromere-enriched' library with 459 clones. The sequencing of 459 clones summed up to 947.9 kb of reads, of which little over half (495.6 kb) were from the centromeric region. The assembly of the centromeric reads yielded 13 contigs that spanned 19 % of the targeted centromere (79.9 kb) [19]. This *Drosophila* minichromosome approach highlighted that centromeric DNA could be sequenced, despite being a laborious process, and gave hope that other similar approaches could be used to resolve naturally occurring centromeres, too.

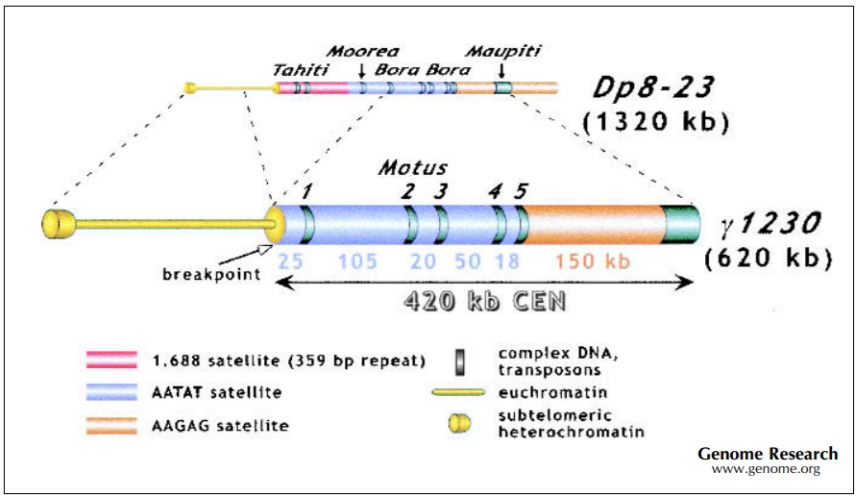

Figure 2. Schematic of the *Dp1187* minichromosome (here named *Dp8-23*) and its radiation-generated deletion derivative *γ1230*. Main features of the minichromosome are euchromatin, subtelomeric heterochromatin, the 359 base pairs satellite and 5 identified 'islands' of complex DNA named Tahiti, Moorea, Bora Bora and Maupiti inserted into AATAT and AAGAG simple satellite blocks. The breakpoints that gave rise to the 620 kb γ1230 derivative are shown. The 420 kb functional centromere was established by examining the transmission of multiple derivatives through mitosis and meiosis.

In 2007, the *Drosophila melanogaster* genome still had low contiguity and quality in the heterochromatic regions of all chromosomes, due to the major challenges regarding cloning and assembly of simple sequence repeats, transposable elements and rDNA repeats. Generally, the shorter the repeats in question (e.g. fragments of transposable elements, satellites), the lower the chance to obtain a good assembly at that locus because it's more difficult to assign reads to a specific repeat copy. Compared to the euchromatic region, heterochromatic scaffolds had 5.8 times as many sequence gaps per Mb, as well as lower sequence quality [20].

The strategy to improve the heterochromatin assemblies involved a 10 kb clone library spanning the genome 15 times which was used to polish the individual low quality regions. For the higher order assemblies, BAC-based sequencing and physical mapping has been used. A total of 15 Mb of heterochromatic assembled scaffolds revealed clusters of complete and incomplete copies of transposable elements but were still inconclusive with respect to simple repeats or satellites [20].

3.2 ULTRA LONG READ SEQUENCING MAY BE THE ANSWER

In 2009, a novel kind of experiment used a transposon-based approach to sequence highly repetitive BAC clones. The general principle is to generate a library through a transposon random insertion step and physically map the precise position of the transposon, which allows to read up to a certain limit the bases around the transposon. This strategy allowed to assemble the Y chromosome of *Drosophila melanogaster* and reveal the telomeric nature of the repeats found within the Y chromosome centromere. While this is a feasible strategy to sequence a BAC containing a highly repetitive insert, sequencing a whole BAC library exponentially increases the time taken to resolve these repetitive regions [21].

Advances in technology such as Oxford Nanopore, however, now allow sequencing of entire BAC libraries such as the one obtained using the transposon approach mentioned above [21]. A very recent report from April 2018 reports obtaining a linear assembly of the human Y centromere from a BAC library using nanopore long-read sequencing to produce high-quality reads that span hundreds of kilobases of highly repetitive DNA. The high-quality BAC consensus sequences were generated by multiple sequence alignment of 60 full-length 190kb BAC reads and were further 'polished' with Illumina data from the same BACs. Following polishing, an assembly was attempted with all the BAC sequences and the outcome spanned the entire centromere, connecting reads from the p and the q arm of the human Y centromere [22]. Even though Nanopore technology makes sequencing through repetitive DNA much simpler, the validation issue and allocating contigs to different centromeres still remains. The human Y centromere success is partially owed to having a good starting BAC library, well-characterized high order repeats structures, as well as extensive work on previous mapping data [22]. Finding out which of these sequences are bound by CenH3 is another intriguing question.

For *Drosophila melanogaster* , the current strategy employs single molecule real time sequencing technology developed by Pacific Biosciences. Initially aimed to resolve autosomal complex satellite DNA [23], the computational approaches have been redirected to be used with CenH3 ChIP-seq data. Several candidate contigs revealed either patterns that match the complex DNA immersed in simple satellite described for minichromosome Dp1187 (Section 3.1, Figure 2) or displayed long signals indicating stretches of consistent CenH3 bound sequences (Figure 3) or both. Several identified transposable elements known to be heterochromatic or centromeric are another indicative

of a good contig candidate for centromeric sequences, and quantitative PCR experiments confirm that these transposable elements are enriched in the CenH3 ChIP samples as compared to the experimental input [24].

However, the same issues as before still stand – allocating these contigs to *Drosophila* chromosomes proves to be a daunting task, since chromosome painting technologies such as Oligopaint, as well as satellite DNA FISH experiments by themselves have a hard time distinguishing the X, 2 and 4 centromeres from each other [24]. Once the contigs are established, the challenge consists in confidently allocating them based on low-copy sequence variants that discriminate the contigs from each other. The level of stringency with which the in situ hybridizations are performed largely affect the outcome of the methods. Too low stringency results in signal across most centromeres, as there are many elements in the hybridization probes that bind to all of them. Increased stringency most likely causes a decrease in the intensity of the signals (as less probe binds). Detection of low signal may require high sensitivity of detection, which in turn means high noise.

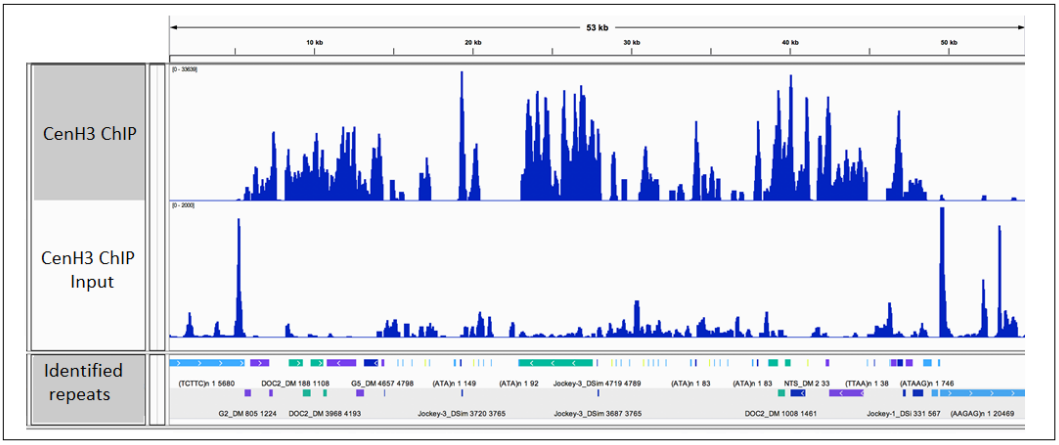

Figure 3. Browser shot of mapped Illumina data from a *Drosophila melanogaster* embryo CenH3 ChIP experiment [24]. Displayed here is the data for the immunoprecipitated sample as well as the DNA input, both in the form of blue peaks indicating regions thought to be bound by CenH3. Identified repetitive elements are depicted at the bottom, as reported by RepeatMasker 4.0.7 configured with the latest RepBase database. The data was mapped to a custom genome containing the latest *Drosophila melanogaster* version to which 173 candidate contigs were catenated. This shot displays a 53 kb portion of a candidate contig for the centromere of chromosome X and its relatively complex DNA with AAGAG satellite repeats at both ends. The peaks from CenH3 ChIP and Input samples are not normalized to each other in order to highlight the uneven signal, possibly attributed to sonication biases and multiple mapping difficulties.

4 CONCLUSIONS AND FUTURE RESEARCH

Since the first description of primary constrictions of mitotic chromosomes in the 1880s [25], centromere research has come a long way. Before the genome projects, many satellites had been independently described and cytogenetically mapped to these primary constrictions [19]. Later, the study of the alpha satellite of humans set the ground for a series of functional assays to determine if live cells are able to recognize such sequences and to assemble centromeres de novo [16] [17] [18]. As the higher order organization of centromeres of humans and other organisms came into light, new challenges aimed to obtain sequence assemblies for centromeres of individual chromosomes in humans and some other well studied model organisms such as *Drosophila melanogaster* , *Schizosaccharomyces pombe* or *Arabidopsis thaliana*.

In 2018, the third generation sequencing techniques such as those developed by Oxford Nanopore and Pacific Biosciences offer a promising future where reliable linear genetic maps connect the p and q arm sequences for entire sets of chromosomes. Development of standard validation methods for such linear assemblies would be of immense value and could, soon enough, allow for sequencing of multiple genomes of different individuals or strains. With it, new functional assays and types of analyses could correlate satellite variation or higher order centromeric structure deviations to diseases involving abnormal chromosome segregation or even to infertility cases.

Of course, it will be very long before personalized medicine can help answer questions about the latest findings in centromere research. In the meantime, basic research got a big boost with the development and perfection of the ultra long read sequencing. Other features, such as genes or noncoding elements yet have to be investigated. For example, masking out repeats in *Drosophila melanogaster* candidate contigs reveal single-copy genes such as klh10 encoding for Kelch-Like Protein 10, whose defective mutants are reported to influence male flies' fertility. It is unknown whether the centromeric chromatin environment is required for the proper functioning of this gene in *Drosophila*, which could very well be the case, given that many genes in the drosophila genome actually reside in the heterochromatin [18].

References

[1] McKinley KL and Cheeseman IM. The molecular basis for centromere identity and function. *Nature Reviews Molecular Cell Biology*, 2015.

[2] Cottarel G, Shero JH, Hieter P, and Hegemann JH. A 125-base-pair CEN6 DNA fragment is sufficient for complete meiotic and mitotic centromere functions in *Saccharomyces cerevisiae*. *Molecular Cell Biology*, 1989.

[3] Rosin LF and Mellone BG. Centromeres drive a hard bargain. *Cell. Trends in Genetics*, 2017.

[4] Zachary Duda, Sarah Trusiak, and Rachel O'Neill. *Centromere Transcription: Means and Motive*, pages 257–281. Springer International Publishing, Cham, 2017.

[5] Ohkuni K and Kitagawa K. Endogenous transcription at the centromere facilitates centromere activity in budding yeast. *Current Biology*, 2011.

[6] Bergmann JH, Jakubsche JN, Martins NM, Kagansky A, Nakano M, Kimura H, Kelly DA, Turner BM, Masumoto H, Larionov V, and Earnshaw WC. Epigenetic engineering: histone H3K9 acetylation is compatible with kinetochore structure and function. *Journal of Cell Science*, 2012.

[7] Chen CC, Bowers S, Lipinszki Z, Palladino J, Trusiak S, Bettini E, Rosin L, Przewloka MR, Glover DM, O'Neill RJ, and Mellone BG. Establishment of centromeric chromatin by the CENP-A assembly factor CAL1 requires FACT-mediated transcription. *Dev. Cell*, 2015.

[8] Aldrup-MacDonald ME, Kuo ME, Sullivan LL, Chew K, and Sullivan BA. Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles. *Genome Research*, 2016.

[9] Garaviś M, Meńdez-Lago M, Gabelica V, Whitehead SL, Gonzaĺez C, and Villasante A. The structure of an endogenous *Drosophila* centromere reveals the prevalence of tandemly repeated sequences able to form i-motifs. *Scientific Reports*, 2015.

[10] Wiens GR and Sorger PK. Centromeric chromatin and epigenetic effects in kinetochore assembly. *Cell*, 1998.

[11] Harmit S. Malik. *The Centromere-Drive Hypothesis: A Simple Basis for Centromere Complexity*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[12] Rosin L and Mellone BG. Co-evolving CENP-A and CAL1 domains mediate centromeric CENP-A deposition across *Drosophila* species. *Dev. Cell*, 2016.

[13] Lindholm AK, Dyer KA, Firman RC, Fishman L, Forstmeier W, Holman L, Johannesson H, Knief U, Kokko H, Larracuente AM, Manser A, Montchamp-Moreau C, Petrosyan VG, Pomiankowski A, Presgraves DC, Safronova LD, Sutter A, Unckless RL, Verspoor RL, Wedell N, Wilkinson GS, and Price TAR. The ecology and evolutionary dynamics of meiotic drive. *Trends in Ecology and Evolution*, 2016.

[14] Iwata-Otsubo A, Dawicki-McKenna JM, Akera T, Falk SJ, Chmatal L, Yang K, Sullivan BA, Schultz RM, Lampson MA, and Black BE. Expanded satellite repeats amplify a discrete CENP-A nucleosome assembly site on chromosomes that drive in female meiosis. *Curr. Biol.*, 2017.

[15] Kasinathan S and Henikoff S. Non-B-form DNA is enriched at centromeres. *Molecular Biology and Evolution*, 2018.

[16] Harrington JJ, Van Bokkelen G, Mays RW, Gustashaw K, and Willard HF. Formation of de novo centromeres and construction of first-generation human artificial microchromosomes. *Nature Genetics*, 1997.

[17] Masumoto H, Ikeno M, Nakano M, Okazaki T, Grimes B, Cooke H, and Suzuki N. Assay of centromere function using a human artificial chromosome. *Chromosoma*, 1998.

[18] Aldrup-MacDonald ME and Sullivan BA. The past, present, and future of human centromere genomics. *Genes*, 2014.

[19] Sun X, Le HD, Wahlstrom JM, and Karpen GH. Sequence analysis of a functional *Drosophila* centromere. *Genome Research*, 2003.

[20] Hoskins RA, Carlson JW, Kennedy C, Acevedo D, Evans-Holm M, Frise E, Wan KH, Park S, Meńdez-Lago M, Rossi F, Villasante A, Dimitri P, Karpen GH, and Celniker SE. Sequence finishing and mapping of *Drosophila melanogaster* heterochromatin. *Science*, 2007.

[21] Meńdez-Lago M, Wild J, Whitehead SL, Tracey A, de Pablos B, Rogers J, Szybalski W, and Villasante A. Novel sequencing strategy for repetitive DNA in a *Drosophila* BAC clone reveals that the centromeric region of the Y chromosome evolved from a telomere. *Nucleic Acids Research*, 2009.

[22] Jain M, Olsen HE, Turner DJ, Stoddart D, Bulazel KV, Paten B, Haussler D, Willard HF, Akeson M, and Miga KH. Linear assembly of a human centromere on the Y chromosome. *Nature Biotechnology*, 2018.

[23] Khost DE, Eickbush DG, and Larracuente AM. Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila melanogaster*. *Genome research*, 2017.

[24] Larracuente AM and Mellone BG. Unpublished data. .

[25] O'Connor C. Chromosome segregation in mitosis: The role of centromeres. *Nature Education*, 2008.