Understanding Meiotic Recombination and Genomic Organisation of Plant Species

Piotr Włodzimierz

Department of Plant Sciences
University of Cambridge

This thesis is submitted for the degree of
Doctor of Philosophy

Darwin College

December 2022

Declaration of Originality

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text

It is not substantially the same as any that I have submitted, or is being concurrently submitted for a degree, diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution.

It does not exceed the prescribed word limit for the Biology Degree Committee.

Piotr Włodzimierz

December 2022

Abstract: Understanding Meiotic Recombination and Genomic Organisation of Plant Species

Piotr Włodzimierz

Reciprocal exchange of eukaryotic genetic material during meiotic crossing over is a major source of genetic variation in sexually reproducing species. Crossover events are not distributed randomly along chromosomes and some regions of the genome, like the centromeres, rarely undergo recombination. Modifying crossover levels and distributions via genetic engineering may provide effective tools for plant breeders to accelerate strain improvement. Despite low recombination rates and their evolutionary conserved function as kinetochore assembly loci, centromeres exhibit some of the highest levels of variation within eukaryotic genomes. Discovery of centromere structure has been hindered by the challenging process of genomic assembly of repetitive regions, as many species contain megabase-long arrays of centromeric tandem repeats. While new long-read DNA sequencing technologies allow for more accurate assembly across the centromeres, methods for their annotation are also required.

In this thesis, I present the development of Tandem Repeat Annotation and Structural Hierarchy (TRASH) software that facilitates analysis of tandem arrays, including centromeric satellite arrays, without prior knowledge of repeat families present in an assembly. I benchmarked TRASH against other software and found it to advance on the current annotation and analysis methods. I used TRASH to analyse in depth the centromeric structures of multiple accessions of metacentric *Arabidopsis thaliana*, *Arabidopsis lyrata*, *Brassica oleracea* and holocentric *Rhynchospora* genus species.

I also present progress towards investigation of the HEI10 meiotic E3 ligase and its role in modulation of crossover levels. Specifically, quantification of the dosage effect of *HEI10* multi-copy lines of tomato and *Arabidopsis* on the crossover recombination landscape.

Together this work contributes to a better understanding of plant centromeric regions and meiotic recombination modulation. It also provides a novel bioinformatics tool for centromere sequence analysis and tandem repeat identification to the scientific community,

Table of contents

List of figures

List of tables

Abbreviations

A	Adenine
bp	Base pair
CG	DNA cytosine methylation in CG contexts
CHG	DNA cytosine methylation in CHG contexts (H = A,C,T)
CHH	DNA cytosine methylation in CHH contexts (H = A,C,T)
ChIP-seq	Chromatin immunoprecipitation-sequencing
cM	Centimorgan
CTAB	Cetyl trimethylammonium bromide
C	Cytosine
CCAN	Constitutive centromere associated network
dHJ	Double Holliday junction
DNA	Deoxyribonucleic acid
G	Guanine
HOR	Higher-order repeat
dsDNA	Double-stranded DNA
D-loop	Displacement loop
DSB(s)	Double strand break
GFP	Green fluorescent protein
Kbp	Kilobase pairs (thousands of base pairs)
LTR	Long terminal repeats
Mbp	Megabase pairs (millions of base pairs)
RNA	Ribonucleic acid
SNP	Single nucleotide polymorphism
SSLP	Simple sequence length polymorphisms
T	Thymine
TAIR	The Arabidopsis Information Resource
TE	Transposable element
TSS	Transcriptional start site
TTS	Transcriptional termination site

Chapter 1

Introduction

1.1 Meiosis and sexual reproduction

Meiosis is a specialised eukaryotic cell division that halves the chromosome number to produce spores/gametes, thereby maintaining the ploidy level in successive generations of sexually reproducing species (Kleckner 1996). Compared to mitosis, meiosis involves two independent rounds of chromosome segregation and cell division, coupled to one round of DNA replication (Murray and Jeyaraman 1985). During the first meiotic cell division (meiosis I), homologous chromosomes segregate to opposite cell poles producing two daughter cells that have one copy each of homologous chromosomes, which is termed a reductional division. The second meiotic division (meiosis II) is similar to mitotic divisions, where sister chromatids separate, resulting in four haploid cells (Mézard et al. 2007). The resulting haploid cells can form gametes that can participate in fertilisation (Bolcun-Filas and Handel 2018). A further key event during meiosis I is induction of DNA double strand breaks (DSBs), which may be repaired using a homologous chromosome to form reciprocal crossovers or gene conversions (Mercier and Grelon 2008, Osman, Voelker and Langer 2011). Reciprocal crossover of genetic material between homologs allows linked genes and genetic variation to be reassorted into new combinations of alleles. The combined processes of independent chromosome segregation and recombination during meiosis have a profound effect on genetic variation, genome evolution and adaptation. Meiotic recombination in most plants, as in many eukaryotes, is required for homolog pairing and synapsis during meiosis I (Gerton and Hawley 2005).

Crossovers provide a physical connection between each pair of homologous chromosomes during segregation at the end of meiosis I, which are cytologically evident as chiasmata at metaphase I. In *Arabidopsis*, formation of chiasmata is required for accurate segregation of homologs to opposite cell poles (Moran et al. 2001, Armstrong and Jones 2003). Therefore, the number of crossovers tends to not be lower than one per pair of homologs per meiosis, which is termed the obligate crossover (Jones and Franklin 2006). Mutants that reduce or fail to crossover result in fertility defects and aneuploidy, due to unbalanced segregation of chromosomes during meiosis I (Grelon et al. 2001, Yin et al. 2002). For example, *Arabidopsis thaliana mnd1* mutants lack chromosome pairing and synapsis, leading to almost complete infertility (Kerzendorfer et al. 2006). Mutations of the *Arabidopsis* genes required for DSB formation, such as *SPO11-1*, *SPO11-2* and *MTOPVIB*, eliminate crossovers and show random segregation of homologs during meiosis I, which produces a large proportion of aneuploid gametes (Grelon et al. 2001, Stacey et al. 2006). In *zip4* mutants, crossover formation is reduced to around 85%, which also causes common aneuploidies and fertility decreases (Chelysheva et al. 2007, Mercier and Grelon 2008, Osman et al. 2011). Interestingly, crossover numbers tend not to be higher than 2 or 3 per chromosome pair per meiosis in most natural species, potentially indicating selection against elevated recombination rates (Mercier et al. 2015, Gaganpreet et al. 2015).

1.2 Meiotic recombination pathways

Interhomolog crossovers are avoided during mitosis as they can cause genome rearrangements that could result in loss of heterozygosity and genome instability (Heyer, Ehmsen and Liu 2010). Hence, DSBs formed in mitotic cells are preferentially repaired using non-homologous end joining or using the sister chromatid as a template for homologous recombination (Anderson et al. 2010). In contrast, during meiosis, interhomolog recombination is actively promoted (Schwacha and Kleckner 1997). Meiotic recombination is initiated by DNA double strand break (DSB) formation, catalysed by a topoisomerase VI-like complex, consisting of SPO11-1, SPO11-2 and MTOPVIB, which also requires the accessory proteins PRD1, PRD2, PRD3 and DFO (Hartung et al. 2007, De Muyt et al. 2007, De Muyt et al. 2009, Grelon et al. 2001, Zhang

et al. 2012, Tang et al. 2017, Vrielynck et al. 2016, de Massy 2013) (**Fig. 2.1**). DNA DSBs generated by the SPO11-1 complex are resected in the 5'-3' direction, leaving 3' single stranded DNA overhangs, which are bound by the RecA-related recombinases RAD51 and DMC1 (Keeney and Neale 2006). The resulting nucleofilament can undergo invasion of another chromatid, either a sister or a homolog, and undergo homology search (Bishop et al. 1992, Shinohara, Ogawa and Ogawa, 1992). DSBs can be repaired using a sister chromatid, or a homologous chromosome as a template, with the latter being promoted during meiosis (Keeney and Neale 2006). The invasion of the ssDNA displaces the intact template DNA resulting in a joint molecule called a displacement loop (D-loop) (Kobbe et al. 2008). D-loops can be processed via different recombination pathways during meiosis. In plants, the majority of strand invasion intermediates undergo non-crossover repair, which can lead to nonreciprocal genetic exchange, including gene conversion (Yang et al. 2012, Drouaud 2013). Alternative pathways that promote non-crossover pathway have been identified in plants: (i) The FANCM helicase and its associated proteins MHF1 and MHF2 are proposed to unwind D-loops and drive them towards non-crossover formation through the SDSA (synthesis dependent strand annealing) pathway (Singh et al. 2010, Girard et al. 2014). In this pathway, a D-loop extends from the 3' via DNA synthesis and is later dissociated from the template DNA, followed by repair with the parental duplex, leading to conversion of a short DNA fragment (McMahill, Sham and Bishop 2007, Crismani et al. 2012; Girard et al. 2014). (ii) The BTR complex consisting of RECQ4/BLM helicases, TOPOISOMERASE3α (TOP3α), and RMI1, is known to unwind D-loops and promote double Holliday Junction (dHJ) dissolution *in vitro* (Reynard, Bussen and Sung 2006, Bachrati, Borts and Hickson 2006, Manthei, Keck 2013). TOP3α and RMI1 appear to have a role independent of RECQ4 in removing recombination intermediates at a later step in meiosis (Séguéla-Arnaud et al. 2015, 2017). (iii) Pathway involving the AAA-ATPase FIDGETIN-LIKE-1 (FIGL1) and its interacting protein FLIP (Girard et al. 2015, Fernandes et al. 2018). Epistasis analysis suggests that FIGL1 regulates the invasion step of homologous recombination by antagonising BRCA2, a positive mediator of RAD51/DMC1 recombinases (Kumar 2019). As FIGL1 belongs to a family of unfoldase proteins, it is possible that it may disassemble RAD51/DMC1 filaments and thus limit aberrant joint molecule formation, thereby regulating crossover formation (Fernandes

et al. 2018). Alternatively, to the non-crossover pathways, the invading 3' ssDNA end may undergo second end capture and double Holliday junction (dHJ) formation, which can be resolved to form crossovers (Mercier et al. 2005).

Crossovers are formed by either interfering Class I or non-interfering Class II repair pathways (Pradillo et al, 2014) (Fig. 1.1). Most crossovers that form in wild type plants are dependent on the Class I repair pathway (Higgins et al. 2004, Mercier et al. 2005, Falque et al. 2009, Shen et al. 2012). For example, around 85% of total *Arabidopsis thaliana* crossover events are Class I dependent (Higgins et al. 2008). The Class I pathway involves a group of genes termed ZMM (an acronym for yeast proteins Zip1, Zip2, Zip3, Zip4, Msh4, Msh5, Mer3) that are thought to stabilise D-loop intermediates and promote the formation and resolution of dHJs as crossovers (Kurzbauer et al. 2012). The identified *A. thaliana* ZMM proteins are SHOC1, HEI10, ZIP4, MSH4, MSH5, MER3, PTD, MLH1, and MLH3 (Higgins et al. 2004, Mercier et al. 2005, Jackson et al. 2006, Macaisne et al. 2008, Higgins et al. 2008, Kuromori et al. 2008, Macaisne et al. 2011, Chelysheva et al. 2012, Mercier et al. 2015, Dion et al. 2007). Class I crossovers can be visualised via MLH1 foci at the pachytene stage of meiosis (Chelysheva et al. 2010). Other ZMM proteins such as MSH4, MSH5 or HEI10 form numerous foci during early prophase, in addition to marking crossover sites during late prophase I (Higgins et al. 2004, Higgins et al. 2008, Chelysheva et al. 2012). Class I crossovers also show interference, meaning that once a crossover becomes designated to form there is a decreased probability of a second crossover occurring in a distance-dependent manner (Mercier et al. 2015, Morgan et al. 2021). Class II crossovers are a minority class, and are dependent on structure-specific endonucleases, including MUS81, and do not display interference (Berchowitz et al. 2007). For example, the Arabidopsis *mus81* mutation reduces recombination by around 10% (Berchowitz et al. 2007). Mutations in proteins of the NCO pathways cause a significant increase in Class II non-interfering crossovers (Girard et al. 2015, Fernandes et al. 2018). The different anti-crossover factors act in parallel to limit crossovers, and when mutations are combined between these pathways, significant additive recombination increases have been achieved (Crismani et al. 2012, Girard et al. 2015, Fernandes et al. 2018).

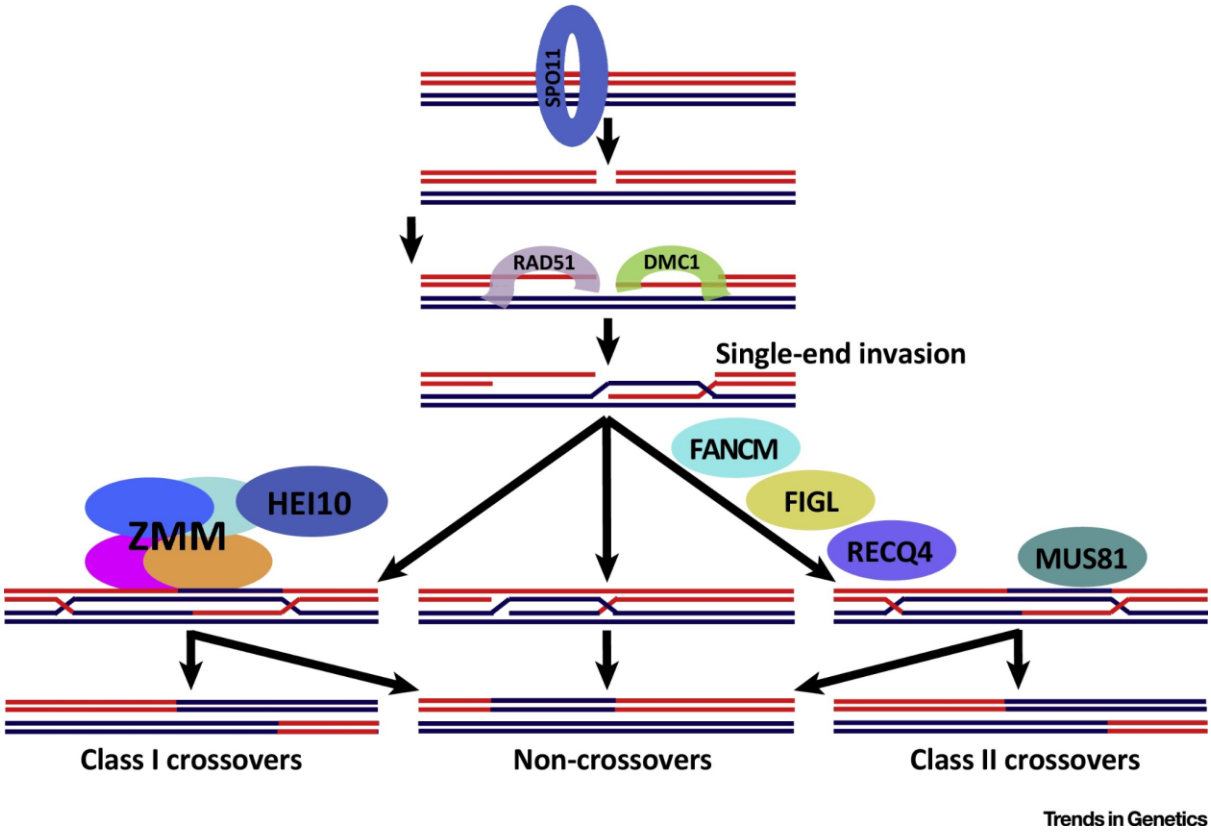


Figure 1.1. Meiotic recombination pathways.
Different pathways of DSB processing function during meiosis.Crossovers are formed by either the interfering Class I or non-interfering Class II pathways. Figure from Zelkowski et al. 2019

During meiosis, recombination occurs as the chromosomes adopt a highly specialised structure (Zickler and Kleckner 1999). Sister chromatids organise along a proteinaceous axis (the axial element) that assembles at the leptotene stage of prophase I (Zickler and Kleckner 1999). Components of the axis include cohesins (including the meiosis specific REC8), that form ring shaped complexes binding sister chromatids together and anchoring axial elements to chromatin (Sun et al. 2015, Lambing et al. 2020). Other components of the axis in plants include the HORMA domain protein ASY1 and its interacting partner ASY3 (Armstrong et al. 2002, Ferdous et al. 2012). Following axis assembly, the paired homologous chromosomes become connected by transverse filaments (e.g., ZYP1 in plants) (Higgins et al. 2005). This process is initiated at multiple sites and proceeds along the chromosomes to form a tripartite structure called synaptonemal complex (SC) (Page and Hawley 2004). Polymerisation of axial elements and SC are required for efficient and high fidelity

interhomolog recombination. In *Arabidopsis*, synaptonemal complex physically connects homologous chromosomes and is essential for the crossover interference, as shown by crossover mapping in the *zyp1* mutant (Capilla-Perez et al. 2021, France et al. 2021). Interestingly, the *zyp1* mutant was also shown to eliminate heterochiasmy and equalise crossover frequency between male and female meiosis (Capilla-Perez et al. 2021, France et al. 2021). Crossover interference is also dependent on the ASY1 protein and its differential enrichment along the chromosome plays a role in shaping recombination distributions (Lambing et al. 2020)

1.3 Distribution of meiotic recombination along chromosomes

In *Arabidopsis thaliana*, the initial meiotic DSBs (~200) outnumber the final crossovers (~10) by ~20:1 (Cifuentes et al. 2013). This suggests that there are active mechanisms that prevent DSBs from maturation into crossovers, including crossover interference and non-crossover repair (Copenhaver et al. 2002, Zhang et al. 2014, Serra et al. 2018). Thus, most of the DSBs are likely channelled to inter-sister repair or are repaired as non-crossovers (Mercier et al. 2015). Apart from crossovers being low in number, the genome-wide distribution of recombination is also uneven (Salome et al. 2012, Rowan et al. 2015) (Fig. 1.2). Narrow regions tend to be overrepresented for recombination (hotspots), and *vice versa* (coldspots) (Mézard et al. 2007). The location and frequency of meiotic DSBs likely has a major effect on possible crossover sites (Fig. 1.2). Crossover sites are enriched for unmethylated DNA, AT-rich sequences, and nucleosome-free regions, and are themselves enriched at gene promoters, terminators, and introns (Wijnker et al. 2013, Choi et al. 2013, Choi et al. 2018). Crossovers also positively associate with poly-A and CTT/GAA motifs (Wijnker et al. 2013, Choi et al. 2013, Shilo et al. 2015).

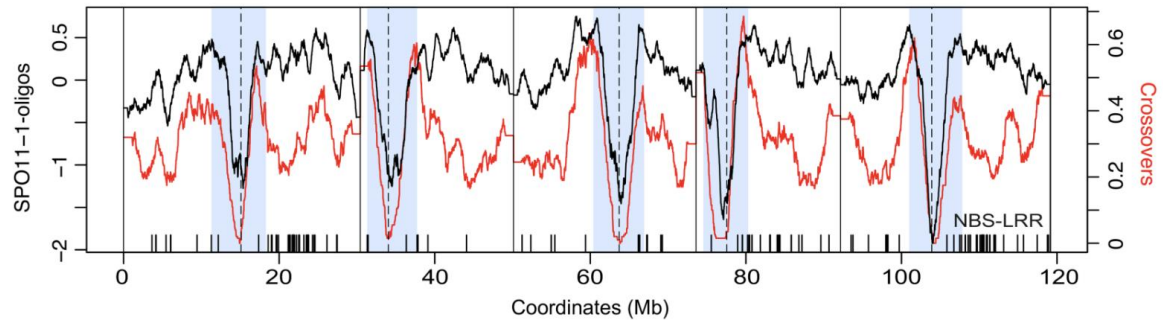


Figure 1.2. Variation in meiotic recombination frequency in *Arabidopsis* and tomato. Distribution of SPO11-1-oligos (black), representing meiotic DSBs across the five *Arabidopsis* chromosomes plotted on a continuous x axis, plotted against crossovers distribution (red) (figure from Choi et al. 2018). These plots were prepared against the TAIR10 reference sequence, which does not include complete centromere sequences. The dotted lines indicate the centromeric assembly gaps.

The repetitive, heterochromatic pericentromeres tend to be cold spots for meiotic recombination (Choi et al. 2013). These regions are nucleosome-dense and modified with DNA methylation and histone modifications, including H3K9 methylation (Choi et al. 2018). Disruption of H3K9 by mutation of *SUVH4*, *SUVH5* and *SUVH6* histone methyltransferases or non-CG DNA methylation by mutation of the *CMT3* methyltransferase increases meiotic DSBs and crossovers in proximity to the centromeres (Underwood et al. 2018). In contrast, disruption of CG context DNA methylation via the *MET1* methyltransferase causes decreased pericentromeric crossovers that are compensated for by a crossover increase in the euchromatic regions (Yelina et al. 2012, 2015). Polymorphism between homologous chromosomes also plays a role in crossover regulation (Dooner 1986, Cole et al. 2010). For example, in *Arabidopsis msh2* mutants with defective mismatch recognition, recombination was repositioned from diverse pericentromeric regions towards less polymorphic sub-telomeric positions (Blackwell et al. 2020). Meiotic DSBs in *Arabidopsis* have been fine-mapped via sequencing of SPO11-1-oligos, which are by-products of DSB processing by SPO11 and can be enriched via immunoprecipitation of SPO11-1 and sequenced (Choi et al. 2018) (Fig. 1.2B). Crossover frequency positively correlates with SPO11-1-oligo density, but there is also significant variation in their relative ratios along the chromosome (Choi et al. 2018). Hence, there are likely additional controls on crossover formation, downstream of DSB formation.

Centromeric regions, where crossovers formation is completely suppressed (Charlesworth et al 1986, Talbert and Henikoff 2010), are responsible for attachment of chromosomes to the microtubule spindle during cell division and for proper disjunction of eukaryotic chromosomes, during both mitotic and meiotic cell divisions (Lermontova et al. 2015; King and Petry 2016). As in most eukaryotic genomes, plant centromeres are often enriched in repetitive satellite sequences (Dong and Jiang 1998). For example, *CEN178* is a tandem repeat present at high copy number within *Arabidopsis* centromeres that is the site of spindle attachment, interspersed with *ATHILA* retrotransposons (Copenhaver and Preuss, 1999; Maheshwari et al. 2017, Naish et al. 2021). Interestingly, despite crossover suppression in the centromeres, the satellite sequences change rapidly between species, indicating that other non-crossover recombination pathways may occur in the centromeres that lead to sequence diversity (Hall et al. 2003, Naish et al, 2021).

1.4 HEI10 is a dosage-dependent regulator of meiotic crossover

Development of high-throughput methods for analysing crossover frequency, like Fluorescent Tagged Lines (FTLs), have allowed identification of quantitative trait loci (QTLs) that influence crossover frequency in specific genomic intervals (Francis et al. 2007, Ziolkowski et al. 2015) (Fig. 1.3). Mapping of the QTL responsible for differences in crossover frequency between the natural *Arabidopsis* ecotypes Columbia (Col) and Landsberg *erecta* (Ler), identified polymorphisms in the putative ubiquitin E3 ligase gene *HEI10*, a previously described ZMM pathway gene (Chelysheva et al. 2012, Ziolkowski et al. 2017). *HEI10* is highly dosage-sensitive and is a limiting factor for Class I crossover formation in *Arabidopsis* (Ziolkowski et al. 2017). The *hei10* mutation was also observed to show semi-dominant effects on crossover frequency (Ziolkowski et al. 2017). Furthermore, overexpression of *HEI10* via transformation of additional copies leads to an increased crossover rate in the genome of *Arabidopsis* by up to 2.7-fold, compared to wild type (Ziolkowski et al. 2017) (Fig. 1.3). A synergistic effect can be obtained by combining *recq4a recq4b* mutations with *HEI10* overexpression (Séguéla-Arnaud et al. 2015, Ziolkowski et al. 2017, Serra et al. 2018). In the case of both anti-crossover mutants and *HEI10* overexpressor lines, crossovers

are most increased in the distal sub-telomeric regions, while they remain suppressed in the centromere-proximal regions (Ziolkowski et al. 2017, Serra et al. 2018). The regions that show the greatest crossover increases are also the least polymorphic and have lowest levels of DNA methylation (Ziolkowski et al. 2017, Serra et al. 2018, Fernandes et al. 2018). This indicates that not all regions of the chromosomes are equally sensitive to increased HEI10 activity.

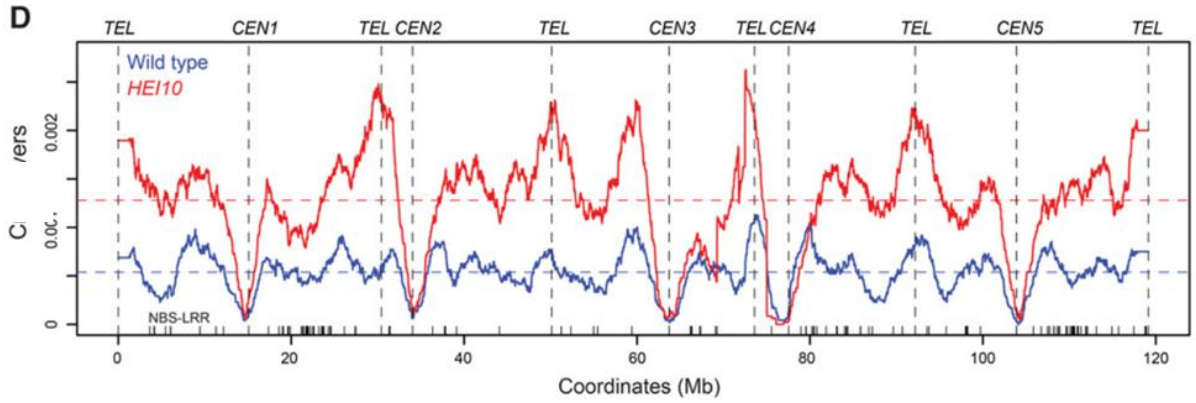


Figure 1.3. Crossover landscape in an *A. thaliana HEI10*-overexpressor line.
Normalized crossovers per megabase ratio plotted over 5 concatenated *Arabidopsis thaliana* chromosomes for wild type (blue) and *HEI10* overexpressing (red) plants (figure from Ziolkowski et al. 2017). Recombination data was acquired using a high-resolution genotyping-by-sequencing (GBS) method.

HEI10 was previously described to function within the ZMM pathway to promote crossovers and belongs to a family of E3 ligases that promote crossovers in diverse eukaryotes (Chelysheva et al. 2012). These proteins have been divided into two broad sub-groups of HEI10-related proteins with proposed ubiquitin-ligase activity, and Zip3/RNF212-related proteins with proposed SUMO-ligase activity, although the differences in their biochemical functionalities are not well defined (Chelysheva et al. 2012, Qiao et al. 2014, De Muyt et al. 2014, Li et al. 2018). Other characterised representatives of this family include ZHP-3 (*Caenorhabditis elegans*), Zip3/Cst9 (*Saccharomyces cerevisiae*), VILYA (*Drosophila melanogaster*) and RNF212 and HEI10 (*Mus musculus, Homo sapiens*), which all play pro-crossover functions (Agarwal et al. 2000, Chowdhury et al. 2009, Jantsch et al. 2004, Kong et al. 2014). In mammals, HEI10 and RNF212 are both present and control the process of recombination (Toby et al. 2003, Reynolds et al. 2013). In mice, it was proposed that these proteins perform

antagonistic functions as ubiquitin and SUMO E3 ligases, where HEI10 promotes dissociation of RNF212 from recombination sites (Qiao et al. 2014). Cytologically, HEI10 foci can be detected throughout the duration of meiotic prophase I in *Arabidopsis* and rice (Chelysheva et al. 2012, Wang et al. 2012). However, HEI10 distribution changes with the progression of meiosis (Morgan et al. 2021). During zygotene, HEI10 is widely localised along axial elements (co-localising with ASY1), while it becomes restricted to MLH1 foci (a Class I crossover marker) at later stages of prophase I (Chelysheva et al. 2012). In *hei10* mutants, the progression of early meiosis is not disturbed, until early diakinesis, where cells show a mixture of univalent and bivalents, suggesting that the pairing of homologs is disturbed (Chelysheva et al. 2012). Indeed, the number of chiasmata in these lines is reduced, which strongly reduces fertility (Chelysheva et al. 2012). This is analogous to other ZMM pathway mutants, for example *zip4* (Chelysheva et al. 2007). In rice, the HEIP1 protein was found to interact with HEI10 via a yeast two-hybrid system (Li et al. 2018). HEIP1 colocalises with HEI10 along meiotic chromosomes and loses this localisation in *hei10* mutants, and *heip1* mutations lead to decreased chiasma frequency (Li et al. 2018). The most likely HEIP1 homolog in *Arabidopsis* is AT2G30480 but is yet to be confirmed and characterised.

1.5 Meiotic recombination as a tool in plant breeding

Modulating meiotic crossover frequency is of great interest for crop improvement, as plant breeding relies on recombination to generate desirable allelic combinations (Wijnker and de Jong, 2008). Most of the data concerning meiotic recombination in plants come from *Arabidopsis thaliana*, which is a model species with a relatively small genome (~130 Megabases), where transcriptionally active euchromatin spans most of the chromosomes. In contrast, other crop species, generally have much larger genomes (e.g wheat=17,000 Mb, tomato=950 Mb and maize=2,300 Mb) and a greater proportion of transposons and heterochromatin (Wang et al 2006, Anderson et al. 2019, Naish et al. 2021). Furthermore, crossover recombination in many of the crops is skewed towards the distal euchromatic ends of the chromosome arms (Lambing, Franklin and Wang 2017), while pericentromeric repetitive regions are silenced for recombination, which may span most of the chromosome in some cases (Schreiber

et al. 2022, Fuentes et al. 2022). Despite this, these regions often contain important genetic variation that are inaccessible to breeders due to recombination patterns. Understanding meiotic recombination may provide knowledge or technology that can unlock, increase, and redistribute crossover events in plant genomes and thereby accelerate crop improvement.

1.6 Centromere function in meiosis and mitosis

Most plant species have strong crossover frequency biases away from the heterochromatic centromeres and telomeres, and recombination 'cold regions' are usually associated with pericentromeric chromatin, which is also epigenetically silent for RNA polymerase II transcription (Fernandes et al. 2019). As a result, breeders have a limited pool of allelic diversity to recombine and select from (Lambing, Franklin and Wang 2017, Choi et al. 2017). Despite crossovers being suppressed within heterochromatin in *Arabidopsis* and maize, DSBs have been detected over the pericentromeric heterochromatic regions, including within specific families of transposons (Choi et al. 2018, Underwood et al. 2018, He et al. 2017). As a consequence of these DSBs, non-crossover outcomes may occur, or inter-sister repair, in the heterochromatin, despite crossover repair being suppressed (Shi et al. 2010, Underwood and Choi 2019). In plants, centromeres play an important role during meiosis, and are vital for recognizing homologous chromosomes, pairing during meiosis, and formation of the synaptonemal complex during meiosis (Da Ines and White 2015). For example, centromere coupling during meiosis is lost in an *Arabidopsis dmc1* mutant (Da Ines and White 2015). DMC1 is a protein loaded onto single-stranded DNA before its invasion of the homologous chromosome during meiotic recombination (Da Ines et al. 2012). A major difference in centromere behaviour between meiosis-I and mitosis, is that during meiosis-I the centromeres of the replicated homologs are mono-orientated to the same cell pole, whereas they are bi-oriented to different cell poles in mitosis (Tanaka, Stark and Tanaka 2005).

1.7 Centromere genetic and epigenetic structure

Centromeres can form a number of structures regarding their position along the chromosomes, including point centromeres, regional centromeres and dispersed holocentric chromosomes (McKinley and Cheeseman 2016) (Fig. 1.4). The function of centromeres is primarily epigenetic and is dependent upon the loading of nucleosomes containing the histone H3 variant called centromere protein A (CENP-A), or CENH3 in plants (Foltz et al. 2006, Lermontova et al. 2006). CENH3 interacts with kinetochore proteins and regulates centromere function and the site of spindle

attachment to the chromosome during mitosis and meiosis. There have been significant advances in our understanding of the molecular basis of centromere function, including their genetic and epigenetic characteristics, and the mechanisms of centromere propagation.

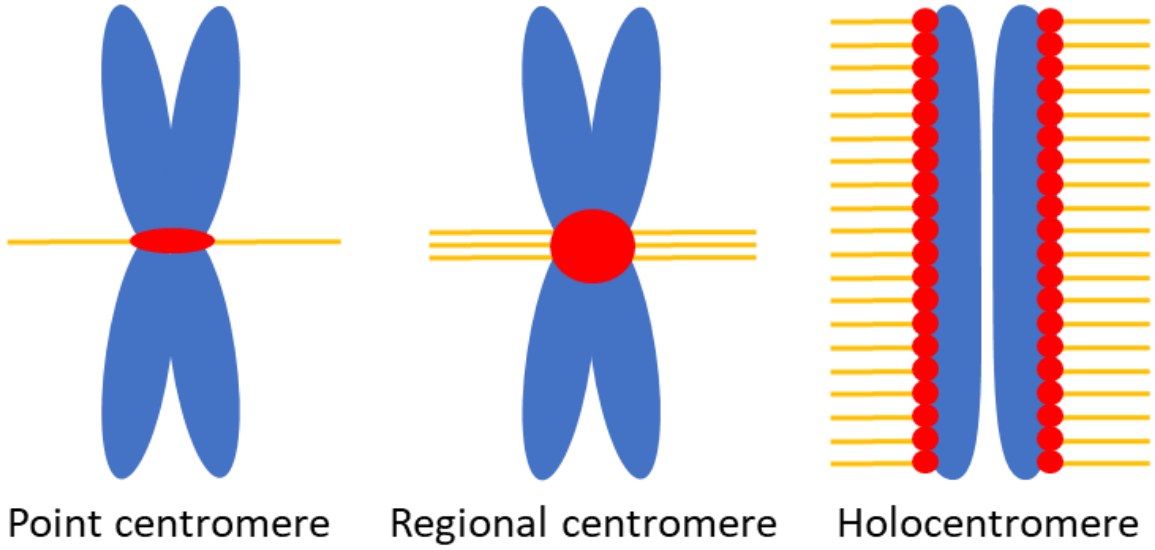


Figure 1.4. Basic structural types of centromere organisation.
Genomic DNA is represented in blue, kinetochore proteins in red and microtubules in orange. The diagrams represent point, regional and holocentric modes of centromere organisation.

Centromeres and pericentromeres are distinguished by the organisation of their DNA sequence repeats and by their distinct chromatin signatures. For example, core centromeres in human have been observed to contain CENP-A-nucleosomes interspersed with canonical H3 nucleosomes that carry transcriptionally permissive marks, including H3K4me2 and H3K36me2, deposited over repetitive alpha-satellite arrays in human centromeres (Sullivan and Karpen 2004, Bergmann et al. 2011, McKinley and Cheeseman 2016). Artificially increasing heterochromatin by targeting histone acetyltransferases to alphoid domains resulting in elevated H3K9 acetylation is detrimental to the CENPA deposition and can induce kinetochore assembly at an ectopic location (Nakano et al. 2008, Ohzeki et al. 2012). Centromere DNA sequence is frequently composed of retrotransposons and/or long arrays of tandem repeats (Plohl, Mestrovic and Mravinac 2014). These tandem repeats are often similar in length to the DNA bound by a single nucleosome (~140-200 nucleotides) or multiples

of this (Gent et al. 2011, Melters et al. 2013). Certain monocentric organisms, however, have non-repetitive centromeres, such as the yeast Candida albicans (Mishra, Baum and Carbon 2007), or a mixture of repetitive and non-repetitive centromeres, such as seen in orangutans, horses, chickens and potatoes (Wade et al. 2009, Shang et al. 2010, Gong et al. 2012).

1.8 Point centromeres

In organisms with point centromeres, specific DNA sequences are necessary for centromere function, these can be referred to as genetic centromeres (Clarke and Baum 1990). For example, the centromere of budding yeast is organised into a single, specialised nucleosome containing Cse4 (the ortholog of the centromere-specific variant of histone H3) that is deposited over a 145-147 bp DNA region called *centromere DNA element III sequence* (*CDEIII*) (Meluh et al. 1998, Meraldi et al. 2006, Furuyama et al. 2007). The CEN nucleosome contains two H2A-H2B dimers and a central Cse4-H4 tetramer, and centromere binding proteins (Cbf1 and CBF3) bind to DNA on the outside of the nucleosome, directly fixing its position (Cole et al. 2011). Point centromeres are believed to have evolved from regional centromeres with the transition from epigenetically defined regional centromeres in other fungi to genetically defined "point" centromeres in budding yeast (Malik and Henikoff 2009). *C. albicans* is an intermediate species, having lost heterochromatin and RNAi protein machineries that are involved in maintaining the epigenetically defined centromeres, and is intermediate both in terms of phylogenetic position, as well as centromere complexity (Sanyal et al. 2004). It should be noted, however, that epigenetic modifications are still essential in species with point centromeres, while regional centromeres are often constrained to specific elements of the centromeric regions, highlighting the interplay of both genetic and epigenetic elements in all centromere types. The coincidence of gaining point centromeres and losing heterochromatin/RNAi machinery in the budding yeast lineage that gave rise to *S. cerevisiae* and *K. lactis* raises an important question about how a transition from a regional centromere that relies on heterochromatin machinery for cohesion could have occurred (Malik and Henikoff 2009). One model of the evolution of the point centromeres highlights the reliance of the 2-micron plasmids found in the budding

yeast on the existing segregation machinery and stipulates that integration of the plasmid DNA into a host chromosome would introduce genetic neocentromere, which would be later adopted to become the main centromere (Ghosh et al. 2007, Hajra et al. 2006, Malik and Henikoff 2009). This is supported by the fact that like 2-micron plasmids, *S. cerevisiae* centromeres rely on unusual DNA-adaptor proteins to recruit conserved centromeric and kinetochore proteins (Ghosh et al. 2007). These genetic centromeres, while functionally conserved, contrast with the megabase scale DNA-satellite rich epigenetic centromeric regions of most plants and animals.

1.9 Regional centromeres

Human peri/centromeric satellite DNAs represents ~6.2% of the genome, which has been recently assembled and analysed in depth by the Telomere-to-Telomere (T2T) consortium (International Human Genome Sequencing Consortium 2001, Nurk et al. 2022, Altemose et al. 2022). All centromeric regions of the human CHM13 assembly contain long tracts of tandemly repeated alpha-satellite monomers (85.2 Mb total genome-wide), and most chromosomes also contain classical human satellite 1, 2 and/or 3, totalling 28.7 and 47.6 Mb, respectively (Altemose et al. 2022). Within the alpha-satellite repeats, evidence was found for an ancient duplication event that predated African ape divergence and involved a large segment of the ancient chromosome 6 centromere, in addition to ~1 Mb of adjacent p-arm sequence (Altemose et al. 2022). The CHM13 assembly revealed regions where combinations of transposon sequences have been tandemly duplicated, forming "composite satellites", and that other satellites often include fragments of ancient transposons as part of their repeating units (Hoyt et al. 2022, Altemose et al. 2022). In a number of species, including human, new-world monkeys and mice, centromeric DNA forms higher order repeat structures, where two or more repeats are tandemly repeated at different points in the centromere (Warburton, Waye and Willard 1993, Sujiwattanarat et al. 2015, Guenatri et al. 2004). Evidence of layered expansions across all alpha-satellite sequences was detected: divergent alpha-satellites flank the central, youngest HOR arrays across the genome and accumulate mutations, inversions, transposon insertions, and non-alpha-satellite satellites over time (Altemose et al. 2022). CENPA was found almost exclusively within the active HOR arrays (Altemose

et al. 2022). This provides a model of the centromere expansion, in which the CENPA is deposited onto the youngest repeats found in the central parts of the alpha-satellite arrays.

In budding yeast, the sequence-specific binding protein Cbf3 recognizes the *CDEIII* centromere sequence, providing a direct link between DNA sequence and function. With regional centromeres, however, predicting the potential roles of a sequence-specific DNA binding protein becomes more challenging due to an extreme divergence of these loci, seemingly uncorrelated with the deposition of centromeric protein, for example stable centromeres can be formed without the usual centromeric repeat arrays in *Equus* species (Wade et al. 2009, Musilova et al. 2013).

1.10 CENPB recognition sequence on human centromeric repeat arrays

Although human centromeres are epigenetically defined by CENPA loading, they also contain DNA-binding motifs important in kinetochore assembly (Okada et al. 2007). The centromere alpha satellite sequences vary between chromosomes, but all human centromeres, except for the chromosome Y, have a CENPB box (or simply the B box) in their alpha-satellites (type B satellites) that bind the CENPB protein (Miga et al. 2014, Altemose et al. 2022). In type A alpha-satellites, a 19-bp motif called n box is found (Talbert and Henikoff 2022). CENPB boxes are most frequently found in dimeric n/B dimers, while they are rarely found in adjacent monomers (Romanova et al. 1996, Alexandrov et al. 2001, Fachinetti et al. 2013). It is likely that the additional kinetochore proteins form complexes with the n/B dimers (Thakur and Henikoff 2018). As a result of their sequence similarity, n/B dimers can be subdivided into SF1 dimers found on chromosomes 1, 3, 5, 6, 7, 10, 12, 16, and 19, while SF2 dimers are found on chromosomes 2, 4, 8, 9, 13, 14, 15, 18, 20, 21, and 22 (Alexandrov et al. 2001; Henikoff et al. 2015). CENPB has been shown to bend DNA by 59° by binding to the CENPB box in the DNA major groove (Tanaka et al. 2001). The protein can also form antiparallel homodimers that bind to two CENPB boxes simultaneously and form a loop between them on the same DNA molecule (Yoda et al. 1998).

During the G1 cell cycle phase, post DNA replication, CENPB is required to restore CENPA loading in the centromere (Fachinetti et al. 2015, Hoffmann et al. 2020). DNA CG methylation within *CENPB* boxes was found to block CENPB binding, suggesting that *CENPB* box epigenetic state can influence CENPA loading, or maintenance on the chromosome (Tanaka et al. 2005). At the chromosome level, CENPA nucleosomes are associated with regions of DNA cytosine hypomethylation in CG sequence contexts relative to surrounding centromeric sequences, called the 'Centromere Dip Region' (CDR), which indicates a correlation between DNA methylation levels and the site of CENPA loading and CENPB domains (Gershman et al. 2022, Altemose et al. 2022). Using a system that allows rapid removal or reloading of CENPA nucleosomes using the auxin (IAA) inducible degradation system, Hoffmann et al. (2020) found that newly loaded CENPA relocalized to the same positions in native centromeres after destruction of existing CENPA. This resulted in *de novo* CENPA deposition without dependence on pre-existing proteins. In the $CENPB^{-/-}$ mutant however, de novo CENPA deposition was strongly impaired (Hoffman et al. 2020). CENPB was also found to recruit CENPC and CENPA when tethered to an ectopic site using a lacO system (Hoffman et al. 2020). In addition, the recruitment of CENPA was dependent on CENPC and could not be achieved directly by CENPB, since the removal of CENPC via siRNA substantially reduced CENPA recruitment (Hoffman et al. 2020).

1.11 Holocentric chromosomes

Holocentric chromosomes show kinetochore complex formation along their chromosomes that can bind to microtubules (Mandrioli and Manicardi 2020). For example, in the *Rhynchospora* genus, holocentric chromosomes have evolved together with inverted meiosis to function with the holocentric chromosome structure (Marques et al. 2016). During inverted, or post-reductional, meiosis, the bivalents align themselves perpendicular to the equatorial plate during metaphase I, with biorientation of sister chromatids forcing them to separate to opposite poles during anaphase I (Wahl 1940, Heckmann et al. 2014). In *Rhynchospora tenuis*, achiasmatic (lacking physical connection between the homologous chromosomes) inverted (sister chromatids segregate at meiosis I, whereas homologous non-sister chromatids segregate at meiosis II) meiosis is possible due to the small number of holocentric

chromosomes inside the nucleus, as random segregation has a high enough ratio of viable offspring (Cabral et al. 2014, Li et al. 2020).

The beak-sedge *Rhynchospora pubera* has repeat-based holocentromeres, which are associated with a 172-bp *Tyba* tandem repeat family and the centromeric retrotransposon of Rhynchospora (*CRRh*) (Marques et al. 2015). Chromatin immunoprecipitation followed by sequencing confirmed that *Tyba* repeats are the main CENH3-binding sites in *R.* pubera and *R. breviuscula*, and that the number of CENH3-binding regions follows a similar pattern to the number of *Tyba* arrays detected (Hofstatter et al. 2022). These repeat-based holocentromeres comprise small islands (20 - 25 kb) of centromeric *Tyba* tandem repeats, and their epigenetic regulation resembles that of monocentric centromeres (Hofstatter et al. 2022). This suggests an evolutionarily conserved epigenetic regulation of repeat-based centromeres in both mono- and holocentric organisms (Hofstatter et al. 2022).

Holocentric species differ from monocentrics not only in chromosomal structure but also in several general karyotypic patterns and the range of chromosome numbers that is nearly continuous, including the example of the largest number of chromosomes in animals (2n=446), found in the blue butterfly *Polyommatus atlantica* (Bures and Zedek 2014). They also exhibit the most extreme size variations. For example, the average chromosome size in the genus *Carex* varies from 2.6 to 122 Mbp, from 8 to 299 Mbp in the genus *Eleocharis*, from 8 to 1324 Mbp in the genus *Luzula* (Bozek et al. 2012, Bennett and Leitch 2012, Bozek et al. 2012).

1.12 Structure of the kinetochore complex

1.12.1 CENPA

CENP-A is a variant of histone H3 that has a long amino terminal tail that differs significantly from canonical H3 in humans and is required to target the centromere and assemble kinetochores (Sullivan, Hechenberger and Masri 1994, Nechemia-Arbely et al. 2017). Compared to canonical nucleosomes, CENP-A nucleosomes are more spatially condensed based on cryo-electron tomography (Geiss et al. 2014). During mitosis, for each sister chromatid centromere to remain epigenetically marked, CENPA-deposition machinery is required (Black and Cleveland 2011). In a process

governed by Cyclin-Dependent kinase 1 (CDK1) and Polo-Like Kinase 1 (PLK1), CENP-A deposition occurs in the G1 phase and is uncoupled from DNA replication (McKinley and Cheeseman 2014, Pan et al. 2018). During DNA replication, CENPA nucleosomes are distributed stochastically to daughter chromosomes, but they are not deposited immediately following replication fork progression (Shelby et al. 2000, Mellone et al. 2011, Lando et al. 2012). After DNA replication, homotypic, octameric CENPA nucleosomes are inherited randomly at each sister centromere but are retained at the same position regardless of the daughter chromosome (Bodor et al. 2014, Rosset al. 2016). The assembly of CENPA nucleosomes is epigenetically influenced by CDK1 phosphorylation of the Mis18 complex, which enables CENPA loading by the HJURP chaperone, and a new nucleosome is assembled near the existing nucleosome (Nardi et al. 2016; Pan et al. 2017, Ross et al. 2016). Through this process, the number of nucleosomes in chromatin is doubled after replication to restore CENPA levels.

As part of the negative regulation of CENPA deposition, CDKs are involved in the degradation of cyclin A, the phosphorylation of M18BP1, as well as the phosphorylation of CENPA itself on residue Serine 68 (Yu et al. 2015). The deposition of CENP-A requires two steps: temporal regulation by CDKs and licensing by PLK1 (Pan et al. 2017, McKinley and Cheeseman 2014). Bypassing both steps by constitutively targeting the MIS18 subunit to the centromere results in CENP-A deposition (Nardi et al. 2016). CENP-A is an essential component of most centromeres, but several other factors also play a role. Aside from CENPB boxes, chromatin remodelers associated with active transcription have also been implicated in the deposition of new CENPA, including RSF1, FACT, CHD1, and RBAP46 (McKinley and Cheeseman 2014). The CENP chaperone HJURP physically interacts with CENPA at centromeres (Dunleavy et al. 2009). It has been demonstrated that HJURP is recruited to centromeres by CENPC and Mis18 complexes and is sufficient for CENPA deposition (Dunleavy et al. 2009, Pan et al. 2017, 2019, Sandmann et al. 2017, Walstein et al. 2021). Thus, a direct or indirect targeting of HJURP to chromatin appears sufficient for CENPA deposition (Barnhart et al. 2011). During G1 phase, the cell can control CENPA nucleosome assembly in chromatin by targeting the Mis18 complex and dissociating the complex upon binding of HJURP (Fujita et al. 2007, Nardi et al.

2016). In another model, CENPC and HJURP bind to adjacent CENPA and H3 dinucleosomes, and CENPA deposition replaces the H3 nucleosome (Pan et al. 2019, Walstein et al. 2021). It is unclear, however, what would prevent CENPA from accumulating between two H3 nucleosomes. CENPC, the Mis18 complex, HJURP, or other factors may distinguish pre-existing CENPA from newly deposited CENPA, limiting uncontrolled deposition of CENPA nucleosomes (Stankovic et al. 2017, French and Straight, 2019, Sundarajan 2022). The functions of CENPC and CENPA are co-regulated, and cells may be dependent on CENPC to restore equilibrium after CENPA is lost from chromatin (Hoffmann et al. 2016, 2020). Notably, HJURP homologs have not to date been identified in plants and CENH3 deposition differs in timing between plants and animals: in plants it occurs before mitotic sister centromere separation, while in animals, after sister centromere separation (Lermontova et al 2006, Ahmad and Henikoff 2001).

Based on the data from the human cell line RPE1, the average centromere contains ~400 CENPA molecules, or ~200 CENPA nucleosomes (Bodor et al. 2014). However, human cell types differ widely in their estimates of CENPA nucleosome abundance, with DLD-1 cells containing as few as 90 nucleosomes per centromere on average (Bodor et al. 2014). The alpha-satellite higher order repeat sequences occupied by CENPA can also vary within a population and sometimes even within an individual (Aldrup-MacDonald et al. 2016). RPE1 cells, for example, have CENPA at different positions on their X chromosomes, indicating that the active CENPA-occupied HOR array may vary within and between populations (Rshrestha et al. 2012, Bodor et al. 2014, Aldrup-MacDonald et al. 2016).

1.12.2 CCAN and KMN networks

The human constitutive centromere associated network (CCAN) consists of 16 proteins located at the centromere throughout the cell cycle (Suzuki et al. 2014). These proteins can be combined into five groups: the CENP-C, CENP-L-N, CENPH-I-K-M, CENPO-P-Q-U-R and CENPT-W-S-X complexes (Foltz et al. 2006, Okada et al. 2006, Hori et al. 2008, Cheeseman and Desai 2008). A robust platform for kinetochore assembly on the centromere is built by CCAN proteins interacting with centromeric chromatin (Suzuki et al. 2014). CCAN recruitment during mitosis is

dependent on CENP-C, a keystone molecule in this assembly (Hori et al. 2013). CENPH-I-K-M is a V-shaped complex formed by CENPH and CENPK, and CENPI, which contains five HEAT-repeat like motifs, associates with CENPH and CENPK to form HIKhead and HIKbase domains, respectively (Tian et al. 2022, Musacchio and Desai 2017). CENPM is located in a pocket on the surface of the HIK base domain (Tian et al. 2022). CENPT-W-S-X forms a heterotetramer with CENPH-I-K-M and shares a similar architecture with the histone H3-H4 tetramer (Tian et al. 2022). CENPT-W-S-X interacts with DNA to create a nucleosome-like structure (Tian et al. 2022). The CCAN provides a platform for the assembly of the outer kinetochore, which is composed of the KNL1 - MIS12 - NDC80 (KMN) network. CENPC and CENPT form parallel but non-redundant pathways to recruit the KMN network, and CENP-U creates a third pathway to recruit the KMN network (Schleiffer et al. 2012, Takenoshita, Hara and Fukagawa 2022).

Some of the homologues of the CCAN and KMN networks have been identified in plants including CENH3 (CENPA variant), CENPC, Ndc80, Nuf2, Spc24, Knl1, Mis12 and Nnf1, but many are still unidentified (Lermontova et al. 2006, Ogura et al. 2004, Allipra et al. 2021, Du, Topp and Dawe 2007, Li and Dawe 2009, Shin et al. 2018).

1.12.3 Physical properties of the human kinetochore

As shown by electron micrographs of metaphase arrested human cells, ~20 microtubules attach to each centromere (McEwen et al. 2001), and CENPA nucleosomes are organised on the surface of the chromosome in three dimensions to facilitate the assembly of the kinetochore and the attachment of microtubules (Marshall et al. 2008). In human cells, chromatin fibre stretching experiments show that CENPA clusters occupy approximately 30-40% of alpha satellite DNA on individual centromere repeat arrays (Sullivan et al. 2011). Many mechanisms could be responsible for dicentric chromosome breakage, including cleavage by the reforming cell wall during plant cytokinesis, the actin-myosin contractile ring, or an endonuclease (McClintock 1938, Lopez et al. 2015, Guérin et al. 2022). According to Novitski's (1952) centromere strength hypothesis, some centromere/kinetochore/spindle combinations break the chromatin bridge. Atomic force microscopy can be used to measure covalent bond strength directly (Grandbois et al. 1999). The strength of

phosphodiester bonds that link adjacent nucleotides in DNA, however, has not been directly tested. If a dicentric bridge does break under tension, the components of the segregation mechanism must be at least as strong as DNA. The strength of the connection between the kinetochore and the underlying chromosome has not yet been measured (Hill and Golic 2022). Microtubule-kinetochore connections appear to be the weakest point based on measurements (Hill and Golic 2022). Nevertheless, the strength of this connection can vary depending on how many microtubules are attached to a kinetochore (Ye et al. 2016).

1.13 Transposable elements

Transposable elements (TE) are mobile DNA sequences capable of replicating themselves within genomes independently of the host cell genome (Wells et al. 2020). Transposons frequently encode multiple biochemically active proteins, as well as complex non-coding regulatory sequences that promote their propagation (Faulkner and Carninci 2009). When transposons insert into the host genome they can result in the evolution of proteins, genes, non-coding RNAs, and *cis*-regulatory elements (Bourque et al. 2018, Cosby et al. 2019). The mobilisation of transposons can also be mutagenic, with potentially severe host phenotypic consequences (Thieme et al. 2017). There is a strong correlation between transposon content and genome size in nearly all eukaryotic genomes examined so far (Lee and Kim 2014). There is, however, no correlation between the proportion of transposons in the genome and the complexity of the organism (Wells and Feschotte 2020). Additionally, transposons are major components of the centromeric and pericentromeric regions of a wide variety of species (Allshire and Karpen 2008, Malik and Henikoff 2009), and have also been proposed to be satellite DNA sources (Mestrovic et al 2015).

Based on the transposition intermediates, eukaryotic transposons can be classified into two major classes: retrotransposons (Class I) and DNA transposons (Class II). Both classes can be subdivided into sub-classes and then into super-families and families (Wicker et al. 2007). It is also possible to classify transposons according to their ability to move autonomously. Many non-autonomous elements, which contain internal deletions, emerge as parasites of autonomous elements (Feschotte et al. 2002).

Based on their mechanism of replication and integration, retrotransposons can be divided into three major subclasses: long terminal repeat (LTR) elements, target-primed non-LTR elements, and YR-mobilised elements (Wells et al. 2020). A non-LTR element is the simplest structurally and usually contains two open reading frames, ORF1 and ORF2 (Burke et al. 1987, Martin 2010). There are two genes inside LTR elements (gag and pol), which are transcribed from a Pol II promoter located within the elements themselves (Bowen et al. 2003). Pol-encoded proteases cleave the polyprotein into multiple proteins (Wilhelm and Wilhelm 2001). Retroviruses replicate and integrate similarly to LTR elements, except for the presence of fusogenic env genes (Eickbush et al. 2002, 2008). Retrotransposons of the YR class are the third major subclass of Class I elements. Despite the fact that they contain terminal repeat sequences, their function and mode of replication remain poorly understood (Poulter and Butler 2015). Penelope elements are unique within the retrotransposon class because they contain pseudo-LTRs and a GIY-YIG endonuclease domain, which is not shared with any other retroelement subclass. Due to this, Penelope-like elements can be considered a separate subclass of retroelements (Evgenev et al. 1997, Arkhipova 2017).

The four major types of DNA transposons are; (i) cut-and-paste elements mobilised by DDE transposases, or by (ii) YR transposases, (iii) rolling-circle elements (also known as Helitrons), and (iv) "self-synthesising" transposons called Mavericks and Polintons (Wells and Feschotte 2020). The DDE transposition process is initiated by nucleophilic attack of a water molecule near the end of terminal inverted repeats (TIRs), which eventually results in the direct removal of the transposon DNA and allows its relocation (Liu, Yang and Schatz, 2019). The number of copies of these elements can increase to form abundant families in the genome (Hickman and Dyda 2016). Although helitrons are abundant in many eukaryotic lineages, they have remained largely uncharacterized until the early 2000s. Their mobilisation mechanism is fundamentally different from that of cut-and-paste elements (Kapitonov and Jurka 2001). For example, Helraiser, an active autonomous element resurrected from inactive Bat genome elements, transposes by a "peel-and-paste" rolling-circle mechanism, but maize genetic data suggests some Helitrons function by excising directly instead of via copying (Li and Dooner. 2009, Grabundzija, Hickman and Dyda

2018). The Maverick class of DNA elements has also been poorly characterised, but they are typified by their large size (15-20 kb) and complexity (Pritham, Putliwala and Feschotte 2007). They share similarities with disparate groups of double-stranded DNA (dsDNA) viruses, including a protein-primed family-B DNA polymerase and they contain a DDE nuclease (Kapitonov and Jurka 2006). In many Maverick elements, double and single jelly-roll capsid-like proteins are encoded, and their close relationships to viruses have led to a suggestion that they could be endogenous viruses or virophages (Krupovic et al. 2014, Koonin and Krupovic 2017).

The observations over the last few decades suggest that all major subclasses of elements are widely distributed throughout the eukaryotic tree, and that elements have evolved in highly modular ways (Wells and Feschotte 2020). Although phylogenomic analyses reveal that there are deep relationships among the core transposition enzymes that define the major TE subclasses, they offer little insight into the deep origins of individual families and superfamilies (Arkhipova 2017). Since chimerism, horizontal and mosaic evolution have a major impact on the evolution and diversification of transposons (Arkhipova 2017), this creates further challenges in understanding their long-term evolutionary relationships.

1.13.1 Transposon insertion patterns along chromosomes

Studies of *de novo* insertions of transposons have documented general patterns, transposons favouring insertion in genomic regions that minimise their deleterious effects, and transposons targeting sites that likely facilitate their subsequent propagation (Sultana et al. 2017). Mechanistically, transposon insertion is dictated by its associated nuclease that catalyses its chromosomal integration (Feng 1996). Several families of LINEs target ribosomal RNA gene arrays and have evolved different site preferences that enable them to coexist within the same genome (Eickbush et al. 2015). Transposons have evolved strategies to avoid genes in compact genomes with little intergenic space, including inserting upstream of Pol III-transcribed genes and within silent chromatin at telomeric regions, such as *Ty1* integration preference to heterochromatin regions in *S. cerevisiae* (Devine and Boeke 1996). Ty1/Copia-like retrotransposons in *Arabidopsis* have evolved a mechanism to favour insertion into a subset of non-essential genes, by targeting H2A-Z containing nucleosomes (Quadrana

et al. 2019). All new transposon insertions are subject to natural selection acting at the level of host fitness. Longer transposons may be strongly selected against due to their increased likelihood of disrupting gene expression and initiating illegitimate recombination, while shorter elements such as SINEs and MITEs accumulate in gene-rich regions (Duret et al. 2000, Wright et al. 2003, Cosby et al. 2019).

1.13.2 Centromeric transposons

Transposable elements can play a major role in the eukaryotic centromeres. For example, in *Dictyostelium discoideum*, identifiable transposons comprise 86% of the centromeres, which are 171–361 kbp in length (Glöckner and Heidel 2009). Centromeres in *Phytophthora sojae* are transposon-rich, especially containing LTR retrotransposons (Fang et al. 2020). Transposons can be also found in the satellite repeat based centromeres, like *Arabidopsis thaliana* (Langdon et al. 2000, Naish et al. 2021), or *Mus musculus* (Bourchis and Bestor 2004). A high number of retrotransposons related to an *Arabidopsis thaliana COPIA* element have been identified in the centromeric sequences of *Arabidopsis lyrata* (Tsukahara et al. 2009). Interestingly, when one of these transposons, *Tal1*, was introduced into *A. thaliana ddm1* mutant, it was found to mobilise and integrate exclusively into the centromeric repeats, despite the diversity in the centromeric sequence of these two species (Tsukahara et al. 2012). Despite this, detailed centromeric maps indicate that *ATHILA transposons of the Ty3/Gypsy family*, and not *COPIA* elements are the main centromeric retrotransposons of *Arabidopsis thaliana* (Naish et al. 2021). Transposons might be inserted in centromeric regions, because they can be transcribed without deleterious effects on transcription of other genes. They may also drive female meiosis by recruiting centromere binding proteins (Kumon and Lampson 2022). To counteract that, host genomes evolved mechanisms to suppress transposon activity within centromeres. RNA-based silencing and protein-based silencing can initiate heterochromatin formation, although these pathways seem insufficient to completely purge transposons from centromeres (Janssen and Colmenares 2018, Rhind et al. 2011). Alternatively, transposons can be domesticated to silence other types of transposons, for example widespread in eukaryotes Pogo-like transposases (Mateo and Gonzales 2014, Gao et al. 2020). CENPB in yeasts and

mammals is one such example of a domesticated Pogo-like transposase that regulates heterochromatin formation (Smit et al. 1996, Kipling 1997, Casola et al. 2008). The question of whether transposons play a functional role in the centromeres or simply hijack these recombination-free zones remains unclear.

1.14 Centromeric transcription

Increasingly, centromeres are viewed as dynamic chromosome regions, rather than being inert, and may include genes that are transcribed at low levels (Su et al. 2019; Henikoff and Talbert 2018). For example, CenRNAs are associated with a broad range of functions, including participating in the regulation of chromosome behaviour, gene transcription, and chromatin architecture (Arunkumar and Melters 2020). The centromeric region of the human genome presents a distinct set of histone modifications, including H3K4me2, which is associated with open, but not actively transcribed, euchromatin (Sullivan and Karpen 2004). In *S. cerevisiae*, *S. pombe*, and human cells, the level of transcription within the centromere is regulated by RNA polymerase II, and competition between the transcription factors Cbf1 and Ste12 and the silencing factors Sir1, Hst1 and Cdc14 (Ohkuni and Kitagawa 2011; Hildebrand and Biggins 2016). Centromeric CENPA tends to associate with RNA polymerase II promoters where RNA Pol-II binding is high (Choi et al. 2011, Ólafsson and Thorpe 2020). The level of centromeric non-coding RNA transcription is dependent on activation of RNA Pol-II and varies between developmental stages and tissues (McNulty et al. 2017, Maison et al. 2010). Excessively high or low levels of transcription lead to centromere inactivation and failures in chromosome segregation (Ling and Yuen 2019; Ohkuni and Kitagawa 2011). Several transcriptional regulators regulate cenRNA levels, including RNA Pol-II, ZFAT (human and mouse, Ishikura et al. 2020), Cbf1, H2A.Z^{Htz1} (budding yeast, Ling and Yuen 2019), MIWI and Dicer (mouse, Hsieh, Xia and Lin 2020).

During mitosis, most regions within the condensed heterochromatin are transcriptionally inactive, while centromeric regions remain active (Chan et al. 2012, Liu et al. 2015). RNA pol II is responsible for this activity and R-loops, a by-product of DNA-RNA hybridization, are necessary for faithful mitosis (Aze et al. 2016, Kabeche et al. 2018, Leclerc 2021). Non-coding RNAs are produced from each human alpha-

satellite array and are required for kinetochore assembly and *de novo* deposition of CENPA into chromatin (McNulty Sullivan and Sullivan 2017, Choi et al. 2011, Bobkov et al. 2018). These RNAs undergo post-transcriptional processing in mice to generate smaller RNAs (Bouzinba-Segard et al. 2006). In many eukaryotes, the RNAi machinery plays an important role in chromosome function (Gutbrod and Martienssen 2020). In *Cryptococcus* yeast, loss of the RNAi machinery triggers the attrition of centromeric retrotransposons, resulting in the shortening of centromere length (Yadav et al. 2018). In higher eukaryotes, the effect of the RNAi machinery is less clear. In mouse embryonic stem cells, the depletion of Dicer leads to an accumulation of centromeric transcripts (Kanellopoulou et al. 2005, Murchison et al. 2005), but in chicken-human hybrid cells, this leads to the diffusion of heterochromatin protein 1 (HP1) throughout the entire chromosome, as opposed to focusing on the centromere (Fukagawa al. 2004). Multiple proteins interact with these RNAs, including the H3K9 methyltransferase SUV39H1, SUV39H2, CENPA, HJURP, Aurora-B and H3K9 chaperone proteins (Johnson and Straight 2017, Maison et al. 2011, Quénet et al. 2014, Blower 2016). In *Xenopus* eggs, the centromeric region is transcribed as cenRNA which localises to mitotic centromeres, chromatin, and the spindle, allowing the activation of Aurora-B kinase (Blower, 2016). In mice, centromeric minor satellite RNAs yield transcripts up to 4 kb long and are required for Aurora-B kinase activity (Ferri et al. 2009, Bouzinba-Segard et al. 2006). A direct RNA-protein interaction between centromeric RNAs and CENPA has been found in many eukaryotes, and the loss of centromeric transcripts leads to the loss of CENPA and its chaperone HJURP to centromeres (Jansen et al. 2007, Dunleavy et al. 2012, Quénet et al. 2014). In plants, centromeric transcription has been identified in maize and *Sorghum* within Ty*1*/Copia elements (Miller et al. 1998, Jiang et al. 2003), centromeric retrotransposon of rice (Neumann, Yan and Jiang 2007), and Arabidopsis *CEN180* centromeric satellite repeats (May et al. 2005).

1.15 Centromere evolution

Centromeres evolve rapidly, either due to adaptive evolution that increases the chance of their inheritance (drive), or lack of constraint on its structure and position (drift) (Melters et al. 2013, Garrido-Ramos 2017). In sexually reproducing species that

undergo meiosis, satellite DNA is predicted to be different between populations but similar within populations, with sexual reproduction driving and fixating sequence variants, while in asexual species it is predicted to be different within a population through individual homogenisation by biased gene conversion (Dover 1986, Cesari et al. 2003, Langley et al. 2019). In addition to centromere drive in female meiosis, other evolutionary forces may select for centromere DNA and protein variants. These forces include selection for non-segregation-related functions, and invasion of centromeres by transposons and extraneous genetic elements (Kumon and Lampson 2022).

1.15.1 Centromere drive

Based on the behaviour of dicentric chromosomes, Sears and Câmara (1952) suggested that centromeres can vary in their strength which can cause preferential segregation of the stronger centromere through cell division. In many eukaryotic lineages, only one meiotic product produces a functional gamete, and the other haploid cells are degraded, which occurs most frequently following female meiosis (Gorelick et al. 2017). In some cases, centromeres appear to be adapted to preferentially segregate into the surviving functional gamete. The underlying centromere drive mechanisms depend on coupling spindle asymmetry to cell division, and asymmetry between centromeres of homologous chromosomes to respond to the spindle asymmetry (Akera et al. 2017). A model was proposed for the evolution of centromeres and kinetochore proteins which implies that satellite arrays differ genetically in their ability to recruit kinetochore proteins, and that kinetochore proteins coevolved to suppress the drive by restoring parity between homologs during chromosome segregation (Henikoff et al. 2001). Indeed, strong experimental evidence for centromere drive has been found in specific hybrids of monkeyflower or mice (Fishman and Willis 2005, Finseth et al. 2015, Chmátal et al. 2014, Iwata-Otsubo et al. 2017, de Villena and Sapienza 2001). For example, the iron mountain population of the wild monkeyflower *Mimulus guttatus* shows dramatic examples of centromere drive (Fishman and Willis 2005). The driving allele *D* is associated with a large expansion on chromosome 11 of the satellite repeat Cent728, found at all centromeres (Fishman and Willis 2008). Allele *D-*, which lacks the satellite expansion, has a low frequency in

the iron mountain population, potentially suggesting coevolved conspecific suppression of drive (Finseth et al. 2021, Talbert and Henikoff 2022).

Support for centromere drive can be found in the analysis of Robertsonian translocations in human or *Dichroplus pratensis*, which fuse two acrocentric or telocentric chromosomes at their centromeres (de Villena and Sapienza 2001, Bidau and Marti 2004). These translocations are preferentially transmitted to the egg cell, compared to native acrocentric chromosomes (Schueler et al. 2010). Positive selection was also observed on the N-terminal tail of CenH3 in Drosophila and several other plants and animals, which were interpreted as DNA-binding domains (Malik et al. 2002, Cooper et al. 2004, Talbert and Henikoff 2022). In several species of Drosophila and mosquitos, CenH3 has been duplicated and may have tissue-specific roles (Kursel Welsh and Malik 2017, Teixeira et al. 2018). This could allow CENH3 paralogs to resolve intralocus conflicts between maternal and paternal centromeric requirements and may allow independent adaptation of paralogs to the deleterious effects of centromere drive (Kursel Welsh and Malik 2017, Kursel et al. 2021).

In mouse oocytes, spindle asymmetry in terms of microtubule post-translational modification (tyrosinylation) is intrinsically coupled to gamete fate asymmetry, and selfish genetic elements can exploit this spindle asymmetry to drive by preferentially orienting towards the detyrosinated side of the spindle (Akera, Trimm and Lampson 2019). Confirmation of centromere strength and preferential segregation on an asymmetric meiotic spindle comes largely from mice (Chmatal et al. 2014, 15). The stronger centromere of CF-1 mouse population is located closer to a cell pole, and the CHPO strain centromere with weaker centromeres has less CENPA, centromere protein B, and CENPC (Chmatal et al. 2014, Iwata-Otsubo et al. 2017). 46 genes were compared from 11 rodent genomes to test for positive selection that may indicate genes with a role in suppressing centromere drive (Kumon et al. 2021). Ten genes were identified, including the chaperone HJURP, the licensing factor KNL2/ M18BP1, the inner centromere protein (INCENP) and SGO2 (Kumon et al. 2021).

In *Arabidopsis thaliana*, full complementation of a null *cenh3* mutant with homologs from *Brassica rapa* and *Lepidium oleraceum* was observed (Maheshwari et al. 2015). However, when the transformed lines were backcrossed to wild type, abnormal

segregation and aneuploidies were observed, indicated by seed abortion and chromosomal rearrangements, with most aberrations coming from the chromosomes that inherited with the transgenic CENH3 (Maheshwari et al. 2015). Altered CENH3 is removed from the egg cell while the wild type variant is maintained in the hybrids, which might explain the differences in centromere strengths (Marimuthu et al. 2021). A similar process of genome elimination and haplotype induction has been observed in barley inter-species hybrids (Sanei et al. 2011), maize *cenh3* heterozygote backcrosses (Wang et al. 2021), and wheat CRISPR-induced *cenh3* mutants in heteroallelic combinations with homoeolog A being in a heterozygous state (Lv et al. 2020). These studies provide direct evidence for the strong effect that CENH3 protein sequence can have on genome inheritance in hybrids.

Plasmids and B chromosomes are extraneous genetic elements that exploit the host replication and segregation machinery for their own inheritance (Jones 1991, Rizvi 2017). B chromosomes are devoid of coding genes and mostly composed of tandem repeats, such as satellite DNA and ribosomal DNA (Camacho et al. 2000). Although plasmids and transposons are also present in bacteria and archaea, only eukaryotes have developed complex centromeres (Jun and Mulder 2006). Eukaryotes require chromosome segregation machinery for meiosis, which makes the associated machinery indispensable and provides an opportunity for selfish genetic elements to cheat (Lenormand et al. 2016). An evolutionary arms race between selfish extraneous genetic elements and centromere binding proteins can potentially lead to rapid evolution of both types of sequence (Kumon and Lampson 2022).

Although centromere position on the chromosome varies, many species have either mostly telocentric or mostly metacentric chromosomes, where centromeres are localised towards the end or middle of the chromosome, respectively (de Villena and Sapienza 2001). Karyotypes have also switched between mostly telocentric and mostly metacentric states (Britton-Davidian et al. 2000, White, Bordewich and Searle 2010). Robertsonian fusions can cause transitions in one direction, from telocentric to metacentric, and were observed to happen in whole populations (Britton-Davidian et al. 2005, Molina et al. 2014, de Villena and Sapienza 2001). Centromere repositioning has been also observed in non-repetitive centromeres of horse (Purgato

et al. 2015), fission yeast (Yao et al. 2013), and between individual chicken cell lines (Hori et al. 2017). Interestingly, when those lines were propagated, the centromere maintained a stable position as measured by CENPA ChIP-seq over 50 generations in wild-type cells, but in the CENPC deficient background centromere repositioning was observed (Hori et al. 2017). This indicates that the proteins of the centromere may constrain the position of CENPA nucleosomes.

1.15.2 Satellite DNA evolution

The formation and evolution of centromeric satellite arrays has been proposed to occur via unequal recombination (Smith 1976). In the unequal exchange model, tandem duplications can be generated by random mutation, followed by unequal exchange between sister chromatids or homologs, generating further reciprocal duplications and deletions of tandem repeat arrays (Smith 1976). With a high enough recombination rate, significant homogeneity of tandem repeat arrays can be maintained in the face of mutation. Dover (1982) proposed that gene conversion, unequal exchange, and transposition are processes that lead to turnover of DNA and may lead to "accidental speciation" due to incompatibility of repeat arrays between separate populations that have diverged (Yunis and Yasmineh 1971, Ferree and Prasad et al. 2012). The discovery of the single-stranded annealing pathway (Lin et al. 1984), suggested that tandem satellite arrays may shrink over time unless a process favouring expansion counteracted this repair pathway (Henikoff et al. 2001). However, the widespread persistence of centromeric satellite arrays indicates that pathways exist that actively create and change these repeats.

Human satellite higher order repeats (HORs), complex nested structures of tandemly repeated monomer blocks, have been proposed to be generated by break-induced replication (BIR), in which the constitutive centromere associated network (CCAN) presents a barrier to replication, resulting in fork pausing and collapse, creating a single-ended DSB. BIR is an alternative homologous recombination DSB repair mechanism, in which one of the single ends of the DSB fails to engage with homologous sequence, which can lead to a non-specific strand invasion (Greenfeder and Newlon 1992, Kramara, Osia, Malkova 2018). In the BIR model, HORs go through a life cycle starting with n/B dimers, which are favoured by centromere drive (Rice et

al. 2020). As HORs increase in length, via BIR, they are also more likely to acquire CENPB box mutations, additional n-box monomers, or other divergences that make them susceptible to replacement by younger HORs (Gamba and Fachinetti 2020). BIR repair may account for the rapid divergence of centromeric HORs at the nucleotide level, which is greater than 10 times the divergence on chromosome arms between humans and chimps (Rice et al. 2020). The BIR model is supported by the recently completed T2T assembly of human Chromosome 8, which shows a symmetrical satellite array with four or five layers of evolutionary structure (Logsdon et al. 2021). The investigators found that humans and chimps share a common ancestor for the monomers in their flanking pericentromeres, and that the highly identical repeats in the q arm of apes appear to displace older repeats out of the centromere, creating a remnant of the ancestral centromere on the flanking edges. As an exception, the 632 kb region of CENPA on chromosome 8 was found in the adjacent fourth layer in a region of great admixture of HOR types (Logsdon et al. 2021).

The BIR model proposes that n/B dimers were acquired through centromere drive, and that CENPB strengthens the kinetochore (Talbert and Henikoff 2022). CENPB is conserved throughout mammals, but CENPB boxes are only present in some mammalian clades (Gamba and Fachinetti 2020). As compared to other centromeres, neocentromeres and the Y chromosome centromere, which lack CENPB boxes, have lower levels of CENPC and higher levels of chromosome mis-segregation, which is consistent with the view that CENPB leads to stronger centromeres, which favours centromere drive (Fachinetti et al. 2015). Existence of CENPB antigen has been described in plants (Barbosa-Cisneros et al. 2002), and *CENPB box*-like domains were found (Weide et al. 1998), but no homolog of CENPB has been identified in plants.

1.15.3 Neocentromeres

Centromeres formed at ectopic locations (neocentromeres) have been described in human (Voullaire et al. 1993), barley (Nasuda et al. 2005) and *D. melanogaster* cells (Williams et al. 1998), when the endogenous centromere is deleted or inactivated. Neocentromere formation over non-repetitive regions of human chromosomes demonstrated that genomic alpha-satellite sequences alone are insufficient to determine centromere location, supporting an epigenetic or chromatin-related

component for centromere function (Sart et al. 1997). Once formed, neocentromeres may either go extinct or increase in frequency by drift or drive, eventually leading to fixation as evolutionary young centromeres (Rocchi et al. 2011). Neocentromeres may also behave selfishly by evolving to recruit centromere proteins to exploit the kinetochore pathway (Kumon and Lampson 2022). Neocentromeres can arise on chromosome fragments, or rearranged chromosomes, where the canonical centromere is no longer functional, or has been removed (Barra and Fachinetti 2018). Satellite DNA can also be used to build human centromeres *de novo* and generate human artificial chromosomes (HACs) (Harrington et al. 1997). These HACs can be inherited by human cells and can be used to study the mechanisms by which satellite DNA initiates centromere formation (Ohzeki et al. 2015). Some centromeres in closely related species adopt new positions over evolutionary time, without transposing the surrounding genetic markers, for example the repositioned centromeres of *A. alpina* were found in similar genetic environments to those found in homoeologous chromosomes of *A. lyrata* (Mandakova et al. 2020). In maize with stable dicentric chromosomes created during a breakage–fusion–bridge process (McClintock 1941), only one of the centromeres is active (Sullivan and Willard 1998), although reactivation of the inactive centromere has been described (Fu et al. 2012).

1.16 Plant centromeres

Most plants are monocentric and many of them contain large arrays of centromeric tandem repeats (Comai, Maheshwari and Marimuthu 2017). Many described plant centromeric satellites range from 150 to 180 bp in monomer size and can occupy several kilobase- to megabase-sized regions, which have been detected in several plant genomes, including *Arabidopsis thaliana* (Nagaki and Murata 2003, Naish et al. 2021), *Oryza sativa* (Lee et al. 2005), and *Triticum aestivum* (Su et al. 2019). *Arabidopsis thaliana* is a major model plant and attempts at sequencing its genome started over two decades ago (Lin 1999). However, the centromere sequence assembly has remained incomplete because of the high repetition and similarity of centromeric satellite arrays. Recently, a number of high-quality assemblies of *Arabidopsis thaliana* have been published using long-read sequencing technologies: Col-CEN (Naish et al. 2021), Col-XJTU (Wang et al. 2021) and Col-PEK (Hou et al.

2022). *Arabidopsis thaliana* centromeres contain millions of base pairs of the *CEN180* satellite repeat, which support CENH3 loading (Maheshwari et al. 2017, Naish et al. 2021). Only a fraction of the total 180-bp repeats are bound by CENH3, suggesting that only subsets of the 180-bp satellite arrays are involved in centromere function, similar to the human deposition patterns described previously (Alexandrov et al. 2001, Maheshwari et al. 2017, Rice et al. 2020, Naish et al. 2021).

Plant genomes may increase in size through the rapid accumulation of LTR retroelement transposons (Lee and Kim 2014). For example, the activity of a single family of CR1 elements in Cap*sella* genus can have drastic effects on genome size (Slotte et al. 2013, Ågren et al. 2014, Ågren Huang and Wright 2016). The development of genomic and epigenomic methodologies has enabled the massively parallel assessment of the epiallelic potential of transposon-containing alleles in plant genomes (Baduel et al. 2021). The epiallelic nature and inheritance of the strong hypomethylation induced mostly in CG sequence contexts by *met1*, or at both CG and non-CG contexts by *ddm1*, was evaluated in *Arabidopsis thaliana* (Reinders et al. 2009, Johannes et al. 2009). One-third of hypomethylated transposon sequences in the *ddm1* parental line were inherited in the hypomethylated state across at least eight generations, and two-thirds regained wild type methylation progressively, within three to five generations, and in either some or all the epiRILs that contain corresponding *ddm1*-derived chromosome intervals (Rigal et al. 2016).

Using large amounts of data from natural *A. thaliana* accessions, genome-wide association studies (GWAS) have identified major *trans* modifiers of DNA methylation variation at transposon sequences (Dubin et al. 2015, Sasaki et al. 2019). These *trans* modifiers are related to RNA-directed DNA methylation (RdDM) and other DNA methylation pathways that target transposable elements (Dubin et al. 2015, Sasaki et al. 2019). DNA methylomes have been obtained for a number of mutation accumulation lines in *A. thaliana* and show that spontaneous heritable epimutations occur at CGs at a rate several orders of magnitude greater than that of genetic point mutations: 30,000 differentially methylated positions were identified vs 30 DNA sequence mutations per strain. (Ossowski et al. 2010, Becker et al. 2011). Spontaneous epimutations in transposon sequences predominantly result in a loss,

rather than a gain of methylation, and occur at rates per methylated region that are orders of magnitude higher than the rate of mutations per nucleotide (Schmitz et al. 2011). In addition to being generated spontaneously, epimutations could potentially be induced by exposure to environmental stresses (Jianget al. 2014). These epigenetic changes, often affecting transposon sequences, are less stably inherited than those resulting from spontaneous epimutations that affect all cytosine sequence contexts (Wibowo et al. 2016, Baduel et al. 2021). In *A. thaliana*, impaired RdDM is sufficient to induce transposition for several transposon families (Herr et al. 2005, Pontier et al. 2012), and natural alleles in these pathways are predominantly found in the extreme ecological environments present at the edge of the species niche (Ito 2012, Baduel and Quadrana 2021).

1.17 Project aims and objectives

Most of the understanding of the centromeres comes from an era where accurate genetic maps were not available for most species. This is because centromeric repeat arrays, common across eukaryotic species, have been challenging to sequence and assemble using the previous generations of sequencing techniques (Amarasinghe et al. 2020). Long-read sequencing technologies have overcome this challenge and an increasing number of high-quality genomic assemblies are being released (Naish et al. 2021, Nurk et al. 2022). This opens up questions about the structure of the centromeric arrays and the processes driving their evolution, and their systematic analysis requires specialised bioinformatics tools. Additionally, *HEI10* dosage has the ability to regulate the level of meiotic recombination affecting the genomic organisation, although its limits and activity beyond *Arabidopsis* are poorly understood. Therefore, my PhD project aims to address the *in-silico* mapping challenges of centromeric repeats, by developing novel tools and techniques in centromere analysis, and to test *HEI10* overexpression in tomato and further characterise it in *Arabidopsis*. In more detail, this project aims to:

A) Develop a novel bioinformatic tool able to *de novo* identify, classify, and characterise tandem repeats and their higher order organisation, with particular emphasis on intuitive usage and capability of automatic and *de novo* analysis, in order to facilitate high-throughput analysis.

B) Characterise centromeric regions of model organism *Arabidopsis thaliana*, its close relative *Arabidopsis lyrata*, the more distant relative *Brassica oleracea*, and holocentric *Rhynchospora* species, in order to deepen the understanding of plant centromere organisation and evolution, assess the functionality of the tool described in the previous aim and develop additional methods of analysis and comparison of centromeric regions.

C) Test for *HEI10* dosage effects on meiotic crossover recombination in tomato by *Agrobacterium* mediated transformation and overexpression, and to model the recombination increase in *Arabidopsis thaliana* by combining *HEI10* overexpressing lines.

Chapter 2

Materials and Methods

2.1 Plant methods

2.1.1 Plant material

Arabidopsis thaliana ecotype seeds Col-0 were used in these experiments, originally obtained from the Nottingham Arabidopsis Stock Centre (NASC). FTL *420* was used for crossover frequency measurements (Ziolkowski et al. 2015). T_2 seeds of *HEI10* overexpressing lines were obtained from Dr Piotr Ziolkowski. Plants were grown in growth chambers at 20ºC with long day 16/8 hour light/dark photoperiods, 60% humidity and 150 µmol light intensity. Seeds were stratified for 3 days at 4ºC prior to germination.

Solanum lycopersicum cultivar Heinz accession 1706 and cultivar Micro-Tom accession LA3911 were obtained from the Tomato Genetics Resource Centre (TGRC). Plants were grown at 26ºC 16 hour light/ 21ºC 8 hour dark photoperiods, with 75% humidity and 400 µmol light intensity.

2.1.2 Tomato seed extraction

Tomato seeds were cleaned by adding 1:1 volume of 6M HCl to the seeds (around 20ml) extracted from the fruits, followed by 20 min incubation with periodic shaking, rinse with water and incubation with 1 volume of 200mM Na-phosphate pH 7.2. After 20 minutes of incubation with periodic shaking, seeds were rinsed with water and left overnight to dry.

2.1.3 Automatic measurements of crossover frequency in *Arabidopsis thaliana*

Crossover frequency in specific intervals can be measured using Fluorescent Tagged Lines (FTLs) (Melamed-Bessudo et al. 2005, Ziolkowski et al. 2015). FTLs are characterised by expression of two different fluorescent proteins (RFP and GFP) from T-DNAs linked on the same chromosome. The fluorescent proteins are expressed during seed development from the seed specific NapA promoter and absorbance of the respective wavelengths can be measured to establish the presence of each marker T-DNA. Fluorescence micrographs taken at each wavelength are analysed using CellProfiler software to count the number of seeds expressing each marker, later called green and red seeds. Some of the seeds from the next generation will display a single-colour phenotype, from the ratio of which recombination can be measured, using the equation:

$$\boldsymbol{cM} = \left(\mathbf{1} - \sqrt{\mathbf{1} - \mathbf{2} * \left(\frac{\boldsymbol{nG} + \boldsymbol{nR}}{\boldsymbol{nT}} \right)}\right) * \mathbf{100}\%$$

Where cM relates to crossover frequency, nG is count of green-only seeds, nR is count of red-only seeds and nT is a total count of all seeds. Line 420 was used in many experiments, which measured 19.71 cM in a Col/Col inbred background. The physical distance on chromosome 3 between the *420* T-DNAs is 5.105 Mb.

The CellProfiler pipeline used for automatic seed counting was developed by Dr Piotr Ziolkowski (Fig. 2.1, Ziolkowski et al. 2015). After correcting the picture quality (luminescence distribution), it identifies seed objects and measures their intensity on the “red” and “green” image. Next, a histogram of mean intensity is displayed (similar to Fig. 2.1F), and the user manually chooses a threshold value between non-fluorescent and fluorescent seeds, based on the plot. This value is then used as an input for the next method, which will score the number of seeds above and below this threshold. To avoid arbitrary user-picked thresholds and make the process fully automatic, I created an alternative pipeline that automatically calculates the image pixel intensity histogram and performs Otsu’s method, which looks for the threshold value that would minimise the intra-class variance, defined as a weighted sum of variances between two classes (Otsu, 1979). As a result, the picture is divided into

classes: (i) background and non-fluorescent seeds, giving this class pixel value = 0, and (ii) fluorescent seeds, setting pixel value to 1 Fig. 2.1D). Later the masks are aligned with identified seeds and when an overlap of at least 10% is present, a seed is classified as fluorescent (Fig. 2.1E). A relatively small value of 10% was chosen, because seeds, being spherical objects, will have the highest intensity of fluorescence in the middle. This way, even seeds with poor fluorescence quality can be adequately scored. After identification of fluorescent seeds, the next method in the pipeline masks “red” seeds against “green”, to obtain the number of seeds with both markers present, which is necessary for subsequent crossover frequency calculations (Fig. 2.1F-H). The cM measurements calculated from the automatic protocol were not significantly different from the values from manual protocol (paired t.test with double-tailed distribution = 0.41, p = 0.34).

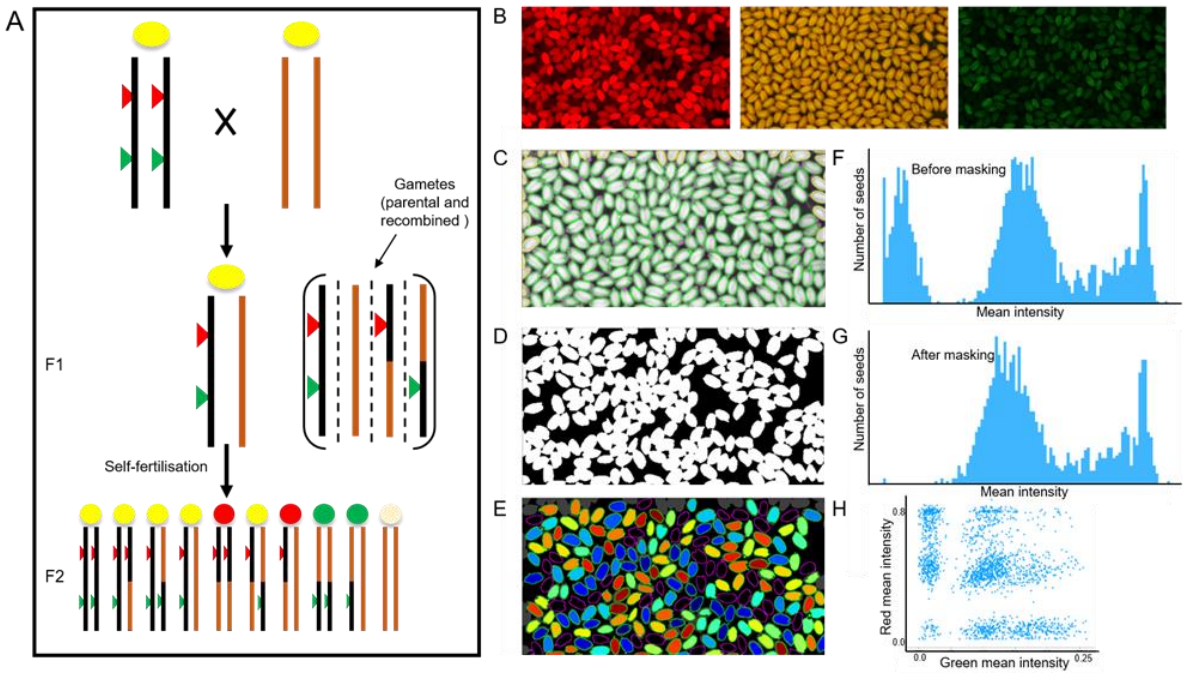


Figure 2.1. Crossover frequency measurements using Fluorescent Tagged Lines (FTL)

A. Molecular basis of FTL recombination measurements. **B**. Examples of images used for processing in Cellprofiler. **C**. Seed object identification. **D**. Thresholding intensity and creating a binary image. **E**. identifying fluorescent seeds based on the binary mask. **F**. Histogram of fluorescence intensity of all seeds. **G**. Histogram of fluorescence intensity of seeds identified to be fluorescent. **H**. Two-dimensional diagram of intensities of red fluorescence (y axis) and green fluorescence (x axis) of all identified seeds as a measure of picture quality.

2.1.4 Progeny testing for antibiotic resistance

Seeds (100 per plate) were surface sterilised by incubation in 50% bleach with 50 µl/L Tween 20 for 7 minutes with gentle shaking. They were rinsed 6 times before sowing on 1% phytoagar with ½ x Murashige and Skoog (MS) salts medium plate with 50 µg/ml Kanamycin. Plates were kept in darkness in 4ºC for 3 days and transferred to a growth chamber with 23ºC, 120-150 µmol/m2s, 16/8 light/dark photoperiod. Phenotype was inspected around 10 days after germination. Plants that were pale exhibited kanamycin toxicity phenotype. Ratio of plants with phenotype to WT phenotype was calculated.

2.2 Molecular biology methods

2.2.1 Cloning of tomato *HEI10* vector

The pCAMBIA1300 based pFGC-pcoCas9 vector was used as a backbone for the construct. pFGC-pcoCas9 was a gift from Prof. Jen Sheen (Addgene plasmid # 52256). The vector was linearised using PCR primers P6F (5'-CCCCTCCATGGAGCCCTTTGGTCTTCTGAGAC-3') and P4R (5'-TGTGCACTAGTGCAGATCGTTCAAACATTTGGC-3'), which resulted in an 8,387 bp product containing the backbone with the bar gene for phosphinothricin (PTT) selection. 7,193 bp insert was amplified from *Solanum lycopersicum* M82 cultivar genomic DNA using primers P1F (5'-TCTGCACTAGTGCACAGAGGAGGTCCATGTAC-3') and P1R (5'-GGGCTCCATGGAGGGGTGAAGAATCTTGGACG-3') (Fig. 2.2). All primers contained 8-base pairs long 3' adapters complementary to the 5' ends of the opposing primers in the cloning reaction, for the total of 18 base pairs overlap. Both PCR reactions were performed using CloneAmp™ HiFi PCR Premix (Takara) high fidelity polymerase using recommended conditions, confirmed by electrophoresis, and purified using Monarch® PCR & DNA Cleanup Kit (NEB). Cloning was performed using Gibson Assembly® method using ClonExpress Ultra One Step Cloning Kit (Vazyme) according to the manufacturer's instructions and the reaction was used to transform *E. coli* DH5α® chemically competent cells. These were plated onto LB plates containing 50 µg/ml kanamycin and grown overnight at 37°C. Colonies were assessed by colony-PCR reactions using DreamTaq™ polymerase, with template DNA

added by gently touching a colony with a tip and mixing it into the prepared PCR reaction. Primers for colony PCR were chosen for the amplicon to overlap the ligation site: M13F (5'-TGTAAAACGACGGCCAGT-3') and SlHEI10RB (5'-CAAGTGGGGGCAGTTTATTTC-3'), for a 168 bp product. Colonies with amplicons of the predicted size were used to inoculate liquid LB containing 50 µg/ml kanamycin grown overnight at 37°C prior to plasmid extraction using the Monarch® Plasmid Miniprep Kit (NEB). Purified plasmid DNA was checked by restriction enzyme digestion using *Acl*I and sent for Sanger sequencing at Source Bioscience.

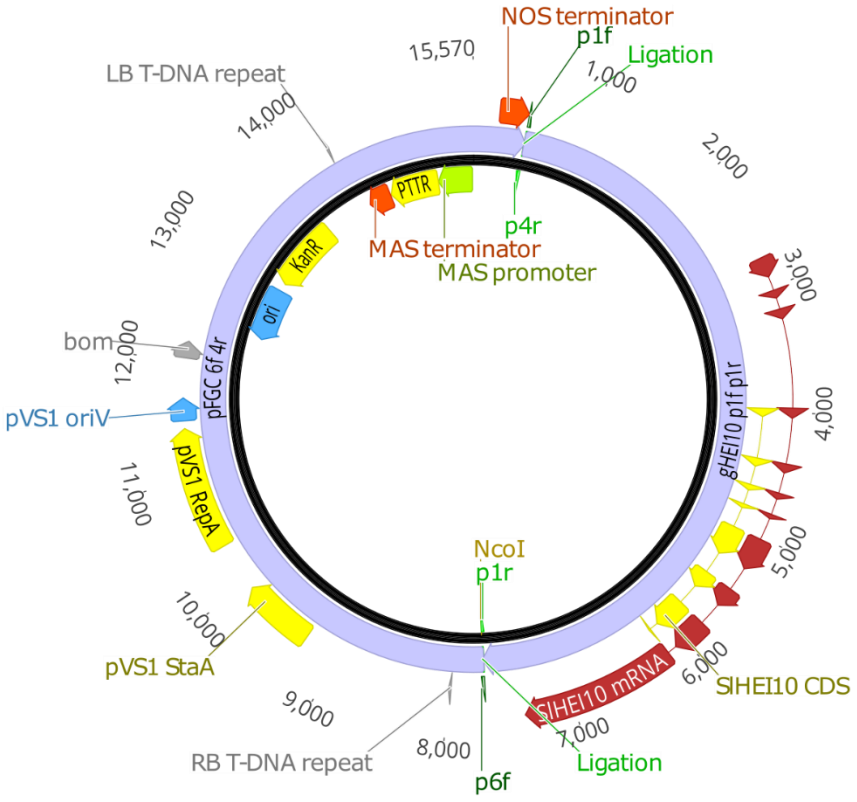


Figure 2.2. Tomato *HEI10* overexpression vector map
pCAMBIA1300-derived genomic *HEI10* carrying vector has been constructed using the Gibson Assembly® method after PCR amplification of 7,193 bp long fragment containing tomato *HEI10* and an 8,387 bp long fragment containing the vector backbone. Resulting binary vector confers resistance to kanamycin (Kan) in bacteria and to phosphinothricin (PTT) in plants.

2.2.2 Preparation of *A. tumefaciens* culture for tomato transformation

After confirmation of the plasmid sequence, plasmids were used for electroporation mediated Agrobacterium transformation. 100 pg of DNA was added to 50 ul of *A. tumefaciens* GV3101 cells, placed in a 1 mm electroporation cuvette and inserted into Gene Pulser Xcell Electroporation System (Bio-Rad) for transformation using a predefined bacterial protocol. 700 µl of liquid LB medium was immediately added to the cuvette, mixed, and transferred to a 1.5 ml tube, which was incubated at 28°C for 3h with 700 rpm shaking using Eppendorf™ ThermoMixer. After that, cells were pulse-centrifuged, most of the supernatant removed, the pellet resuspended in the remaining 100 ul volume and plated onto LB plates containing 50 µl/ml Kanamycin, 25 µg/ml Rifampicin and 30 µg/ml Gentamicin. After 3 days of incubation at 28°C individual colonies were used to inoculate 4 ml of LB with the same antibiotics added. This was further used to inoculate 50 ml of LB after 2 days at 28°C with 200 rpm shaking. Cells were collected by 10 min centrifugation at 3,000 G, resuspended in 5% sucrose until OD600 reached 0.4-0.5 AU and used immediately in the tomato transformation.

2.2.3 Tomato transformation

Transformation of tomato was performed using a modified protocol described by McCormick et al. 1986. Briefly, surface-sterilised seeds were germinated on 1 x MS salts medium with 1% phytoagar. Cotyledons of 6-day old plants were carefully cut, to acquire square 5-10 mm in size explants. These were submerged in *Agrobacterium* suspensions (5% sucrose, OD600 = 0.4–0.5) for 5-30 seconds and blotted on a sterile filter paper to dry. Next, they were transferred to feeder plates containing 0.6% phytoagar, 1xMS medium supplemented with 0.6 mg/L 2,4- Dichlorophenoxyacetic acid. After 40 hours, explants were transferred to regeneration plates containing 1xMS salts, 100 mg/L myo-inositol, 1xNitsch vitamins, 20 g/L sucrose, 4 g/L phytoagar, 500 µg/ml augmentin and 15 µg/ml phosphinothricin (PPT) at pH 6.0. After regeneration from callus, shoots were cut from the explants and transferred into rooting medium containing ½xMS, 5g/L sucrose and 2.25 g/L gelrite at pH 6.0

2.2.4 DNA extraction for tomato transgene and *Arabidopsis* genotyping

A 2-mm wide leaf tissue disc was collected during the rooting stage for the tomato transgene genotyping or from the rosette leaf for the Arabidopsis *HEI10* transgene genotyping. Genomic DNA for genotyping was extracted using the protocol described by Edwards et al. 1991, which was modified for a 96-well plate format. Plant tissue was disrupted using 3 mm borosilicate glass beads in 200 µl of extraction buffer (200 mM Tris-HCl pH 7.5, 250 mM NaCl, 25 mM EDTA) with a TissueLyser II (QIAGEN) for 2 min at 30 Hz. A further 200 µl of extraction buffer supplemented with 1% SDS was added, and the plate was centrifuged at 3,000 G for 5 minutes. The supernatant was transferred to a new plate and one volume (ca. 350 µl) of isopropanol was added and left to precipitate for 10 minutes at room temperature. After centrifugation for 35 min at 3,000 G, the pellet was washed with 150 µl ethanol (70%) before being left to dry and resuspended in 100 µl of water.

2.2.5 CTAB DNA extraction

For transgene mapping and qPCR applications, high quality gDNA was extracted from 4–6-week-old tomato plants using a protocol adapted from Clarke 2009. 2-3 1 cm long leaves or leaf cuttings were collected into 2 ml Eppendorf tubes containing 4 glass beads (3 mm). Samples were snap frozen in liquid nitrogen and stored at -80 °C before grinding in a TissueLyser II (Qiagen) for 2 rounds of 2 min (30 Hz), changing the positions of the samples and cooling them in liquid nitrogen in between to ensure even and complete disruption. 700 µl of CTAB buffer (140 mM sorbitol, 220 mM Tris pH 8.0, 22 mM EDTA, 800 mM NaCl, 0.1 % (v / v) N-Lauryl sarcosine, 4 % (w/v) CTAB (cetyl trimethyl ammonium bromide)) warmed up to 60°C was added to each Eppendorf, which were immediately inverted until resuspension of the plant material was complete. Samples were incubated for 30 minutes at 65 °C with 700 rpm mixing in a Thermomixer (Eppendorf), with mixing by inversion after 15 minutes. After cooling to room temperature, samples were pulse-spin centrifuged to pellet insoluble debris. 650 µl of the supernatant was transferred to a fresh 2 ml Eppendorf containing an equal volume of chloroform (650 µl), and vortexed vigorously until mixed. Samples were then centrifuged at 13,000 G for 5 minutes at room temperature. 550 µl of the upper aqueous layer was removed to a new 1.5 ml Eppendorf containing an equal volume of isopropanol. Tubes were vortexed, left at room temperature for 5 minutes and then

centrifuged at 13,000 G for 20 minutes at 4 °C. The supernatant was poured off and the pellet washed with 500 µl of 70 % ethanol and centrifuged at 13,000 G, at 4 °C. The supernatant was poured off and the Eppendorf left to air dry for 20 minutes at room temperature. The pellet was resuspended in 100 µl water containing RNase A (1 µl 100 mg / ml RNase A per 1 ml water) and incubated at 37 °C for 30 minutes. gDNA was precipitated by addition of 0.1 volumes of 3 M Sodium Acetate Solution (11 µl) and 2.5 volumes of 100 % ethanol prior to incubation at -20 °C for at least 30 minutes. The tubes were centrifuged at 13,000 G for 15 minutes at 4 °C. The supernatant was poured off and the pellet washed with 500 ul of 70% ethanol before pouring off, pipetting out remaining supernatant and allowing it to air dry for 20 minutes at room temperature. The final pellet was resuspended in 20 µl of water.

2.2.6 RNA extraction and cDNA synthesis for tomato and *Arabidopsis* expression assays

Extraction of RNA from A. thaliana buds and leaves was performed using TRIzol™ Reagent (Invitrogen) according to the manufacturer's instructions. *Arabdiopsis* bud tissue was collected by separating the central, youngest buds from 4-10 inflorescences and snap-frozen in liquid nitrogen. cDNA synthesis was performed using SuperScript™ IV Reverse Transcriptase (Invitrogen) according to the manufacturer's instructions. Random hexamers were used for the priming reaction, to avoid potential incomplete reverse transcription when oligo-dT primers are used. Because tomato meiotic expression measurement has been poorly described in literature, and a limited number of buds are available for collection at any time from a single plant, buds were divided based on their size and tested for *HEI10* and *DMC1* meiotic genes expression to find optimal collection conditions (Fig. 2.3).

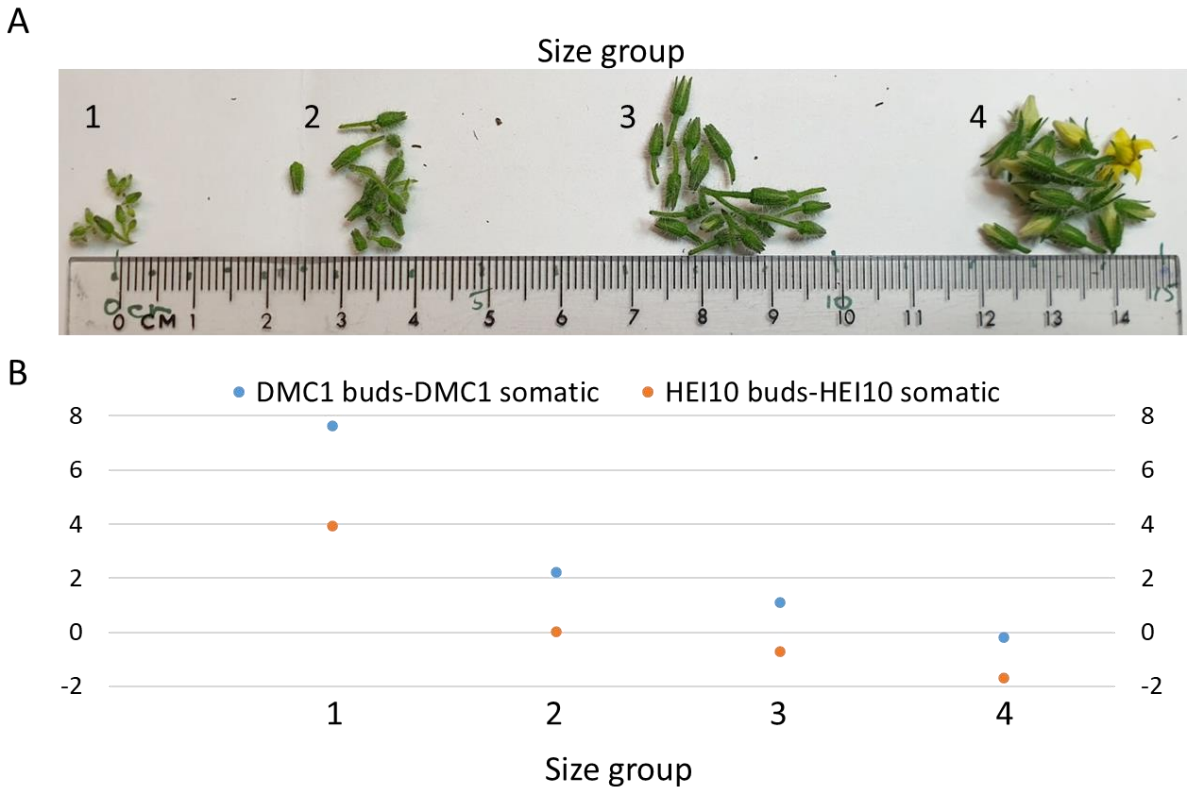


Figure 2.3. Size of tomato buds in collection for meiotic RNA enrichment
A. Tomato buds collected from 6 wild type Micro-Tom plants divided into 4 groups based on their size. Each group was divided into 3 replicates for extraction and cDNA synthesis **B**. Relative expression (ΔCt) of *HEI10* and *DMC1* from tomato buds at different developmental stages. Measurements were normalised with somatic base expression levels using RNA extracted from leaf tissue and with input RNA in cDNA synthesis reaction.

Since meiotic expression of both *DMC1* and *HEI10* was the highest in the smallest buds, but at the same time 1-3 buds were sufficient to acquire sufficient amounts of cDNA, these collection conditions were used for all tomato extractions.

2.2.7 Quantitative PCR for genomic copy number analysis

50-150 mg of leaf tissue was harvested, and DNA was extracted using the CTAB method. 0.25 ng/µl dilutions were used for qPRC reactions with with two primer pairs: the HEI10RB (5'-CTTTTTCACTCACTGCAAATACC-3') and M13ext (5'-AGGAAACAGCTATGACCATG-3') primers were used to amplify a 181 bp fragment specifically from HEI10 transgenic locus. ACT11-qF2 (5'-

GAGGCTCCATTCTAGCATCAC-3') and ACT11- qR2 (5'-GGACTATTGATGGCCCTGAC-3') create an amplicon of 180 bp from the *ACT11* loci which serves to normalise the sample, as copy number of this gene should be the same in all plants. 3 technical replicates were performed. 2^ΔΔCt values were calculated using the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001):

$$\Delta\Delta Ct = (Ct^{ACT11\ OX} - Ct^{HEI10\ OX}) - (Ct^{ACT11\ WT} - Ct^{HEI10\ WT}),$$

where results from HEI10 overexpressing plants are marked with "OX" and wild type plants with "WT".

2.2.8 Quantitative PCR for *HEI10* expression

Tomato and *Arabidopsis* qPCR reactions using the synthesised cDNA were performed using Luna® Universal qPCR Master Mix (NEB) on CFX96 thermal cycler (BioRad), following the manufacturer's instructions. Four biological samples and three technical replicates per sample were used for each experiment. *SlHEI10* was amplified using primers SlHq1_F (5'-GCTAAGAAGTGAGTATGAGTCAG-3') and SlHq1_R (5'-GAACTGTTCTGTCTTGCTGGC-3'), *SlDMC1* was amplified using primers SlDq1_F (5'-TGAAGAAACGAGCCAGATGC-3') and SlDq1_R (5'-GCATCACTTCCAGTCATATATCC-3'). Primer efficiency was determined beforehand using a serial dilution curve. The fold change in *HEI10* expression relative to *DMC1* was calculated using the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001) in both species:

$$\Delta\Delta Ct = (Ct^{DMC1\ OX} - Ct^{HEI10\ OX}) - (Ct^{DMC1\ WT} - Ct^{HEI10\ WT}),$$

where results from HEI10 overexpressing plants are marked with "OX" and wild type plants with "WT".

2.2.9 DNA genotyping

PCR genotyping was conducted using DreamTaq™ polymerase according to the manufacturer's instructions. Melting temperatures Tm for each primer pair was calculated using Geneious Prime® 2019 software and used as a starting point for PCR reaction optimization. Range of 5 temperatures was tested for each primer pair: Tm-3°C, Tm-1°C, Tm, Tm+1°C, Tm+3°C. PCR products were separated on an 1% agarose

gel (1×TBE, 1/10,000 Midori Green stain (Nippon Genetics) and visualised under UV light. SSLP markers PCR products were separated on a 3% agarose gel instead.

2.2.10 Nucleic acid quantification

Plasmid DNA after alkaline lysis-based extraction and gDNA extraction for genotyping was quantified using a NanoDrop™ 1000 Spectrophotometer (Thermo Scientific). For gDNA extracted using the CTAB method, a Broad Range DNA Qubit® Fluorometer (Life Technologies) was used. For DNA used in Next Generation Sequencing (NGS) library preparation, a High Sensitivity DNA Qubit® Fluorometer (Life Technologies) was used, in conjunction with an HS DNA Agilent 2100 Bioanalyzer system (Agilent Technologies). RNA used for cDNA synthesis was quantified using High Sensitivity RNA Qubit® Fluorometer (Life Technologies).

2.2.11 Transgene mapping by Tail-PCR

Mapping of tomato transgenes were performed using Thermal asymmetric interlaced PCR (Tail-PCR) (Liu and Whittier 1995), according to an updated protocol from Liu and Chen 2007. Gel-extracted PCR products were Sanger sequenced by submitting samples and primers to Source Bioscience or Azenta (GENEWIZ). Sequencing analysis was conducted using Geneious Prime® 2020.

2.3 Bioinformatics methods

2.3.1 Tomato SSLP markers design

Chromosome 5 was chosen for the Simple Sequence Length Polymorphism (SSLP) marker design based on its highest count of SNPs out of all 12 chromosomes (Kobayashi et al. 2014). 16 primer pairs were designed based on the Micro-Tom variant data called on the Heinz SL2.5 assembly (Kobayashi et al. 2014, The International Tomato Genome Sequencing Project) by manual searches using Geneious Prime® 201 software (9), with preference to exon locations. Since most mapped indels were relatively short, designed markers differed by only 6-42 bp in amplicon size between the Heinz and Micro-Tom templates. 12 further primer pairs

were designed after a Micro-Tom assembly became available (Genbank accession number GCA_012431665.1), which allowed chromosome 5 alignment with Heinz SL3.5 reference assembly using Mauve (Darling et al. 2004). Amplicons of these primer pairs differed by 61-936 bp. Primer positions, sequences and amplicon sizes are summarised in Table 2.1.

Name	Sequence (5'-3')	Heinz amplicon size (bp)	Micro-Tom amplicon size (bp)	F primer start coordinate (bp)
5.000-F	CCTGCATAAGTAATACTACC	280	341	10,510
5.000-R	ATTTGAGGGCGACTTCTC			
5.007-F	CTTCCCTCTATACGTTTCAG	92	98	754,869
5.007-R	CAATTGACAATGCAAAATACG			
5.002-F	GAACTACCAATGACCACC	138	359	1,879,633
5.002-R	AAGTGTGTTTGGTACTAGG			
5.019-F	CCTCATCAGGTCTCTAATAAAC	114	104	1,965,120
5.019-R	CTCGTGAAGTCTCCATAATG			
5.036-F	TTGTCATTATTTAATACCTACACC	149	135	3,662,251
5.036-R	TCTACTTATTTATTATTTTGAGATCG			
5.004-F	CAGGCACTTTGTTCATAATG	261	159	3,941,207
5.004-R	TCGATTAATTTGGGTTATCGT			
5.040-F	GGATGAAGACGAGATGTATTAAAG	87	98	4,034,641
5.040-R	GCTTCGTTGAGCTTCAAC			
5.049-F	CATTTTATGGCCGCCTTAG	65	78	4,972,711
5.049-R	ACCATGGCTGTCACTTTG			
5.057-F	GGACAAAAATGTTTGGCAAC	121	146	5,778,892
5.057-R	CCGAGATCATAAAATCAACTACC			
5.006-F	TGCAGGAAAGAAACTATCAC	920	202	5,909,290
5.006-R	TCTGGAAGCAAGATGGTG			
5.068-F	GTCCTCTCTTTGGTAGGTC	167	193	6,805,387
5.068-R	GAGGAGGTAGTTTTTGAAATAG			
5.076-F	CAATAACAACTAACCCTTTATTCC	80	61	7,621,162
5.076-R	TTTACTTATGGGGCTTTTGG			
5.008-F	TCTGGATAGAGCAAGTGTAC	165	196	7,965,010
5.008-R	CGCTTTACAGGTGTTTGG			
5.012-F	ATGCTCCCATCACATGA	335	273	11,867,728
5.012-R	TCAAGGCTTATTCAACATTTG			
5.055-F	CCAATTCAGTAGGCAATCTC	147	229	54,088,528
5.055-R	TCAAGCACTATTTCTCCAAG			
5.578-F	CAGAAGATAGGCAATGTTTAC	158	183	57,860,587
5.578-R	TTGTGAGATAAAGTTGATTTGC			
5.590-F	CAAGAATCATGCGACGAA	139	1002	57,956,150
5.590-R	CAAGCCAACATCCACAAG			
5.588-F	CACACTTATACTATACTAAGGTCC	93	51	58,884,055
5.588-R	ACAAGATAGTGCCACGTAG			
5.061-F	GGTCACACAAATGCACG	470	611	60,069,383
5.061-R	TGAAGATGTAGAAGATTCGG			
5.600-F	TATGTGTTTATTGTGCGCG	85	112	

5.600-R	GGTCACATTACAAGTGTCC			60,083,110
5.611-F 5.611-R	CATGTTTGGTGCCTAGTC AGCGATGAGATAAACTGTTC	131	154	61,106,644
5.615-F 5.615-R	TCGATACATTAGGTGTAAATTCG GAAGCAAATACACAACCATTC	138	166	61,867,142
5.620-F 5.620-R	GGGTGAGTAATTATGAAGAGC AGAGTGTTCATAGAGACCG	108	176	61,995,667
5.631-F 5.631-R	CCCCCCTAACCAATATGTG CTACAAAAATAGAGAGTCAACTAC	69	94	63,172,258
5.065-F 5.065-R	CATACCGCCATTTACGAT GGGCTGTGACATCCTTG	163	360	64,135,044
5.641-F 5.641-R	GTTGTTAAGTCATTAAAGGAAAG ACCATTGTATTGAATGATCCC	60	73	64,143,049
5.655-F 5.655-R	AGCAGTCAGATCCAGTAC ATCAAGGATTAGTGTAGGGA	111	88	65,504,915
5.670-F 5.670-R	TCACAAGAGAAAATGAGAAGC CACAAAGAAATTCAGATGCC	361	1297	66,603,266

Table 2.1. Simple sequence length polymorphism (SSLP) markers designed for tomato chromosome 5 Heinz x Micro-Tom genotyping.
Expected amplicon size from Heinz and Micro-Tom sequence and coordinates (bp) of the first nucleotide of the F primer according to the SL4.0 Heinz reference assembly are displayed.

2.3.2 Software and computing systems used

TRASH was developed using the R programming language (version 3.6.1) and R packages (collections of functions and compiled code) that expand its DNA sequence analysis capabilities:

- “remotes” by Gábor Csárdi et al. allows for installation of earlier releases of other packages, ensuring TRASH working as developed,
- “base", version "4.0.3”, the base R package,
- “stringr", version "1.4.0” by Hadley Wickham, a wrapper to “stringi” package (Gagolewski 2022) for fast string processing by using embedded c and c++ compiled code,
- “stringdist", version "0.9.8” by Mark van der Loo et al. (2014) for calculation of edit distance between strings,

- “circlize", version "0.4.15” by Zuguang Gu for compact sequence annotation visualisation,
- “seqinr", version "4.2.8” by Delphine Charif, predominantly used for fasta sequence handling,
- “doParallel", version "1.0.17” by Folashade Daniel for multithreading ability,
- “BiocManager", version "1.30.16” (http://www.bioconductor.org/), required to install and load “Biostrings” package,
- “Biostrings” (Pagès et al. 2022) for manipulation of biological sequences.

TRASH testing was performed using two high-performance computing clusters: Hydrogen in the Department of Plant Sciences of the University of Cambridge using the HTCondor job manager system, and Cambridge Service for Data-Driven Discovery (CSD3) using the Slurm job manager system. Both are Unix-like and no differences in the results were observed between the two systems. Code availability is described in Table 3.2

Analysis	GitHub repository	Comment	Status
TRASH	/vlothec/ TRASH	Main TRASH commit	in active development
Col-CEN	/vlothec/ col0610analysis	Repeat analysis for Naish et al. 2021	read-only
66 Arabidopsis thaliana accessions	/vlothec/ pancentromere	Repeat analysis of 66 *Arabidopsis* accessions. Contains code from collaborators: Fernando Rabanal, Robin Burns, Alexandros Bousios and Andrew Tock	in active development for publication purposes
Brassica oleracea	/vlothec/ bOle	Early development stage	in active development
Rhynchospora	/vlothec/ HoloRhynchospora	Early development stage	in active development
Other	/vlothec/ PhD	Other work presented in this thesis	read-only

Table 2.2. Public repositories containing software developed during work on this thesis

Chapter 3

TRASH: Tandem Repeat Annotation and Structural Hierarchy

In this chapter the development of a novel software called Tandem Repeat Annotation and Structural Hierarchy (TRASH) is presented. The main task of TRASH is to annotate tandem repeats in genomic assemblies. The need for such software arose with the increasing availability of long-read based assemblies, which include extensive regions of tandem repeats, one of which was generated in our lab by Dr Matthew Naish.

This chapter is based on a prepared manuscript entitled “TRASH: Tandem Repeat Annotation and Structural Hierarchy” which was submitted on 15th of October 2022 and currently is under revision. One of the authors, Michael Hong, wrote and implemented the circos plot module of the program. The results of analyses based on TRASH are presented in the next chapter.

3.1 Introduction

The extreme diversity of the centromeric DNA organisation necessitates a new depth of sequencing and analysis, which will not be restricted to individual genome analysis. Instead, a whole population from a species might need to be sequenced at the highest level, to discover the nuance variations that can lead to the description of the evolutionary mechanisms standing behind them. Due to their sequence repetition, it has been challenging to correctly assemble large tandem repeat arrays like ribosomal DNA coding, telomeric or centromeric arrays (Miga and Sullivan 2021, Rabanal et al. 2022). However, the advent of long-read DNA sequencing technologies, including Oxford Nanopore and PacBio HiFi, have allowed accurate and complete assembly of

complex satellite arrays for the first time (Miga et al. 2020, Logsdon et al. 2021, Naish et al. 2021, Altemose et al. 2022, Nurk et al. 2022). Further improvements to these technologies might allow for sequencing using minimal amounts of material, allowing for a single-individual sequencing and high-throughput sequencing of multiple individuals (Lebrigand et al 2020). The availability of complete assemblies necessitates development of specific tools able to identify and annotate tandem and other sequence repeats. A range of existing software exists for repeat annotation. For example, RepeatMasker uses Basic Local Alignment Search Tool (BLAST) and a library of transposable elements (Smit et al. 2014), Tandem Repeats Finder (TRF) uses an algorithm which *de novo* extracts tandem repeat families (Benson 1999), and RepeatExplorer2 uses graph-based clustering to annotate repeats (Novák et al. 2013). Although these tools are effective for *de novo* identification of tandem repeat regions, they do not precisely annotate individual repeat locations, or higher order repeats.

More recently, specific tools have been developed to annotate the human centromeric alpha satellite arrays and their higher order structures, including HORmon (Kunyavskaya et al. 2022), centroFlye (Bzikadze and Pevzner 2020), Alpha-CENTAURI (Sevim et al. 2016), HiCAT (Gao et al. 2022) and CentromereArchitect (Dvorkina et al. 2021). Although these tools are effective in human genomes, they rely on prior mapping of repeats and in some cases monomer definitions, understood as division of individual repeats into highly similar classes that define HOR subunits, which limits their wider applicability. Other software designed to annotate tandem repeats are also available, including PHOBOS that focuses on short repeats (1-50 bp) (Mayer, http://www.rub.de/ecoevo/cm/cm_phobos.htm), and TRAL that is designed to identify internal tandem repeat in proteins (Schaper et al. 2015). In summary, a method for annotation and analysis of megabase tandem arrays, which does not rely on previously identified repeats, is required.

I wrote TRASH: Tandem Repeat Identification and Structural Hierarchy to address these challenges and specifically to facilitate analysis of tandem arrays, including centromeric satellite arrays, without prior knowledge about repeat families present in an assembly. Additionally, I designed the software for ease of use, so that the tool can be widely used by the community.

A further consideration for centromere satellite arrays is that they are often characterised by Higher Order Repeats (HORs), which are multi-repeat length structures superimposed on monomer arrays (Huntington 1987). Initially defined in human chromosome-specific alpha-satellite DNA units, HOR blocks were later described to contain a varying number of repeat subunits (monomer classes), with internal monomer class identity levels higher than average identity between all repeats (A. R. Mitchell 1985, HF Willard 1985, Vladimir Paar 2007). The existence of HORs together with high repeat diversity between species (Melters et al. 2013), while repeat unit lengths are constrained within species, has been suggested to result from recombination pathways including unequal crossover, gene conversion and break-induced repair (Tinline-Purvis et al. 2009, Koumbaris et al. 2011). Description of HORs can be advantageous for analysis of centromere evolution and ancestry (Altemose et al. 2022, Logsdon et al. 2021, et al. Miga 2019). For example, human alpha-satellite HORs were found to be predominant in the central parts of the centromeres, typically associated with higher CENP-A occupancy and lower methylation relative to the rest of the centromere (Altemose et al. 2022).

To expand TRASH, I decided to implement a HOR identification module. In the predominant definition of HORs, derived from the human alpha-satellite studies, blocks of monomer classes form HORs that are arranged in tandem and are unique to a chromosome (Sevim et al. 2016, Dvorkina, Bzikadze and Pevzner 2020). Therefore, their identification can be approached by identification of monomer classes, assigning repeats to these classes, and searching for patterns in the monomer class strings. This method is used by most software handling HOR identification like HORmon or Alpha-CENTAURI (Kunyavskaya et al. 2022, Sevim et al. 2016). This method, despite performing well on human centromeres, assumes that HORs are arranged tandemly and that repeats can be reliably divided into monomer classes. Therefore, I decided to use an alternative method, where monomer identification is omitted, and all repeats are compared against each other in a dot-plot-like fashion. Then, two blocks of repeats that have high similarity of their respective repeats can be defined as a HOR. This way, a repeat can be a part of multiple HORs, potentially highlighting its evolutionary history.

3.2 Results

Tandem Repeat Annotation and Structural Hierarchy (TRASH) software has been developed using R and C (see Methods for details). In this chapter, intermediate steps and results of TRASH workflow are showcased using *Arabidopsis thaliana* Col-CEN assembly (Naish et al. 2021).

3.2.1 TRASH input and parallelisation

TRASH is built to accept a fasta-formatted file, or number of files, and creates a list of all sequences present in these files. Then, in the user-specified directory, it creates output sub-directories where temporary files will be kept and some output files saved, while the main output files and plots are placed in a user-specified directory. The next step of repeat identification is, if possible, performed in parallel for each of the sequences. After all sequences are analysed, the output is formatted and plotted for the user.

3.2.2 Identification of tandemly repeated DNA segments

In any tandemly repeated region, sufficiently short *k*-mers will be repeated approximately every N-bp, where N is the period of the array (Fig. 3.1). This characteristic is used to identify regions that contain tandem arrays and to find said N value. In the first step of TRASH, each sequence is divided into adjacent windows of 1,500 bp by default, which should allow for identification of the majority of the centromeric repeats (Melters et al. 2013), but is not too large to slow down the analysis considerably. This value, like most others described here, can be adjusted by the user if, for example, longer repeats are expected. In each window, the repeat content score is measured by calculating the proportion of non-unique *k*-mers relative to the window size (*k*=10 nucleotides by default). The window size can be modified and is affecting the maximum repeat size that TRASH can identify, as the identical *k*-mers will only be found within the window of sufficient size. Additional filtering for repeat size can be set by the user and is performed at a later stage. Figure 3.2 presents the reasoning behind choosing the optimal *k* value. It can be modified by the user without impacting the runtime (Fig. 3.2A). However, too short *k* values might give a large false-positive rate (when $k < 6$, Fig. 3.2B), while too long *k* values will raise the

stringency until no repeats can be identified unless they are identical. Nonetheless, lower end *k*-values are advised since additional annotation of repetitive windows is not detrimental to the result if later stages of the script are not able to identify repeats over too loosely identified regions. The only cost is additional runtime, which is negligible and such false-positive regions tend to be very short. Repeat content scores tend to approach 0% for windows without repeats, whereas values will be in the range of 80-100% for windows occupied by tandem repeats (Fig. 3.3).

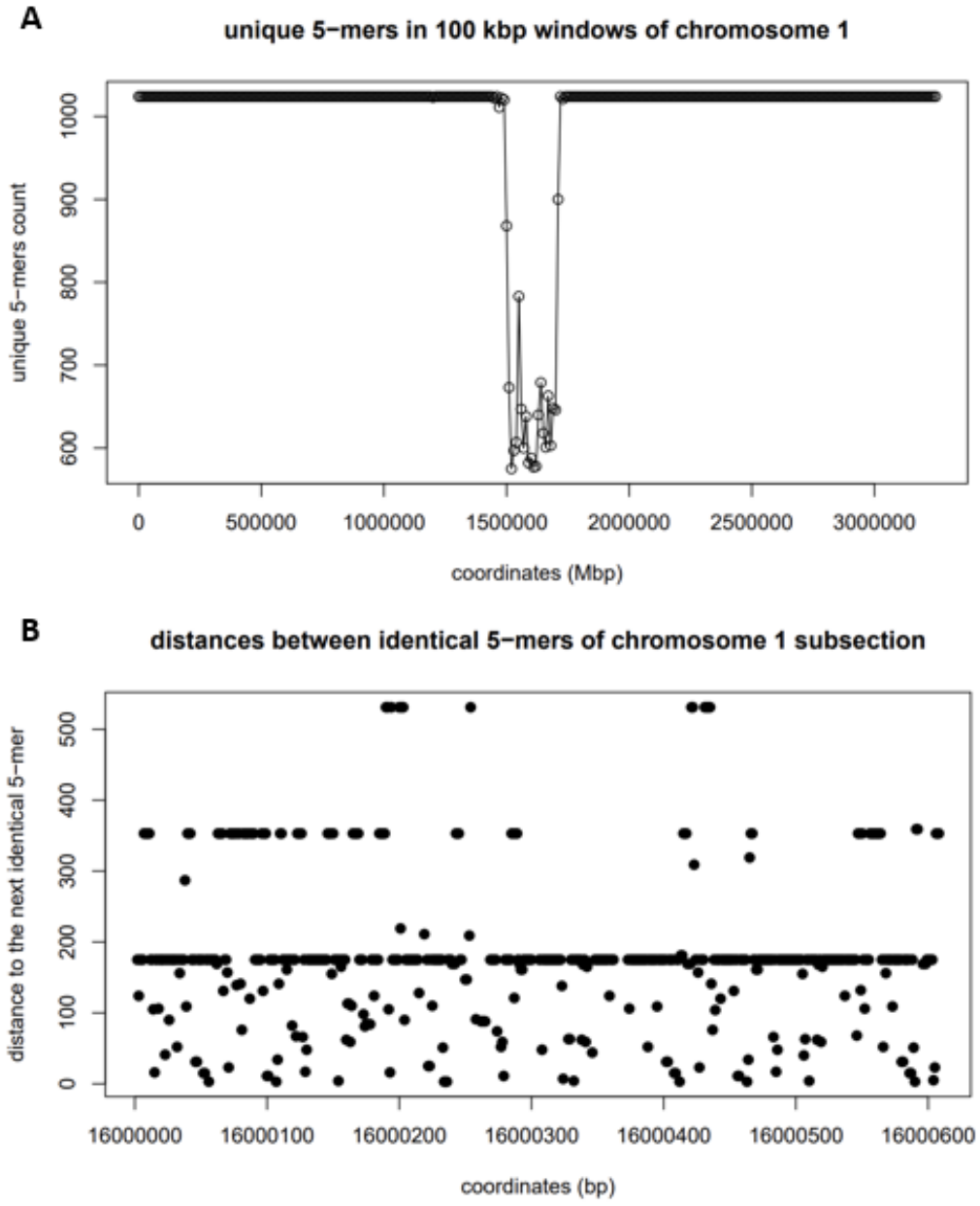


Figure 3.1. *K*-mer counting as a method of identifying repetitive DNA.
A. Unique 10-mer counts in 100 kbp windows along chromosome 1 of the Col-CEN assembly are plotted. *k*-mer counts are strong indicators of underlying sequence

repetitive regions. **B**. Distances between non-unique 10-mers of a sub-region of chromosome 1 (16,000,001:16,001,001 bp), which can indicate the period of underlying tandem repeats. The plotted sub-region contains repeats of a 178 bp period.

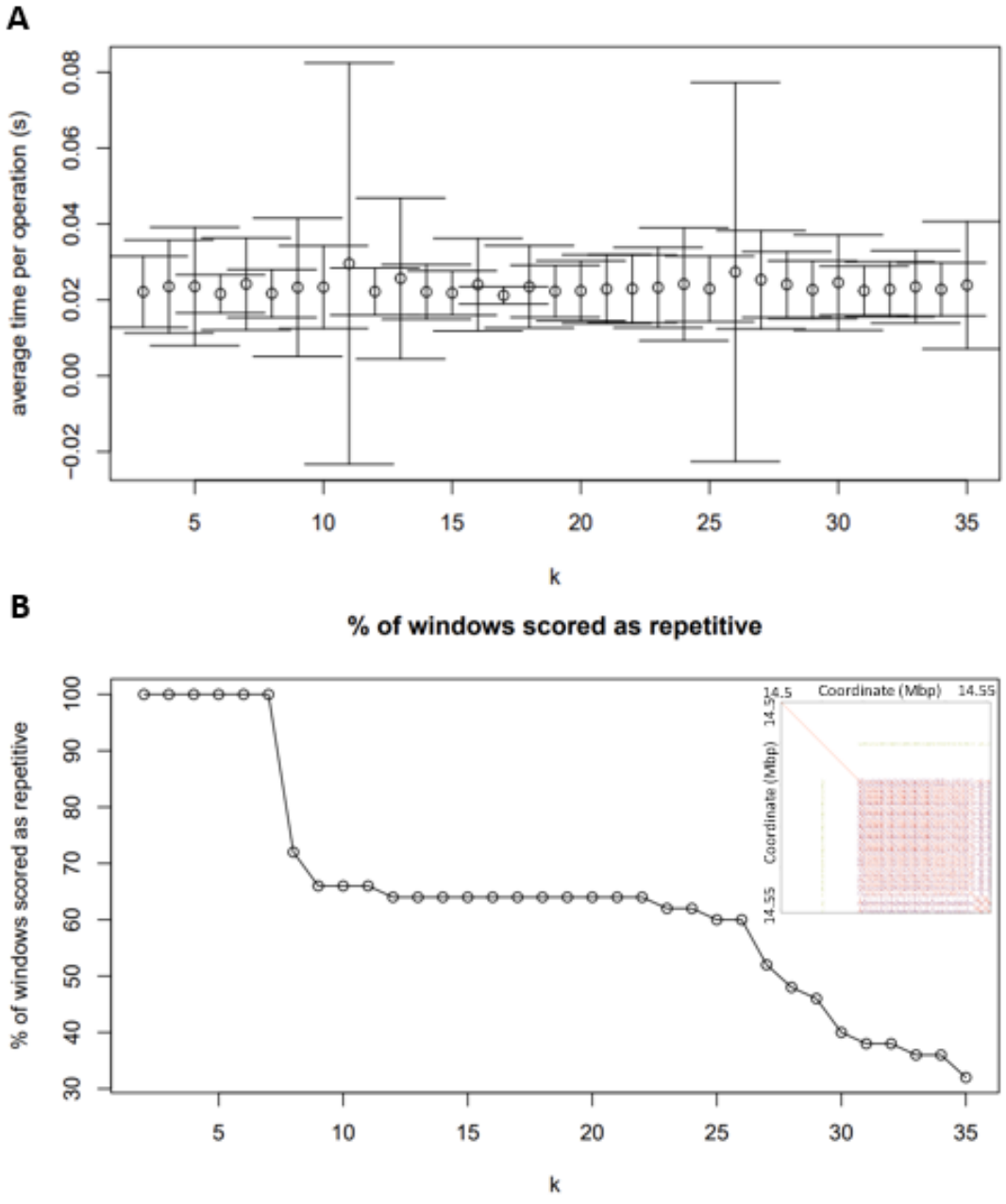


Figure 3.2. Determining the best value of *k* for the analysis.
A. *K* value vs runtime of non-unique *k*-mer distance calculations is plotted, which is the most time-consuming step of TRASH. For each *k* value, 100 random windows were checked and the average runtime with standard deviation is plotted. The *k* value appears to not consistently affect the runtime. **B**. *k* value vs % of identified windows under the threshold for repeat content from a region of chromosome 1 (14,500,000:14,550,000 bp). Approximately 64% of this window consists of *CEN178*

repeats, as shown on the inset dot plot of the same region. This suggests that *k*-mer values in the range of 9-26 would be appropriate for this assembly.

Otsu's (1979) method is used to find the optimal threshold that divides the bimodally distributed window scores into those containing repeats and those that do not. All windows above this threshold are marked as repetitive, as they contain a high number of internally repeated *k*-mers, and those that are physically adjacent are concatenated (Fig. 3.4A). After filtering regions under the allowed region size minimum (3000 bp by default, used to shorten analysis time by removing regions containing small numbers of repeats), the result is a list of repetitive regions, to which subsequent analysis is restricted.

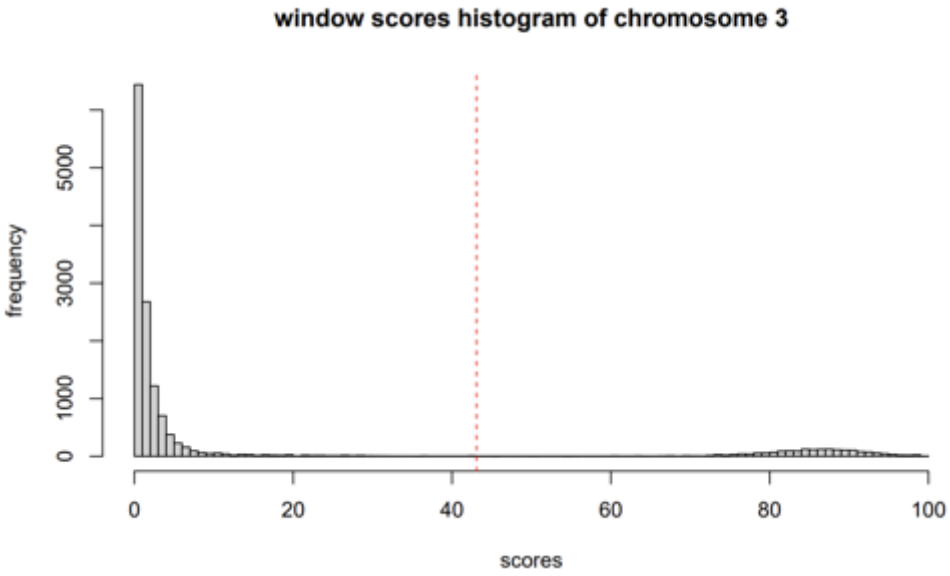


Figure 3.3. Scoring regions of the assembly for their repeat content and dividing into repetitive and non-repetitive components.
Chromosome 3 of the Col-CEN assembly was divided into adjacent 1 kbp windows and each window was scored for the proportion of non-unique 10-mers. The distribution of these scores is plotted with the division between two peaks calculated using Otsu's method (red dotted line).

3.2.3 Identification of the tandem repeat period, mapping corrections and primary consensus generation

Assuming the identified repetitive regions consist of one or more tandemly repeated arrays, it should be possible to determine the period of the repeats by calculating the most common distance between pairs of consecutive identical *k*-mers. (Fig. 3.4B). The search is performed by mapping each *k* length subsequence to a downstream region limited by minimum and maximum repeat size settings (4 and 1000 bp by default respectively). The N value is the most common distance found within the region and is the approximation of the periodicity of underlying DNA repetition.

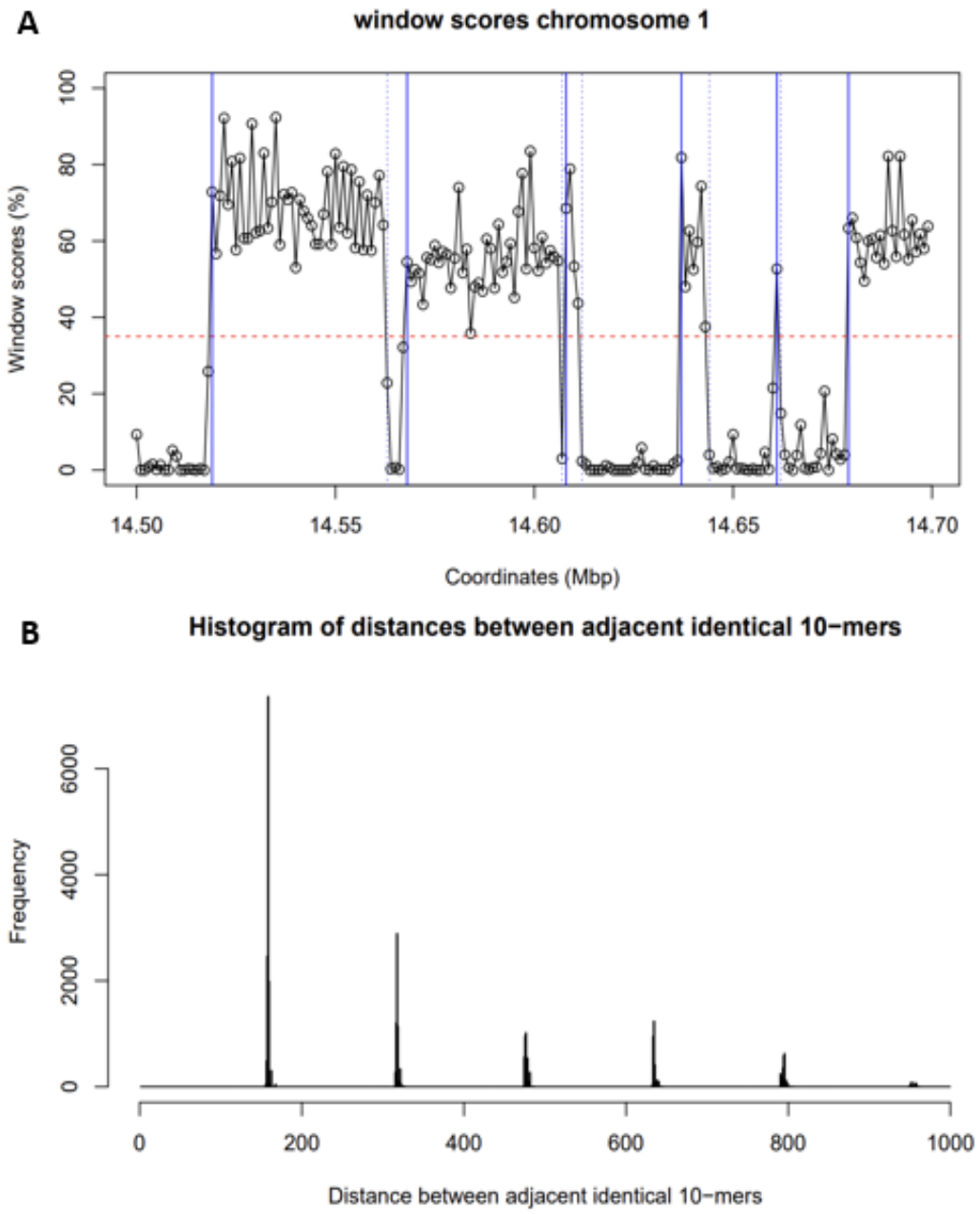


Figure 3.4. Concatenating repetitive windows into repetitive regions and identifying the underlying repeat period.
A. The percentage of non-unique 10-mer scores in 1 kbp adjacent windows are plotted from an extraction of chromosome 1 of the Col-CEN assembly (14,500,000:14,700,000 bp). Repetitive regions are formed from adjacent windows with scores above the threshold (red dashed line) and are marked using blue vertical lines. The solid line marks the beginning, whereas the dashed line marks the end of each window. **B**. Histogram of distances between adjacent identical 10-mers taken from the first repetitive region identified in A (14,519,001:14,564,000 bp). The most frequent distance is 159 bp, followed by multiplications of that value, suggesting the most common repeat in that region has a period of 159 bp.

If a region is occupied by an N-bp tandemly repeated sequence, any random N-sized sub-sequence is likely to be representative of the entire region. It is therefore used to sample the region a number of times (5 by default), to randomly extract N-length subsequences. These are mapped back to the region using the matchPattern function from the R Biostrings package (Pages et al. 2022) (Fig. 3.5A). Each set is refined by looking for overlaps and gaps between consecutive repeats. Shorter overlaps (under 0.65 * N bp by default) are divided equally between the repeats, longer overlaps (equal or larger than 0.65 * N) are handled by removing the shorter repeat. This has an ability to correct for repeat dimers being identified over monomers, since the potential overlap will cause the repeats to split in half. Short gaps (under 10bp) are handled by extending the neighbouring repeats to cover the gap evenly (Fig. 5.6B). The set of matches that covers the greatest part of the region is then extracted and aligned using MAFFT (settings: --kimura 1 --retree 1), consensus of which becomes a primary consensus of the region.

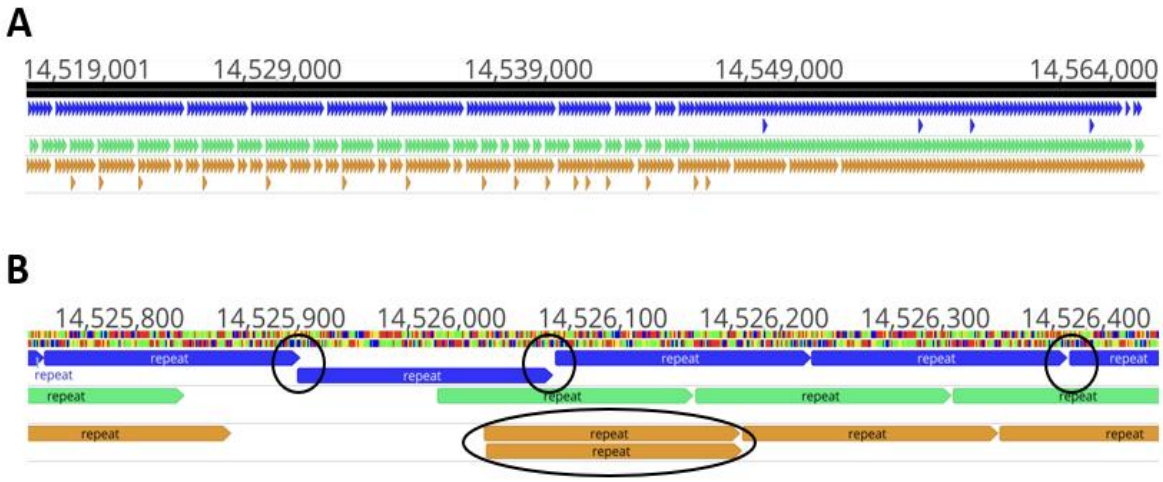


Figure 3.5. Identifying the primary repeat consensus.
A. The results of mapping 3 random substrings of the size 159 bp back to the sequence of the repetitive region shown in Figure 5.4 (14,500,000:14,700,000 bp). A random substring of a tandem array should be representative of the whole sequence. Different coloured tracks represent 3 different mapping attempts using different 159 bp substrings. **B.** Section of the region shown above highlighting mapping imperfections (circled). They are handled by splitting the short overlaps between adjacent repeats, filling short gaps by extending adjacent repeats, or removing one of the annotated repeats with long overlaps.

3.2.4 Splitting adjacent regions with more than one repeat family

In some cases, distinct repeat families are found immediately adjacent to one another (Fig. 3.6A). To identify both families, TRASH checks the coverage of the primary consensus mapping, and if a continuous sequence of more than the allowed region size remains, it splits the region. The newly created region goes through the same N value and primary consensus identification process, until no further repeats can be found, or no more sequence with unmapped repeats remains (Fig. 3.6).

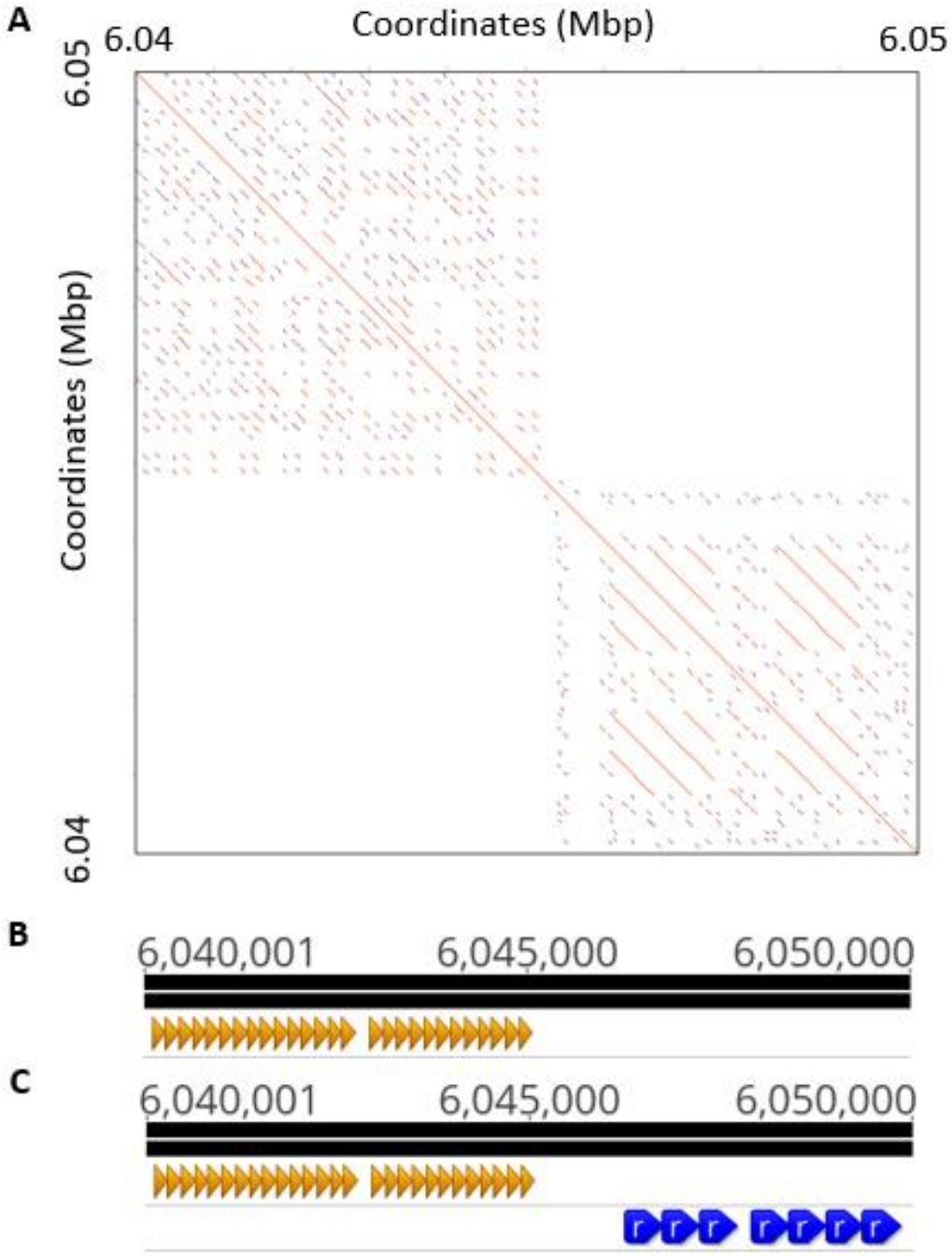


Figure 3.6. Handling two distinct tandem repeat arrays located in proximity.
A. Dot plot of the Chromosome 2 6,040,000:6,050,000 bp from the Col-CEN assembly showing an example of two different tandem repeat arrays located proximally. **B**. Mapped repeats after the initial identification step. **C**. Mapped repeats after the splitting step and second identification. After initial identification TRASH checks whether there is a substantial, uninterrupted sequence remaining with no repeats annotated. If so, TRASH splits the region, or multiple regions, to process them in the same way as before.

3.2.5 Consensus shifting, family templates and secondary consensus generation.

To ensure that repeats of the same family identified in independent regions can be directly compared, their relative start positions should be consistent, and they should be identified in the same orientation (i.e., they should be in the same shift). This problem arises due to tandem repeats not having intrinsic start positions. Applying a correction to map repeats with the same shift means downstream processing is simplified. For example, an alignment between repeats in the same relative shift allows a full comparison with reduced end gaps present in alignments (Fig 3.7). A query repeat can be moved to the same shift of a subject repeat by creating all possible shifts of the former (including reverse complementary), making pairwise alignments with the latter and choosing the shift that creates the highest alignment score, or contains the least free end gaps. Unfortunately, this approach cannot be applied when there are multiple sequences to adjust and they are likely not to be of the same repeat family, as is the case with a list of primary consensus sequences identified in the initial stages of TRASH. Instead, I decided to employ a method of changing the shift independently for each primary consensus, in a way that related repeats would end up in the same shift without pairwise comparisons. In short, the algorithm is hashing repeat *k*-mers and for each possible shift, these hash values (K) are multiplied by the *k*-mer position in that specific shift resulting in a shift score (S). In the first step, 6-mers are extracted starting at each position, with final *k*-mers taking nucleotides from the beginning of the sequence. Then, each *k*-mer is assigned a sequence-based score $\mathrm{K_n}$ by dividing it into nucleotides at all positions *i* for which a value V_i is assigned: 'A' = 0, 'C' = 1, 'T' = 2, 'G' = 3. Non-standard nucleotides are removed from the analysis at an earlier stage. $\mathrm{K_n}$ is then calculated using equation:

$$K_n = \sum_{i=1}^{6} V_i i^4.$$

This results in a list of $\mathrm{K_n}$ values which are then transformed *n - 1* times, each time moving its start position by 1 and shifting the last value to the beginning. This corresponds to *n* possible shifts of the primary consensus sequence (including the

original one). To consider the reverse complement orientation, each of the calculated lists is reversed. $m=2n$ lists have their unique S_m score calculated with the equation:

$$S_m = \sum_{n=1}^{n} n * K_n.$$

The shift with the lowest score S_m is then replaced as the primary consensus. This secures similar shifts for related repeats. An example of alignment of a group of *5S* rDNA repeats with and without shifting shows that the algorithm can adjust multiple consensus sequences, bringing the number of free end gaps to the minimum (Fig. 3.7).

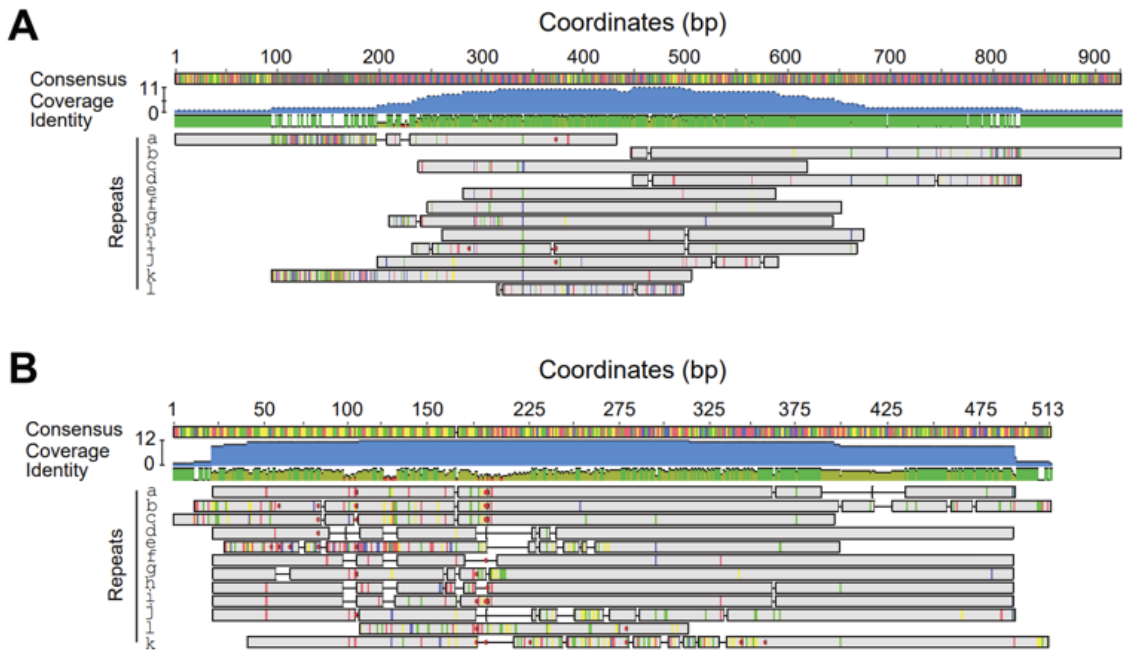


Figure 3.7. Shifting the frame of related repeats using TRASH.
A. A multiple pairwise sequence alignment of 12 *5S* rDNA repeats is shown, prior to shifting the repeat consensus. Variable start positions cause uneven alignment coverage, making sequence comparisons less accurate and causing 'end gaps' when coverage against the consensus is plotted. **B**. A multiple sequence alignment of the same *5S* rDNA repeats is shown, following shifting the repeat start positions via TRASH. This causes repeats to have maximised alignment coverage, which allows for accurate comparison of the repeats. The figure is adjusted from the TRASH submission manuscript.

The downstream TRASH module of HOR identification requires that only one family of repeats is used, to ensure the multiple pairwise alignment is optimal (i.e., least gaps coming from aligning unrelated sequences). Classification can also be helpful to assess the size of a repeat family present in an assembly. For that, a table can be provided by the user with information on sequences and names of putative repeat families found in the assembly (called a 'template') (Table 3.1). Templates may be previously known repeats, or abundant repeat families identified in previous TRASH analysis. The first step is to identify which primary consensus sequences are related to the template. Since the shift between these sequences might be different, direct alignment to calculate the similarity is not feasible. Instead, a *k*-mer approach is used where both query (primary consensus) and subject (template sequence) are divided into k-mers, similarly to the shifting method. The two *k*-mer sets are compared for identical elements (with duplications) and the Jaccard similarity index is returned (Fig. 3.8). The same is performed between the query and 1,000 random permutations of the subject. The first score is compared to the distribution of permutation scores to test for a significance with a 95% confidence interval. In the case of multiple sequence templates being provided, the template with the highest score is chosen and the name of the template is assigned to the primary consensus as its 'family'. Figure 3.8 shows the similarity levels between related (A) and unrelated (B) sequences with varying *k*-values. Based on this analysis, 6 was determined to be the default value of *k*.

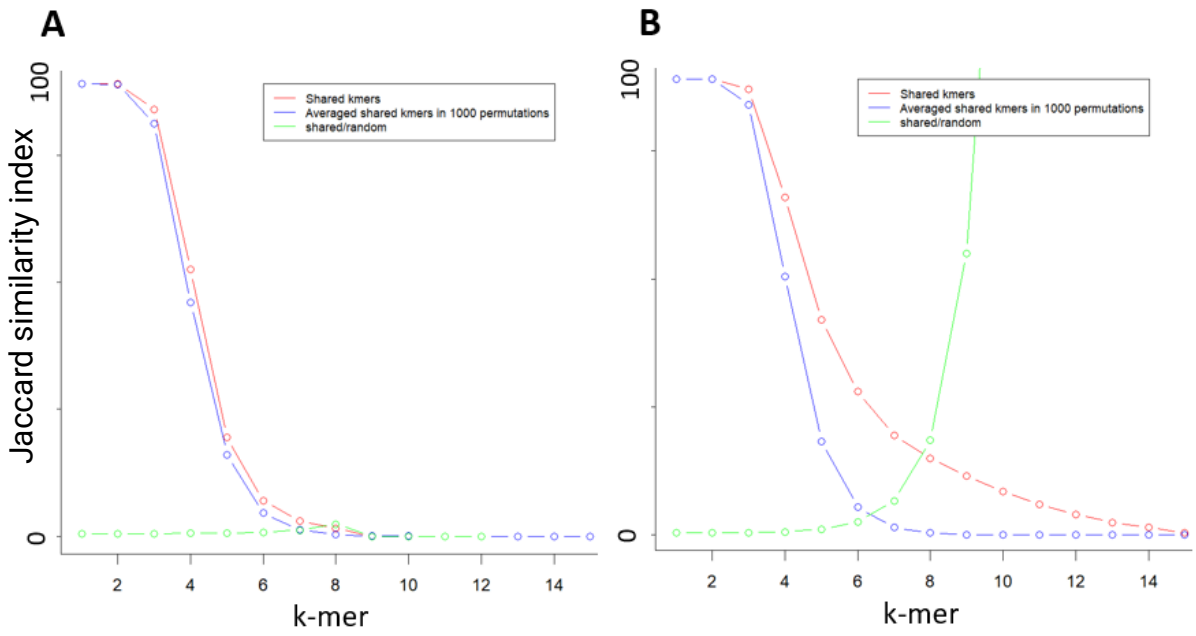


Figure 3.8. Jaccard similarity index as a function of k-mer value.
A. In red, comparison between two sequences with no expected similarity (subject: *CEN178* consensus and query: its reverse complement). In blue, the average score of 1,000 permutations of the query against the subject. The green line is a ratio of the base similarity and permutation-based scores. **B.** Equivalent comparison between *A. thaliana CEN178* and *A. lyrata CEN179* sequences with 80.8% percentage identity (significant at $p = 0.01$ in a 500 permutations test).

When a primary consensus is assigned to a repeat template, it also has its shift changed to best align with the sequence template. This is done with the previously mentioned method of creating all possible shifts, aligning to the sequence template, and extracting the one with the highest score. Table 3.1 is an example table that was used in repeat identification in the *Arabidopsis thaliana* Col-CEN assembly (Naish et al. 2021) presented later in this chapter.

Sequence	Repeat family name	Length (bp)
AGTATAAGAACTTAAACCGCAACCGATCTTAAAAGCCTAAGTA GTGTTTCCTTGTTAGAAGACACAAAGCCAAAGACTCATATGG ACTTTGGCTACACCATGAAAGCTTTGAGAAGCAAGAAGAAGG TTGGTTAGTGTTTTGGAGTCGAATATGACTTGATGTCATGTGT ATGATTG	*CEN178*	178
TTGGGAGAAAATGGGTATAAGTGTTGTCTAAACACTCCTAATC CATCTCTAACTCTTATAATTAGTCAAATGCATTGGATTGTGAC ACATTTTGACCATAGAAACACTAACAAAGCTATTTACTGCTTC TAAGCAATTTTTTGTTGGTTTTAGCCTCTT	*CEN159*	159
TCGGAGGGCTGTCTTTGGGCTTTCCGAAAAGGTATCACATGC CAAGTTTGGCCTCACGGTCTAAAAGTTATGGAGTCATAAAGT TTTAACCAAAAAAAAAAAAGGTTAAACATAAAAGAGGGATGCAA CACGAGGACTTCCCGGGAGGTCACCCATCCTAGTACTACTCT CGCCCAAGCACGCTTGACTGCGGAGTTCTGATGGGATCCGG TGCATTAGTGCTGGTATGATCGCATCCGTTAGTATATGCAATG CAATCGTATATATTCTTTTTTGAAGACTTGATGAACCATTCGC CGTGGGTCCCACCCGCTATGTAGGGATACCCCATCTAGTCTT AACGAGCTTTGATGCATGAAAAAATTCGAAAACAATGCTTGAA CAAGTAATTTTGGGTCCGTAATATAGCCCAAATCACGAAAATG CCCGAAAAAGTACTTAAAGGTCAAAATTTGGGGTCGACAAAA AGTCAATGGAAAAGTTCCATTGTCCTGCTTCTTTCG	*5S*	502

Table 3.1. Sequence templates used for analysis of the Arabidopsis thaliana Col-CEN assembly by TRASH.
These are used to inform TRASH repeat classification module after *de novo* identification of repeats.

Modified primary consensus sequences resulting from the shift adjustment and family classification are processed as before, by mapping them to their regions, polishing the mapped repeats and extracting the consensus of the mapped repeat alignments. All repeats of the same family (when applicable) from the whole chromosome/sequence are also extracted to be aligned to create a family-wide consensus. Each repeat is then compared to that consensus to calculate the Levenshtein edit distance to facilitate downstream analysis.

3.2.6 Tandem repeat identification output

TRASH analysis of Col-CEN using default settings resulted in the identification of 96,793 repeats having a total length of 13,257,524 bp, which corresponds to 10.03%

of the 132,081,078 bp genome. 64,791 repeats were classified as *CEN178*, 1,479 as *CEN159* and 1,262 as *5S* rDNA, based on the provided sequence templates (Table 3.2). A histogram of repeat sizes confirms the high relative abundance of *CEN178* centromeric repeats (Fig. 3.9B).

The main output of TRASH is a comma-separated values (csv) file with the details of the identified repeats. Table 3.2 presents an example from the chromosome 1 of *Arabidopsis thaliana* Col-CEN assembly. Family is assigned to a repeat when the region's consensus matches one of the provided templates. In this case, edit distance is also assigned and repetitiveness is a score that can be calculated at the later stages of the run following the HOR module. Visualisation of the identified repeats is performed with a circos plot, where the most common repeats (grouped by the length) are plotted for each of the fasta file sequences (Fig. 3.9). The circos plotter was implemented by TRASH co-author, Michael Hong. Additionally, linear plots with all repeats from individual sequences can be produced with start positions on the x axis and repeat sizes on the y axis (Fig. 3.10).

start	end	width	seq	strand	class	seq.name	edit.distance	repetitiveness
4	9	6	TAGGTT	-		Chr1	NA	NA
11	16	6	TAGGTT	-		Chr1	NA	NA
18	23	6	TAGGTT	-		Chr1	NA	NA
25	30	6	TAGGTT	-		Chr1	NA	NA
32	37	6	TAGGTT	-		Chr1	NA	NA
38	44	7	TAGGGTT	-		Chr1	NA	NA
45	51	7	TAGGGTT	-		Chr1	NA	NA
53	59	7	TAGGGTT	-		Chr1	NA	NA
60	66	7	TAGGGTT	-		Chr1	NA	NA
67	73	7	TAGGGTT	-		Chr1	NA	NA
74	80	7	TAGGGTT	-		Chr1	NA	NA
81	87	7	TAGGGTT	-		Chr1	NA	NA
88	94	7	TAGGGTT	-		Chr1	NA	NA
95	101	7	TAGGGTT	-		Chr1	NA	NA
102	108	7	TAGGGTT	-		Chr1	NA	NA
110	115	6	TAGGTT	-		Chr1	NA	NA
116	121	6	TAGGTT	-		Chr1	NA	NA
122	128	7	TAGGGTT	-		Chr1	NA	NA
129	135	7	TAGGGTT	-		Chr1	NA	NA
136	142	7	TAGGGTT	-		Chr1	NA	NA

Table 3.2. Example 'repeats' output generated by TRASH on the Col-CEN assembly. TRASH was run with the HOR module activated and using table 5.1 as a template input.

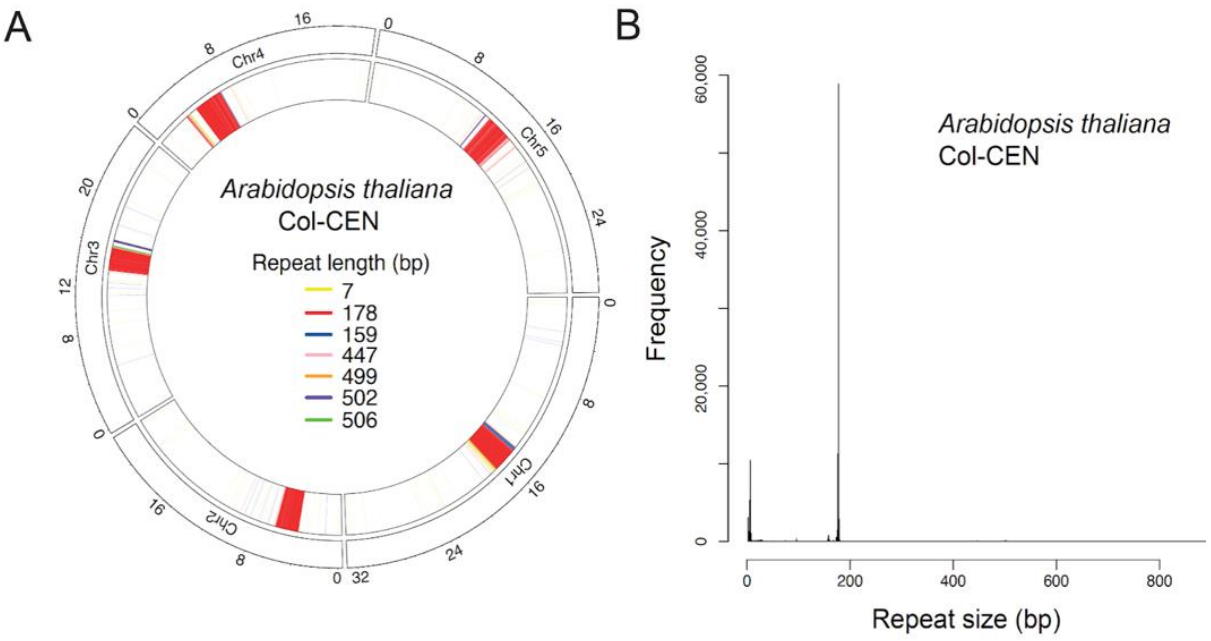


Figure 3.9. Tandem repeat identification in the *Arabidopsis thaliana* Col-CEN genome by TRASH.

A. A circos plot of tandem repeats identified by TRASH in the *A. thaliana* Col-CEN genome assembly (Naish et al., 2021). This is the default output of the TRASH repeat identification module. Repeat shading is coloured according to repeat length (bp). **B**. Histogram of tandem repeat lengths (bp) identified in Col-CEN. Visible peaks correspond to telomeric (7 bp), *CEN159* (159 bp), *CEN178* (178 bp) and *5S* rDNA (~502 bp) repeat families. This figure is adjusted from the TRASH submission manuscript.

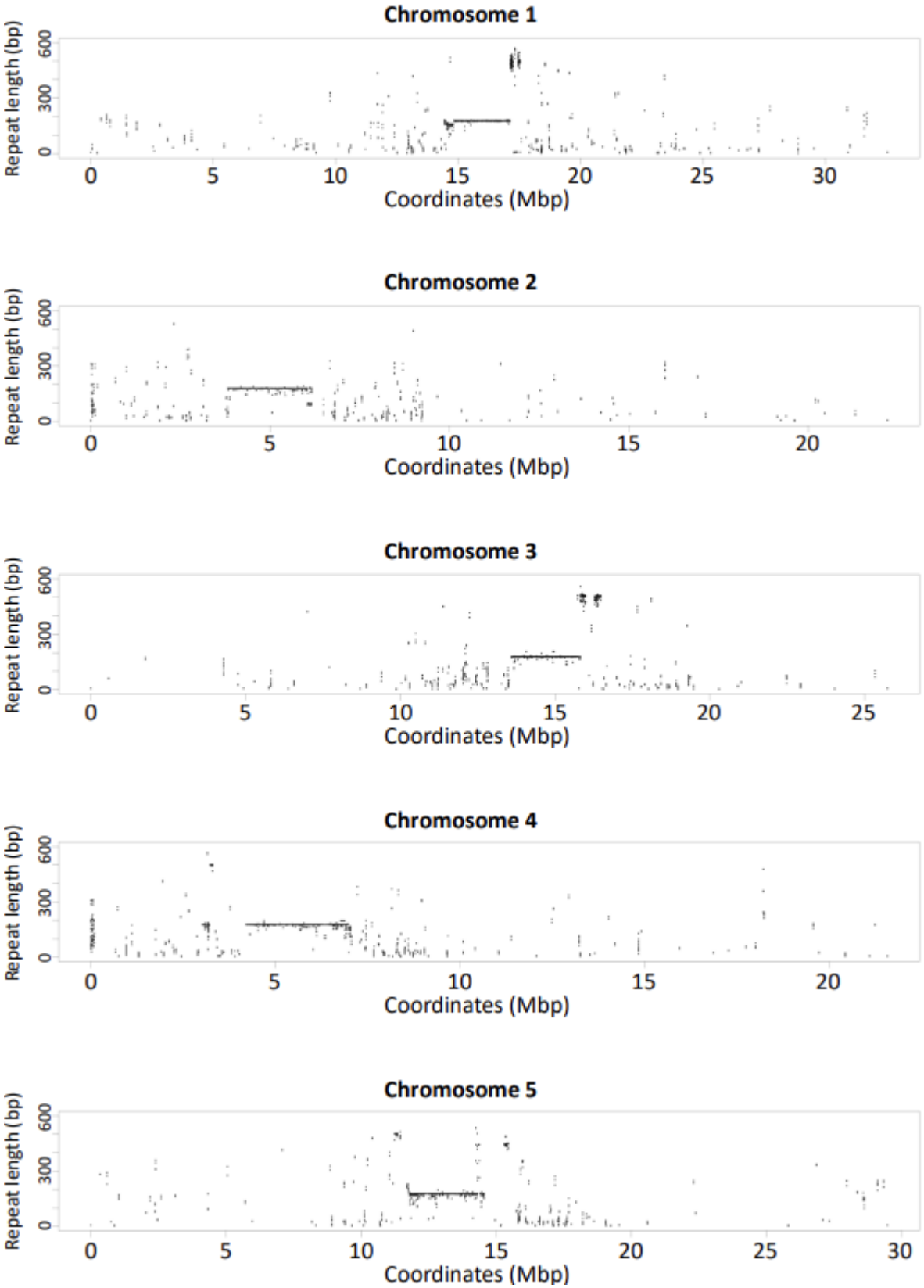


Figure 3.10. Linear plots of tandem repeats identified in the Col-CEN assembly using TRASH.

Plots of tandem repeats are shown along each chromosome. In these plots the y axis range is equal to the maximum TRASH repeat size setting used. This figure is adjusted from the TRASH submission manuscript.

3.2.7 Higher Order Repeat (HOR) identification

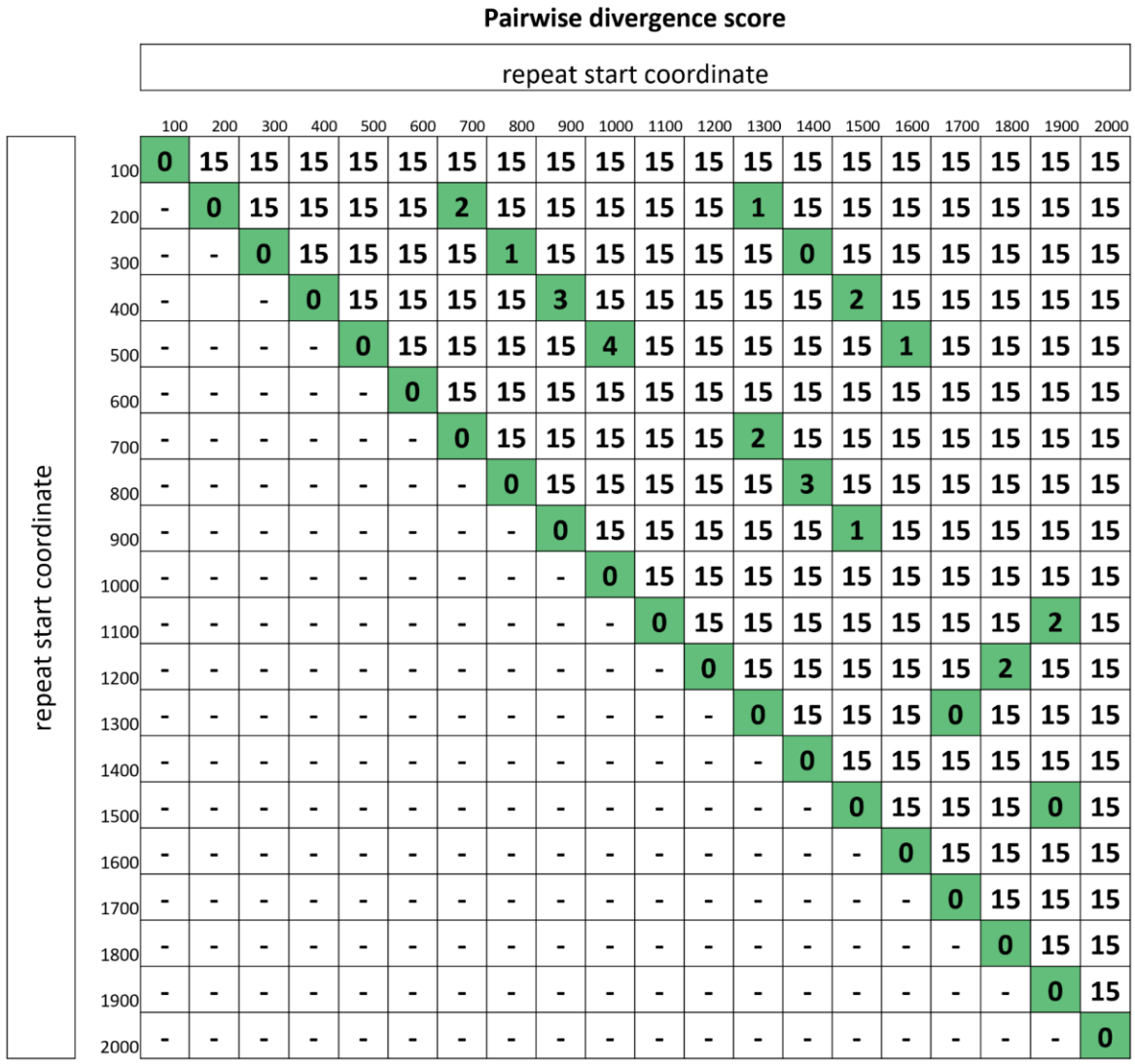
Higher Order Repeats (HORs) are a way of describing the organisation of repeats in a tandem array at different scales. As mentioned before, I use a broad understanding of the HOR term, meaning any duplication of pairs of similar repeats. First, all repeats of the same family are aligned using MAFFT (settings: --kimura 1 --retree 2) (Katoh and Standley, 2013). This alignment is used as input for the HOR identification module. There, each position of each pair of repeats is compared site-by-site searching for variants. The total number of variants identified defines a variant score (*VS*), and a pair of repeats can be a part of a HOR if the *VS* value is lower than the set threshold (5 by default). After an initial match between a pair of repeats with *VS* < threshold is found, a HOR instance is created. That HOR can be extended by testing the adjacent repeats downstream of both initial members, with the same conditions as before. This continues until a pair of repeats do not meet the conditions, or the end of the repeat array is reached for one of the blocks. Information on the direction of each repeat, which is previously identified based on the sequence template, is used to determine in which direction the HOR should be extended. Since *VS* can be low on average, which creates a high number of short HORs, filtering by a minimum number of repeats comprising a HOR can be set (3 by default), to make further analysis less computationally demanding.

Since HOR identification is performed using a 2-dimensional matrix, the number of HORs can increase exponentially with addition of new repeats. This makes it difficult to directly compare the output between chromosomes or between genomes with a varying number of repeats. Additionally, when one repeat can be a part of many HORs, a single score that could summarise its involvement is preferred. Because of that, two metrics are used for normalisation of the identified HOR counts informing on how many repeats are involved in HOR structures:

- Repetitiveness (*Rs*): per repeat sum of all lengths (in monomers) of HORs that repeat is a part of, normalised against the number of repeats of the same family within the analysed chromosome/sequence:

- HOR abundance (*Ha*): per region (centromere/genome) sum of all lengths (in monomers) of HORs present within the region, normalised against the theoretical maximum of this score if all repeats were identical and forming HORs with all other repeats.

R_n is the number of repeats considered for the HOR calculations. H_L is the length in monomers of a HOR. An example theoretical analysis with VS scores matrix visualisation and HOR abundance calculations is presented on Figure 3.11.



orientation	total_variant	start.A.bp	start.B.bp	end.A.bp	end.B.bp	block.units	block.A.size. bp	block.B.size. bp
1	9	200	700	599	1099	4	499	499
1	6	700	1300	999	1599	3	399	399
1	4	200	1300	599	1699	4	499	499
2	4	1100	1700	1399	1999	3	399	399

Figure 3.11. Identification of higher order repeats.
A theoretical region of 20 repeats is shown, with each repeat being 100 bp in length. Pairwise divergence scores between each repeat are shown by the numerical values. When a divergence threshold of 5 is used, green shading is used to show repeat pairs that could potentially comprise a higher order repeat (HOR). Immediately below this matrix, repetitiveness scores for each repeat are shown, which is a sum of all HORs that a repeat is a part of divided by repeat number of the region (20 in this case). The table beneath the figure shows a summary of what TRASH would output for the identified HORs. In the first column, '1' represents 'head-to-tail' orientation and '2'

represents ‘head-to-head’ orientation. ‘Total.variant’ shows the sum of all *VS* scores of the repeat pairs forming the HOR. This figure is adjusted from the TRASH submission manuscript.

An additional output of the HOR identification is a dot plot of start coordinates of all HOR blocks (Fig. 3.12). Repetitiveness and edit distance values can be directly compared and plotted to inform on the characteristics of the tandem arrays (Fig. 3.12). To test the HOR identification module, the dot plot output was compared to a sequence identity heat map of the same region produced by StainedGlass (Vollger, 2022). Consistently, regions with higher pairwise similarity identified by StainedGlass correspond to regions that TRASH annotated with *CEN178* HORs (Fig. 3.12).

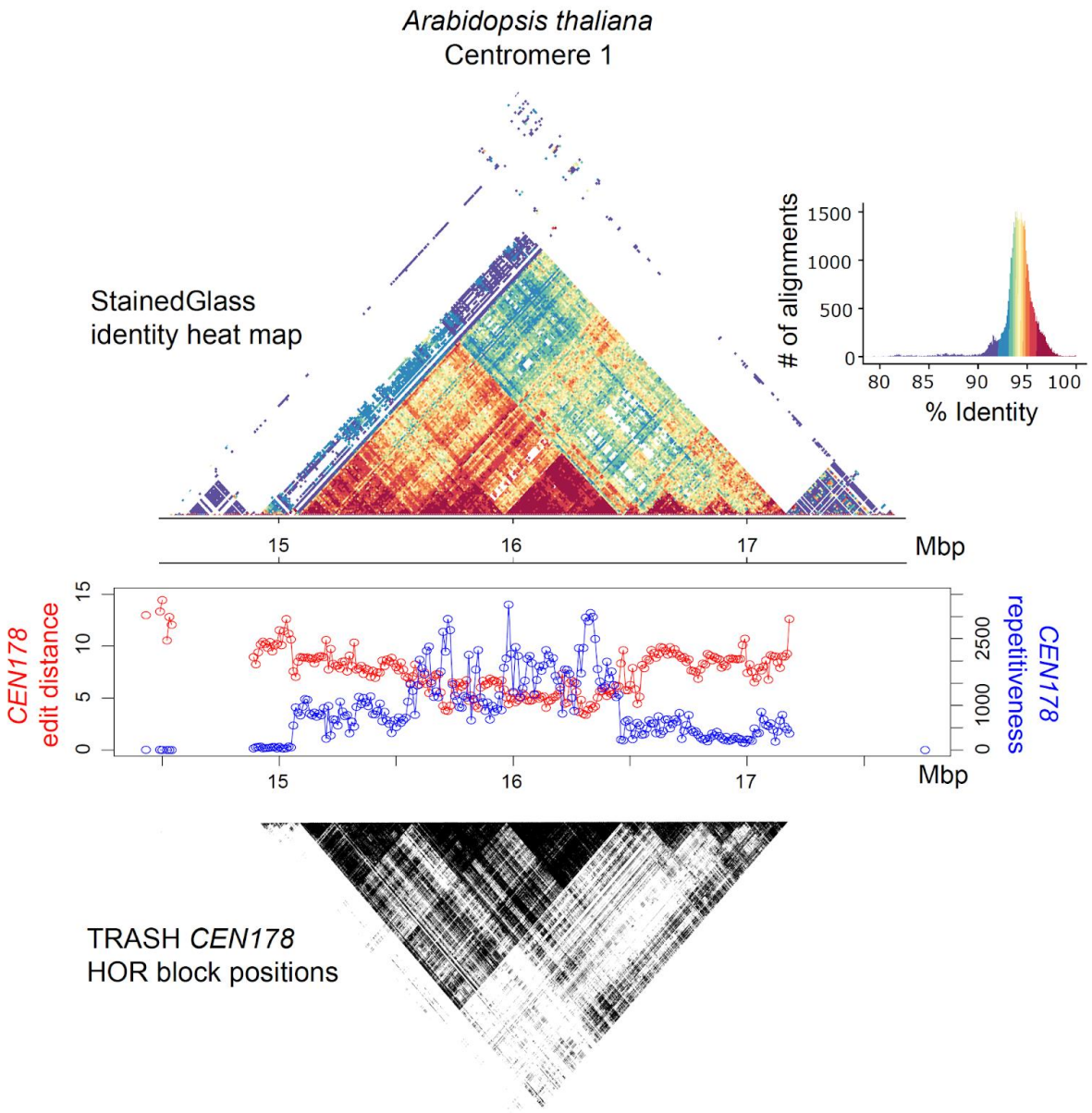


Figure 3.12. Higher order repeat analysis using TRASH.
A StainedGlass sequence identity heat map of *Arabidopsis thaliana* centromere 1 (14,442,038-17,870,129 bp) from the Col-CEN assembly is shown (Vollger et al., 2022). In the centre, characteristics of repeats of the *CEN178* family calculated by TRASH are plotted using a moving average in windows of 20 kbp. Levenshtein edit distance between each repeat and the family-wide consensus is plotted (red), in addition to HOR repetitiveness (blue). Beneath is a dot plot showing *CEN178* HORs identified by TRASH over the same region. This figure is adjusted from the TRASH submission manuscript.

3.2.8 Benchmarking using *Arabidopsis thaliana* Col-CEN assembly

To assess the quality of repeat annotation and to compare it with alternative software, the results of the *Arabidopsis thaliana* Col-CEN genome run were compared to TRF (Tandem Repeat Finder) and RepeatModeller (REF) as alternative *de novo* methods. Additionally, HMMER mapping of a *CEN178* consensus (Maheshwari et al., 2017) was performed as an independent alignment-based method to assess all the benchmarked software *de novo* ability to identify the main family of repeats.

All methods were able to correctly identify the majority of the *CEN178* repeats using the HMMER search as a baseline (Fig. 3.13). TRASH identified repeats that overlapped with 98.29% of the total coverage identified by HMMER, compared with 98.32% for TRF and 99.97% for RepeatModeller (Table 3.3). The advantage of TRASH is the precise mapping of each repeat. All *CEN178* repeats that were identified by TRASH were around 178 bp in size. In comparison, TRF output included overlapping annotations of 178 bp period and 356 bp period, due to improper merging of repeats into dimers. Individual repeats were also not mapped but annotated as entire tandemly repeated regions. RepeatModeller, while having almost perfect coverage of HMMER identified repeats, falls behind regarding its description of the repeats. The regions annotated as “rnd-1_family-1” were the predominant annotation type overlapping with *CEN178* repeats, but its period was described as 1,070 bp, indicating a merge of 6 monomers. Accurate individual repeat mapping of TRASH is crucial in any downstream analysis and is the main advantage of TRASH in comparison with these alternative methods.

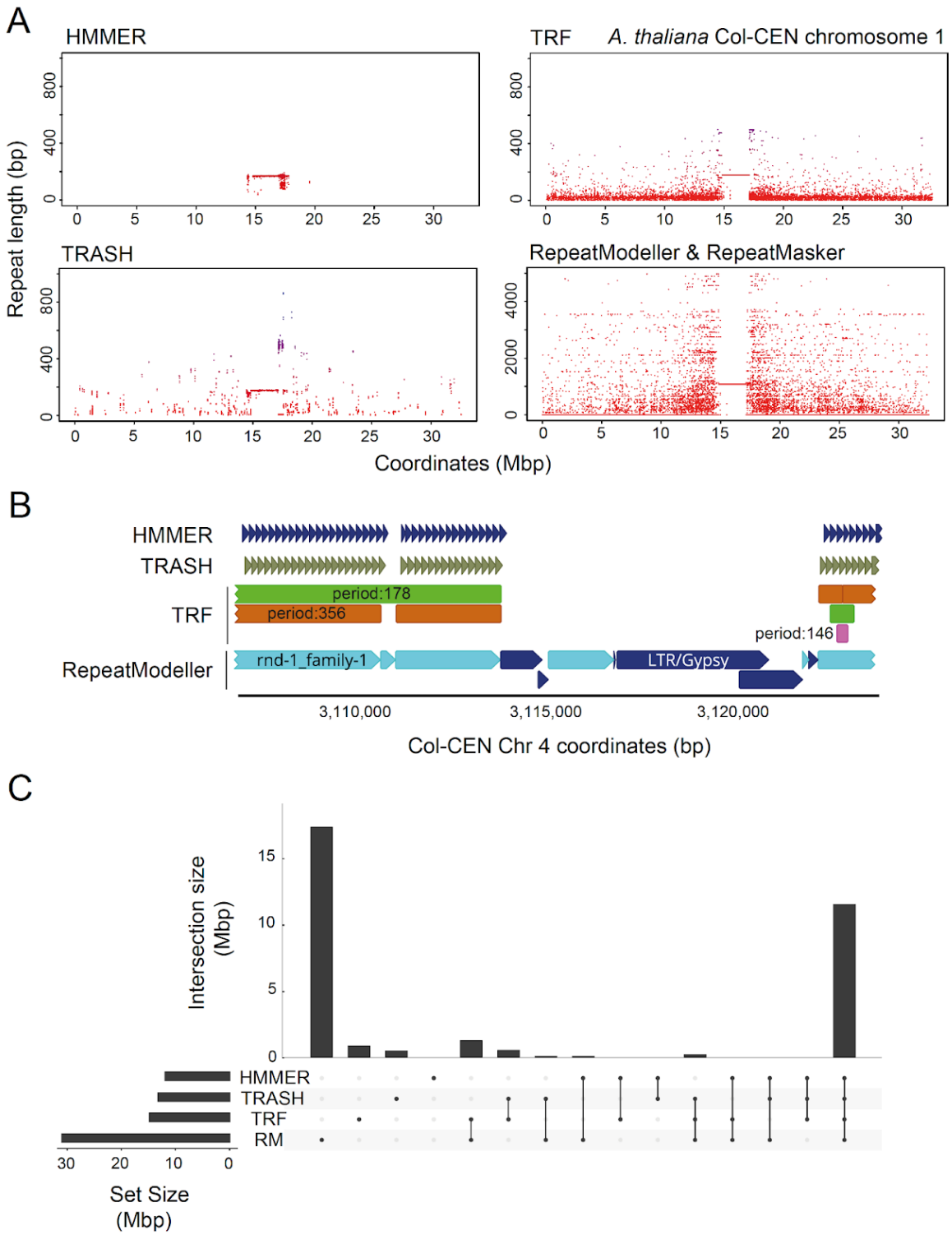


Figure 3.13. Benchmarking TRASH against alternative software for de novo tandem repeat identification.

A. Tandem repeats (red) plotted along chromosome 1 of the Col-CEN assembly identified by TRASH, Tandem Repeat Finder (TRF) and RepeatModeller with RepeatMasker, or by searching a *CEN178* satellite consensus using HMMER. The y axis represents repeat unit length (bp). **B**. An example region of *Arabidopsis thaliana*

chromosome 4 (3,106,910-3,123,586 bp) showing repeat annotations generated by TRASH, RepeatModeller, TRF and HMMER. The TRF annotation includes information on the periodicity of the annotated regions. 'rnd-1_family-1' corresponds to a repeat family of size 1,054 bp from the RepeatModeller library output. **C**. Upset plot showing overlaps between base-pair coverage of Arabidopsis thaliana repeats de novo identified with TRASH, TRF and RepeatModeller, and *CEN178* consensus-based alignment with HMMER. The Set Size plot shows the total base pairs of repeats (Mbp) identified by each method. The Intersection Size plot shows the total base pairs (Mbp) of repeats that are uniquely annotated by the software, or combinations of software. The software being considered are indicated by the black dots below. The figure is adjusted from the TRASH submission manuscript.

The lower relative repeat coverage of both TRASH and TRF can be attributed to inability of both software to identify interspersed repeats, which are present in the Col-CEN assembly (Naish et al. 2021). The ability of RepeatModeller to identify repetitive elements that are not tandemly arranged results in significantly higher identification coverage compared to TRF and TRASH and is mostly attributed to dispersed transposable elements.

The overlap between base-pair coverage of the identified repeats by each method is presented on an upset plot (Fig. 3.13C). It is a bar-plot visualisation of a Venn diagram, where all possible combinations of sets are presented on the x axis (with their description below) and their sizes on the y axis. The overlap of all sets is 11.5 Mbp, which represents the main *CEN178* array, since HMMER mapping is a part of it. A notable set of positions identified solely by RepeatModeller represent transposable elements and other non-tandemly arranged repeats. Overall, TRASH can identify satellite regions and accurately map individual repeats, providing this information in a tabulated form to facilitate downstream analysis.

Method	Repeat positions identified (bp)	*CEN178* overlap with 'HMMER' consensus search (bp)	Runtime (hh:mm)
'TRASH' (*de novo*)	13,227,720	11,214,232	02:15
'TRASH' (templates provided)	13,226,439	11,214,127	03:49
'TRF'	14,799,833	11,125,819	00:05
'RepeatModeller' and 'RepeatMasker'	31,003,608	11,850,842	14:02

Table 3.3. Annotation of the Arabidopsis thaliana Col-CEN genome assembly by TRASH and alternative software.
The calculated runtimes used a 16 GB RAM 8 core 3.2 GHz machine. The TRASH run with templates included HOR identification.

Runtime differences align with the purpose of each software (Table 3.3). TRF is a rapid method of tandem repeat identification but doesn't provide detailed information outside of the position and repeat period. RepeatModeller and RepeatMasker together have long runtimes which allow for good coverage over both interspersed and tandem repeats, additionally assigning them to the families, as long as they match the available database information. The inability to *de novo* identifies repeats and properly assign a period size for tandem repeats makes it less applicable to analysis of centromeric satellites. TRASH runtime falls between these two and while not as fast as TRF, together with its ease of use, it can be used for studies involving a large number of assemblies.

3.2.9 TRASH workflow settings

The main workflow of TRASH is summarised in Figure 3.14. It follows repeat identification and HOR identification when sequence templates are provided and at least one of the repeats has a family that has been specified to be used as an input for the HOR module. TRASH can be also run using alternative workflows to the main one described above. Especially, when repeat identification has been performed and additional HOR analysis is required. Command line arguments controlling the workflow and changing the settings of the analysis are specified in Table 3.4.

flag	setting
--def	use the default R packages path.
--rmtemp	remove the *_out directory after run completion.
--horclass name	set the name of the repeat family that should be used for HOR calculations, required for the HOR module to be activated.
--limrepno x	limit alignment sizes (in bp of total sequence) used during the run to calculate consensus, samples repeats to avoid large alignment operations. 78000 by default
--horonly x	skip the repeat identification if performed earlier and only calculate HORs, needs to be used togehter with -horclass flag.
--minhor x	HORs shorter than this value will be discarded, 3 by default.
--maxdiv x	pair of repeats with divergence score higher than this value will not be considered as a potential HOR, 5 by default.
--maxchr x	total number of sequences that should be analysed. Usefull when assembly contains large number of contigs. Sequences are chosen based on their size.
--k x	kmer size, 10 by default. Decrease if more degraded arrays should be identified, increase for extra stringency (range of 8-16 recommended).
--t x	threshold score to choose repetitive windows, 5 by default. Change will work similar to the kmer size changes.
--win x	window size to use for initial count of repeat content, 1000 by default. Identified repeats will not be bigger than this value.
--m x	max repeat size to be identified, hard capped by -win setting.
--freg x	regions smaller than this will be filtered out at initial steps (some might remain if they come from splitting of a larger region).
--frep x	repeats shorter than this will be filtered out, 4 by default.
--o path	output path where repeats will be saved and temporary directories created.
--seqt path	path to the file with repeat family templates, the file needs to be formatted as described below.
--par x	max number of cores used for multithreading, defaults to 1. If set as 0, TRASH will try to register as many cores as there are sequences, or maximum available, whatever is smaller.
--randomseed x	set a random seed for reproducibility of the repeat identification, seed from the previous run can be found in TRASH_YYYYMMDDHHMMSS.out
--simpleplot	output a plot with repeat coordinates and their sizes for each sequence (additionally to the circos plot)

Table 3.4. User available flags controlling the workflow and settings of TRASH.

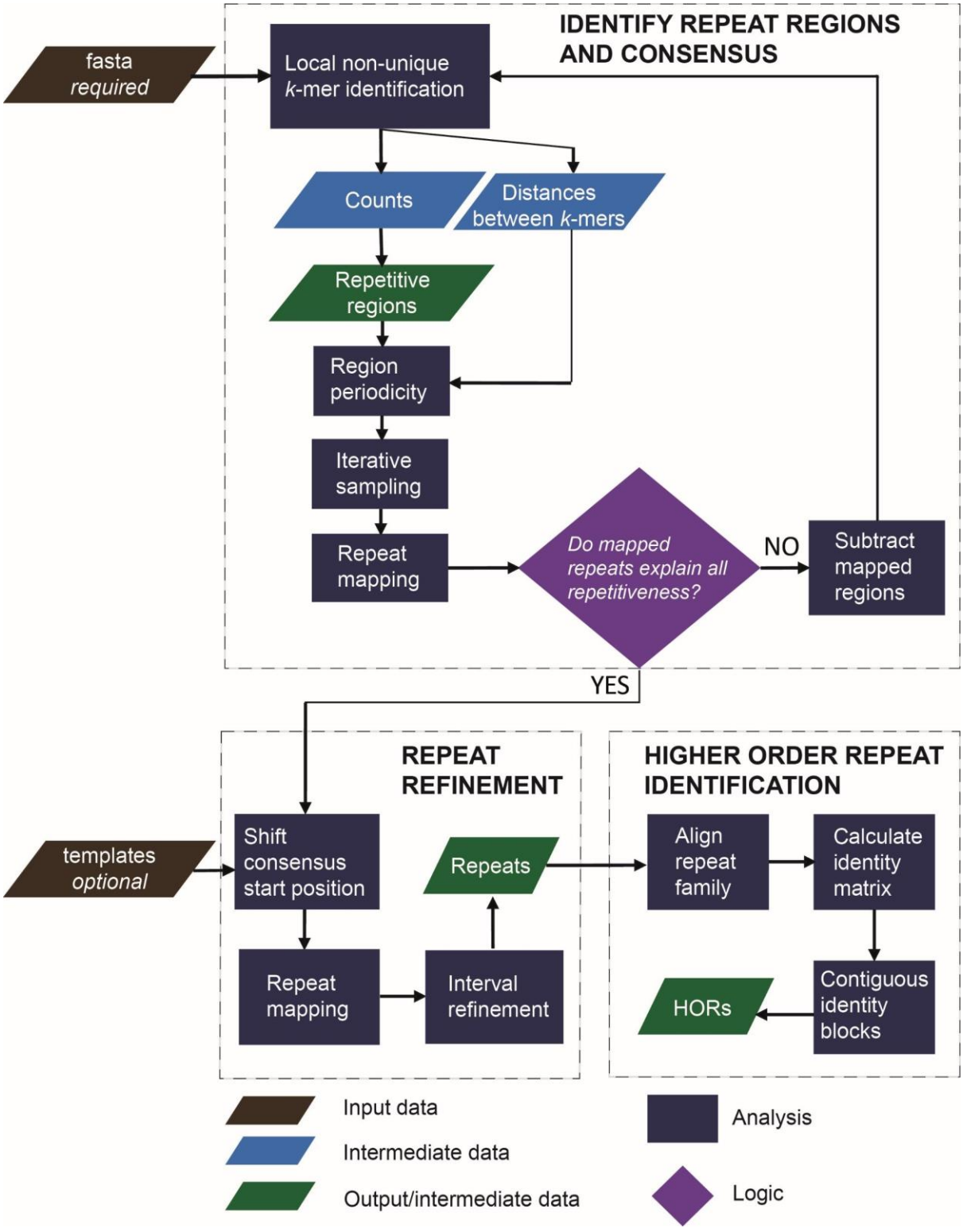


Figure 3.14. A simplified TRASH workflow diagram.
Individual steps of TRASH workflow are represented, starting with the input fasta file, and indicating all output TRASH produces. Repeat identification and repeat refinement modules indicated by dashed line boxes are always performed together, although repeat refinement can use optional input of repeat templates. The HOR identification module can be used only when those templates are provided. After the

initial run of TRASH which included repeat identification and repeat refinement, the HOR module can be used independently on the data provided before.

3.3 Discussion

TRASH can robustly identify tandem repeats in *Arabidopsis thaliana* chromosomes, on par with the alternative available software. Additionally, it can analyse their higher order structures, without a need for monomer class definitions. Importantly, prior information about repeat families is not required for TRASH to operate. Therefore, it can be used in a streamlined manner without much user input, which simplifies analysis making it both accessible and scalable, which will be shown in the next chapter.

The main disadvantage to TRASH is that it is restricted to detecting continuous arrays of tandem repeats. Interspersed repeats, including transposable elements, will not be annotated, unless they are arranged in tandem arrays. This is outside of the scope of TRASH, since the aim is to identify satellite repeats and software that handles interspersed repeats mapping exists, like RepeatModeller.

While the runtime of the repeat identification module scales linearly with the number of repeats it identifies, HOR module has an exponential complexity due to the pairwise comparisons and can become an issue with larger data sets. Potential solution could involve application of chained guide trees to find HOR seeds that could be expanded (Yamada 2016).

Future development of TRASH will focus on:

- Accessibility, by making TRASH available on Windows and iOS systems,
- Scalability, by code optimisations allowing for even shorter runtimes,
- Bug fixing and community support.

3.4 Acknowledgements

Michael Hong conceptualised and wrote the circos plots module and helped with testing of TRASH.

Chapter 4

Tandem repeat identification in plant species

This chapter presents tandem repeat profiles and their analysis from diverse plant species with varying relatedness levels. My aim was to understand the repeat composition of centromeric regions of plants, and to investigate patterns of their evolution. From these analyses, I speculate about the role of tandem repeat evolution in centromere function and chromosome segregation.

This chapter includes data and analysis from several publications:

- “The genetic and epigenetic landscape of the Arabidopsis centromeres”, Matthew Naish*, Michael Alonge*, **Piotr Wlodzimierz***, Andrew J. Tock, Bradley W. Abramson, Anna Schmücker, Terezie Mandáková, Bhagyshree Jamge, Christophe Lambing, Pallas Kuo, Natasha Yelina, Nolan Hartwick, Kelly Colt, Lisa M. Smith, Jurriaan Ton, Tetsuji Kakutani, Robert A. Martienssen, Korbinian Schneeberger, Martin A. Lysak, Frédéric Berger, Alexandros Bousios, Todd P. Michael, Michael C. Schatz*, Ian R. Henderson, Science, 2021, of which I am a co-first author.
- “Meiotic recombination within plant centromeres”, Joiselle B Fernandes, **Piotr Wlodzimierz** and Ian R Henderson, Curr Opin Plant Biol., 2019, of which I am a co-author.
- “Rapid cycles of satellite homogenization and retrotransposon invasion drive Arabidopsis pancentromere evolution”, **Piotr Wlodzimierz***, Fernando A. Rabanal*, Robin Burns*, Matthew Naish, Elias Primetis, Alison Scott, Terezie Mandáková, Nicola Gorringe, Andrew J. Tock, Daniel Holland, Katrin Fritschi, Anette Habring, Christa Lanz, Christie Patel, Theresa Schlegel, Max Collenberg,

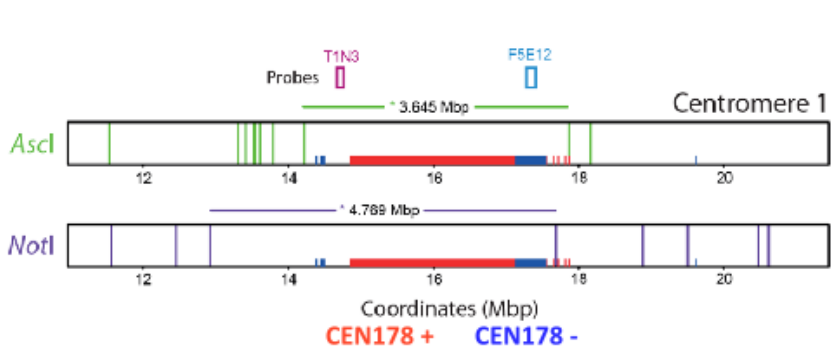
Miriam Mielke, Magnus Nordborg, Fabrice Roux, Gautam Shirsekar, Carlos Alonso-Blanco, Martin A. Lysak, Polina Novikova, Alexandros Bousios, Detlef Weigel and Ian R. Henderson, manuscript submitted, of which I am a co-first author.

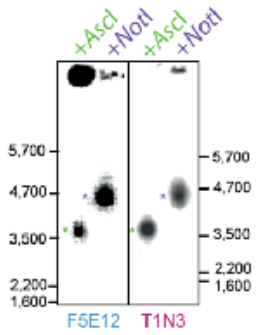
4.1 Results

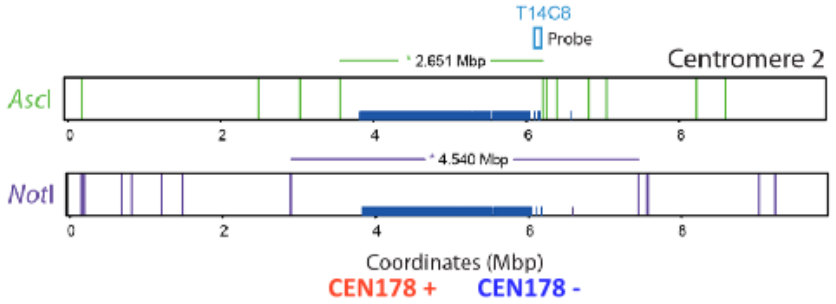
4.1.1 Repeat libraries of *Arabidopsis thaliana* Col-0 centromeres

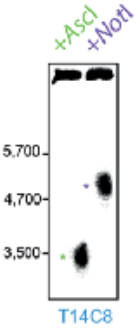
Complete, or nearly complete assemblies are essential in analysis of tandem repeat-rich centromeric regions. For *Arabidopsis thaliana,* this was recently achieved by our group and collaborators with the Col-CEN assembly (Naish et al. 2021). After initial genome assembly using Oxford Nanopore Technology (ONT) reads, polishing using PacBio HiFi reads and manual curation, chromosomes 1, 3 and 5 were completely sequence resolved, and chromosomes 2 and 4 had minor unresolved regions within *45S* ribosomal DNA (rDNA) (Naish et al. 2021). TAIR10 was the gold standard for the *A. thaliana* Col-0 accession genome (Lamesch et al. 2012), and in comparison, Col-CEN adds over 12 Mbp of genomic sequence (131,559,676 bp vs 119,146,348 bp). Because of this, all 5 centromeric sequences can be used for analysis, without concerns of their continuity affecting the results.

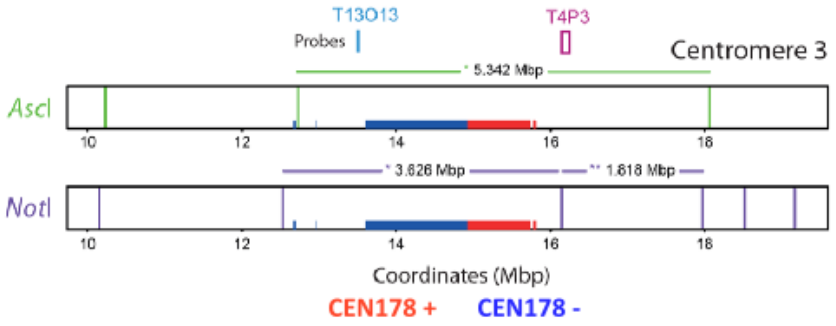
To validate the centromeric assembly, I compared *in silico Asc*I and *Not*I digested Col-CEN sequence with previously published pulsed-field electrophoresis and Southern blot analysis of bacterial artificial chromosomes (BACs) digested with the same enzymes (Kumekawa et al. 2000, Kumekawa et al. 2001, Hosouchi et al. 2002) (Fig. 4.1). All reported digestion fragments correspond to the predicted *in silico* digestion, with the exception of *CEN1* BAC F8L2. However, after inspection, this BAC was deemed to contain an incorrect *Not*I site, which would lead the original authors to believe that there is a 4.7 Mbp duplication in *CEN1* (Hosouchi et al. 2002). When accounting for that site, the Southern blot data is fully concordant with my *in-silico* analysis (Fig. 4.1).

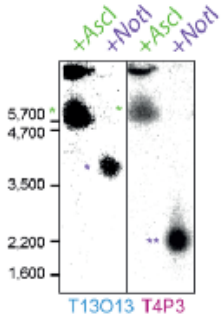

T1N3
F5E12
Probes
Centromere 1
AscI
NotI
* 3.645 Mbp
* 4.789 Mbp
Coordinates (Mbp)
CEN178 +
CEN178 -

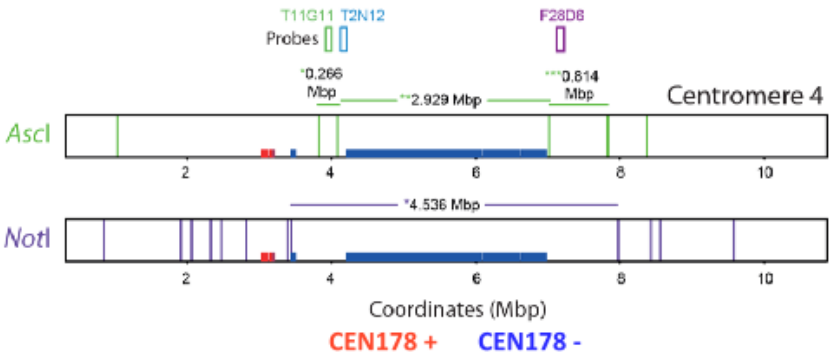

+AscI
+NotI
+AscI
+NotI
5,700
4,700
3,500
2,200
1,600
F5E12
T1N3

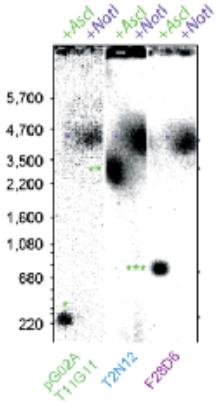

T14C8
Probe
Centromere 2
AscI
NotI
* 2.651 Mbp
* 4.540 Mbp
Coordinates (Mbp)
CEN178 +
CEN178 -

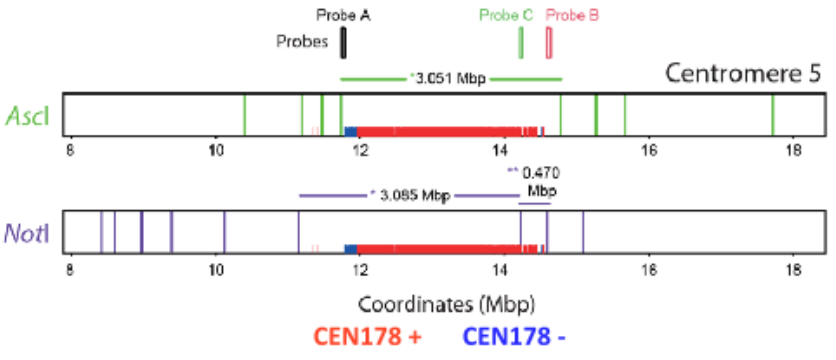

+AscI
+NotI
5,700
4,700
3,500
T14C8


T13O13
T4P3
Probes
Centromere 3
AscI
NotI
* 5.342 Mbp
* 3.626 Mbp
** 1.818 Mbp
Coordinates (Mbp)
CEN178 +
CEN178 -


+AscI
+NotI
+AscI
+NotI
5,700
4,700
3,500
2,200
1,600
T13O13
T4P3


T11G11 T2N12
F28D6
Probes
Centromere 4
AscI
NotI
*0.266 Mbp
**2.929 Mbp
***0.814 Mbp
*4.536 Mbp
Coordinates (Mbp)
CEN178 +
CEN178 -


+AscI
+NotI
+AscI
+NotI
+AscI
+NotI
5,700
4,700
3,500
2,200
1,600
1,080
680
220
pG02A
T11G11
T2N12
F28D6


Probe A
Probe C
Probe B
Probes
Centromere 5
AscI
NotI
*3.051 Mbp
** 0.470 Mbp
* 3.095 Mbp
Coordinates (Mbp)
CEN178 +
CEN178 -

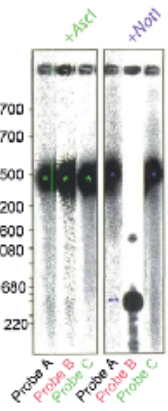

+AscI
+NotI
5,700
4,700
3,500
2,200
1,600
1,080
680
220
Probe A
Probe B
Probe C
Probe A
Probe B
Probe C

Figure 4.1. Comparison of the Col-CEN assembly with physical maps derived from pulsed-field gel electrophoresis and Southern blotting.
This figure was prepared by Prof. Ian Henderson according to the analysis I performed and published as a supplementary figure in Naish et al. 2021. On the right-hand side of the figure published pulsed-field gel electrophoresis and Southern blotting data are shown, where genomic DNA was digested using either *Asc*I or *Not*I (Kumekawa et al. 2000, Kumekawa et al. 2001, Hosouchi et al. 2002). The probe used for hybridization is labelled underneath the blots. To the left are physical maps of the Col-CEN assembly that have been virtually digested for *Asc*I (green) or *Not*I (purple) and restriction site locations indicated relative to chromosome coordinates. The position of plus strand (red) and minus strand (blue) *CEN178* are indicated on the x axis. Above each physical map the location of the probes used for Southern blot hybridization are indicated. The predicted size of cross-hybridizing fragments following restriction digestion are annotated above the physical maps, for comparison with the reproduced data.

Col-CEN repeat analysis was presented in the previous chapter (subchapters 3.2.6-8), where it was used to benchmark 'TRASH' performance and present its functionality (Figs. 3.9-12). In this section, the results of *CEN178* analysis performed using an early-development version of 'TRASH' are presented. Compared to the benchmarking analysis, these results contain additional mapping of *CEN178* repeats that are not arranged in tandem arrays, which were identified using the 'MatchPattern' function from R package Biostrings (Pagès et al. 2022). A total of 66,131 *CEN178* repeats were identified in Col-CEN and the overlap with the benchmarking identification is 96.42%.

The ability to distinguish chromosomes based on centromere-unique repeats was further investigated by calculating the number of shared and unique CEN178 repeats between the centromeres (Table 4.1). Each centromere contains a subset of private *CEN178* monomers, with only 0.6% of all repeats sharing an identical copy (or copies) on a different chromosome (Table 4.1). This is consistent with the model that satellite homogenisation occurs primarily within chromosomes. Despite this, diversity within chromosomes is also high, with the number of unique variants per chromosome reaching 44.4 to 58.9% of all chromosomal repeats (Table 4.1). Uniqueness can be achieved by just a single base pair discrepancy, so to better understand the similarity of repeats within each chromosome, percentage identity score (PID) was calculated for an alignment of repeats from each chromosome, defined as the percentage of pairwise residues that are identical in the alignment, including gap versus non-gap

residues, but excluding gap versus gap residues (Table 4.1). Consensus sequences derived from the five alignment were aligned together to highlight specific intra-chromosomal differences (Fig. 4.2).

Chromosome	*CEN178* repeats	Shared *CEN178* with chromosome:					Unique *CEN178* sequences	Alignment PID score
		Chr1	Chr2	Chr3	Chr4	Chr5		
Chr1	13,578	-	0	31	0	1	6,035	90.7%
Chr2	12,293	0	-	15	24	9	5,739	92.7%
Chr3	11,848	234	5	-	0	3	5,634	92.7%
Chr4	15,613	0	2	0	-	1	7,394	89.3%
Chr5	12,799	1	23	4	20	-	7,544	89.7%
Total:	66,131					Total:	32,346	

Table 4.1. *CEN178* repeats shared across chromosomes of *A. thaliana* Col-CEN assembly.
The number of identified *CEN178* sequences per chromosome; *CEN178* repeats from each of the “Chromosome” rows identified in each of the “Shared” columns; and number of unique *CEN178* sequences per chromosome. Alignment percentage identity score (PID) describes the percentage of pairwise residues that are identical in the alignment, including gap versus non-gap residues, but excluding gap versus gap residues.

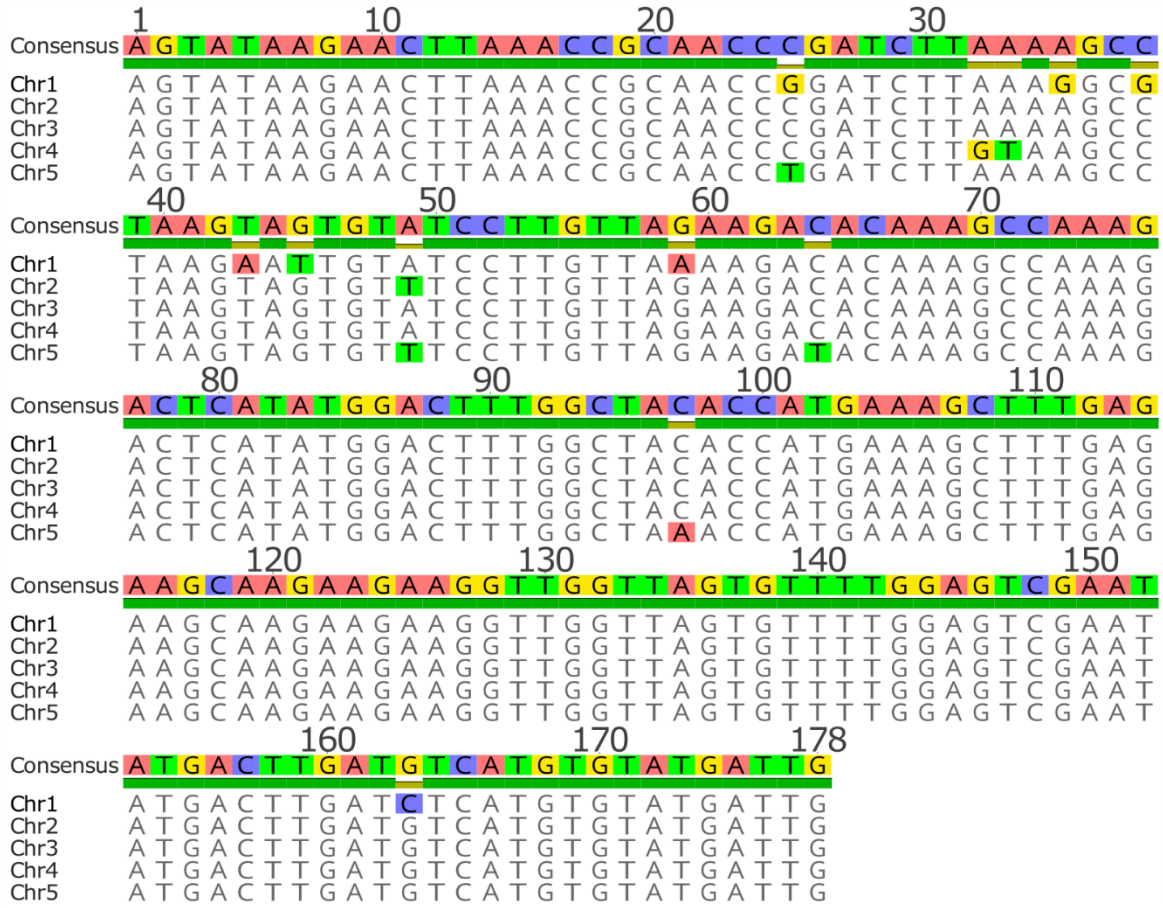


Figure 4.2. Alignment of the *CEN178* consensus sequences from each *A. thaliana* chromosome

CEN178 repeat sequences from each chromosome were aligned using mafft and their consensus were aligned together using mafft and visualised using Geneious Prime.

To validate these observations, DNA fluorescent *in situ* hybridization probes designed to recognize satellites unique to chromosome 1 (*CEN178-α*) and chromosome 5 (*CEN178-β*) were designed by finding all unique CEN178 sequences and extracting those in the highest copy number. FISH of pachytene-stage chromosomes was performed by Dr Terezie Mandáková and Prof. Martin Lysák from the Central European Institute of Technology, Czech Republic. The imaging included chromosome 1 specific BAC probes, which co-localised with the *CEN178-α* probe, while the *CEN178-β* probe signal could be found on separate chromosomes (Fig. 4.3).

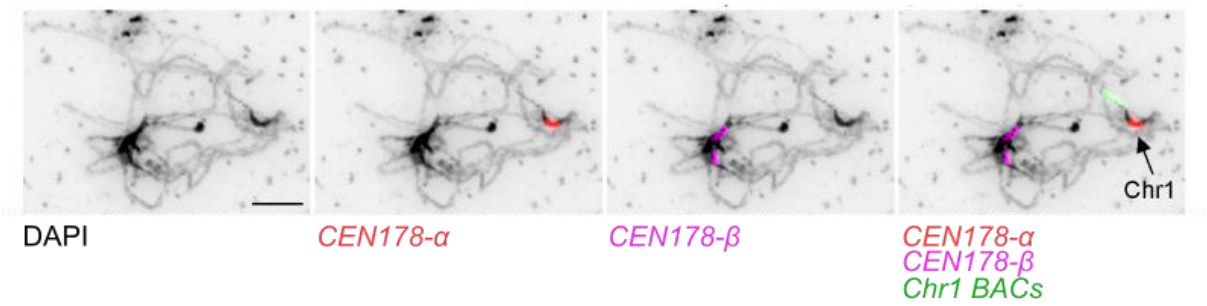


Figure 4.3. Cytological validation of repeat chromosome specificity by fluorescence *in situ* hybridisation (FISH).
Pachytene-stage meiotic chromosomes spreads stained with DAPI (black), and FISH performed using probes designed to label chromosome 1 *CEN178* (red), chromosome 5 (red) and chromosome 1 specific BACs (green). The scale bar represents 10 μM. The spreads were prepared and imaged by Dr Terezie Mandáková and Prof. Martin Lysák.

To define *CEN178* higher-order repeats, monomers were considered the same if they shared five or fewer pairwise variants, defined as alignment disagreements between them. Consecutive repeats of at least two monomers below this variant threshold were identified, yielding 2,408,653 higher-order repeats ('TRASH' settings –t 5 –c 2) (Fig. 4.4A). 95.4% of *CEN178* were involved in at least one HOR. Like the *CEN178* monomer sequences, higher-order repeats are largely chromosome specific, with only 1.71% of HORs (41,221) identified between chromosomes. Most higher-order repeats were short, with 3-monomer blocks being most common (Fig. 4.4B). The frequency of HOR block sizes (H_f) show a negative exponential distribution that fits a $H_f = 5^{10} \times H_s^{-4.452}$ trendline with $R^2 = 0.998$, where H_s is the number of monomers in HOR blocks (Fig. 4.4B). The largest HOR block was formed of 60 monomers (equivalent to 10,689 bp). Many higher-order repeats are in close proximity (26% are <100 kbp apart), although they are found to be dispersed throughout the length of the centromeres (Fig. 4.4C). I also observed that higher-order repeats with blocks further apart showed a higher level of variants per base pair between the blocks (Wilcoxon test $P<2.2\times10^{-16}$) (Fig. 4.4D), consistent with the idea that satellite homogenization is more effective over repeats that are physically closer.

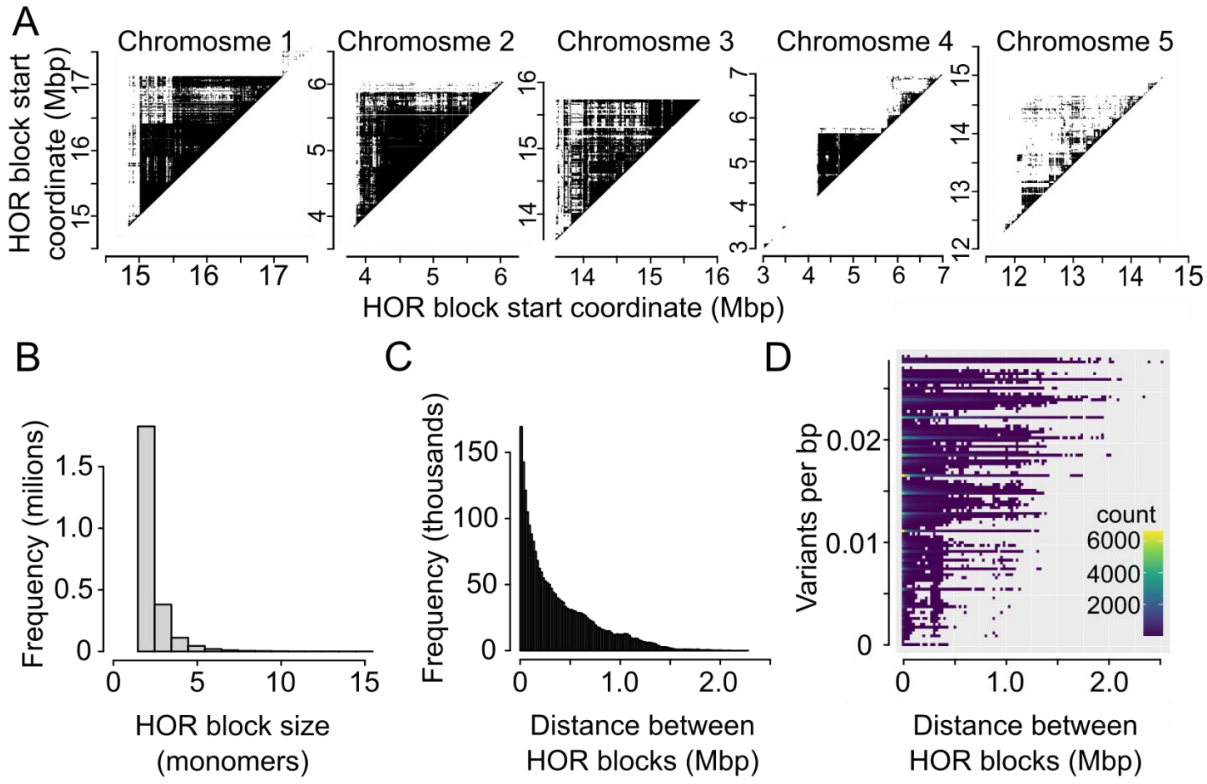


Figure 4.4. Col-CEN *CEN178* higher order repeat (HOR) identification and properties. **A**. *CEN178* HOR dot plots of the five A. *thaliana* centromeres, with the start coordinates of two blocks forming a HOR plotted. **B**. A histogram of the *CEN178* HOR block size distribution, with sizes above 15 monomers excluded for clarity. **C**. Histogram of the frequencies of distances between *CEN178* HOR blocks positioned on the same chromosome. **D**. Plot showing the relationship between the distance between HOR blocks and number of variants per bp (divergence metric).

CEN178 arrays were mostly uninterrupted, with the exception of 111 gaps that were over 1 kbp long. Within these gaps, 53 intact and 20 fragmented *ATHILA* long terminal repeat (LTR) retrotransposons of the Gypsy superfamily were identified by Dr Alexandros Bousios from the University of Sussex, United Kingdom. LTR comparisons indicate that the centromeric *ATHILA* are young, with, on average, 98.7% LTR sequence identity, which was significantly higher than that for *ATHILA* located outside the centromeres (96.15%, n=58, Wilcoxon test P=4.89×10^{-8}). I tested whether each *ATHILA* element per chromosome is an independent integration, or a duplication of *ATHILA*-containing repeats. In the second scenario, repeats surrounding two

centromeric *ATHILA* would be highly similar, and their insertion sites would align. Insertion sites relative to *CEN178* repeats can be mapped precisely due to the integration mechanism creating target site duplications (TSDs), where duplicated 4-bp genomic fragments flanking the element are observed (Voytas and Ausubel 1997, Linheiro and Bergman 2012). When considering pairs of *ATHILA* from the same chromosomes, their surrounding repeats did not share higher similarity levels than randomly extracted repeats from the same chromosome (Fig. 4.5A). When 20-bp flanking regions of each *ATHILA* were extracted to identify insertion sites, one pair on chromosome 4 had identical insertion site sequences, and also two independent pairs on chromosome 5, and one triplet on chromosome 5 (Fig. 4.5B). This suggests that most centromeric *ATHILA* have integrated independently, but they can be copied post-integration, potentially by the same mechanism that generates *CEN178* higher order repeats. Interestingly, the pair on chromosome 4 that shares flanking insertion sequences, contains one full length *ATHILA* and one solo-LTR element, and they are only 2.3 kbp apart. One of the *ATHILA* pairs on chromosome 5 is also a full length/solo-LTR pair, which are over 1 Mbp apart. With the frequency of *ATHILA* elements sharing insertion sites being relatively high (13 out of 29), one explanation could be that there is a preference in the *ATHILA* integration location along the *CEN178*. However, 11 out of the 13 *ATHILA* were also positioned within the same centromere, which on top of the integration preference along the CEN178, would also require preferential integration into a specific chromosome. Alternatively, duplication events happen post integration and involve surrounding repeats, which preserves the insertion site of the new element. Since repeat expansion is concentrated within a chromosome, it also explains why most shared insertion *ATHILA* are within the same chromosome. While centromeric *ATHILA* elements are young, with, on average, 98.7% LTR sequence identity (n = 53), those found outside of the centromeres show a significant decrease in the LTR identity (96.9%, n =58) (Naish et al. 2021). Whether the greater identity level in the centromeres is a result of higher integration levels or post-integration copying mechanism is unknown.

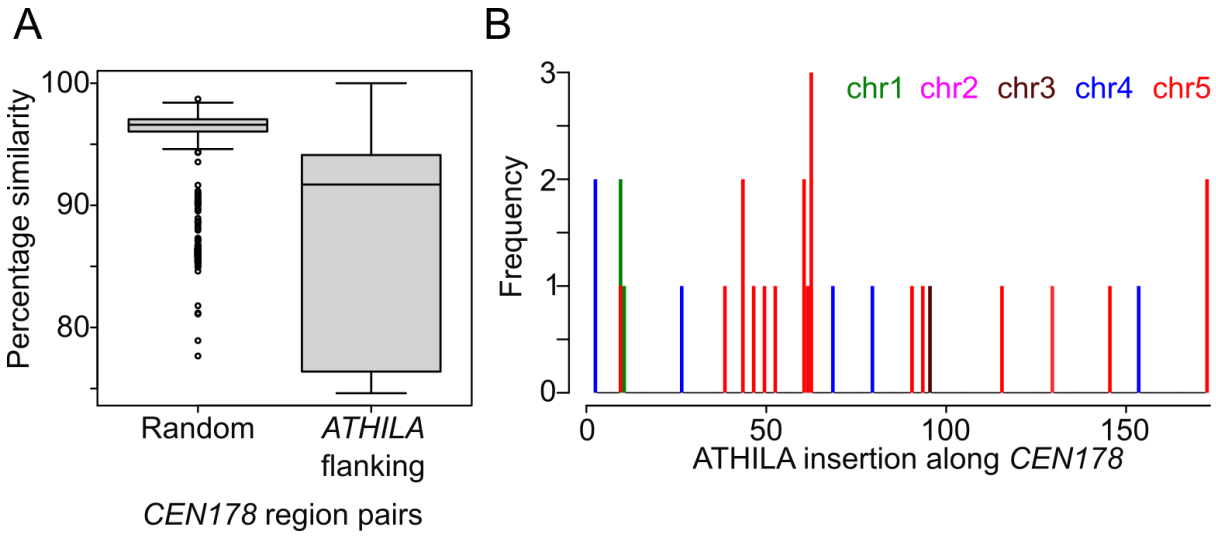


Figure 4.5. *ATHILA* integration patterns within the Col-CEN centromere arrays. **A**. The percentage similarity of repeats (20 monomers) between randomly extracted repeats (n=1,092) and repeats flanking *ATHILA* elements (n=546). Pairwise comparisons of *ATHILA* flanking regions were constrained to same-chromosome pairs. The distribution of random *CEN178* repeat pairs per chromosome was identical to that of *ATHILA* pairs. **B**. Insertion positions of 28 centromeric *ATHILA* along the *CEN178* consensus sequence. Only those elements that had their insertion site successfully mapped using both upstream and downstream sequence were considered. Highlighted are insertion sites of *ATHILA* elements sharing both chromosome and insertion site. Considering the scarcity of ATHILA sharing the insertion site sequences throughout this data, these instances suggest relatively rare post-integration duplication carrying the element together with neighbouring repeats.

CENH3 ChIP and methylation data was obtained by Dr Matthew Naish and mapped to the Col-CEN assembly and averaged across *CEN178* repeats by Dr Andrew Tock and Prof. Ian Henderson (Fig. 4.6). All chromosomes contain a similar level of CENH3 ChIP enrichment, and its deposition patterns overlap with the location of *CEN178* repeat arrays (Fig. 4.6A). At the chromosome scale, *CEN178* satellite numbers track closely the CENH3 enrichment (Fig. 4.6B). DNA methylation in CG, CHG and CHH profiles are also relatively increased in centromeric and pericentromeric regions (Fig. 4.6B). However, CHG DNA methylation shows relatively reduced centromeric frequency compared with CG methylation (Fig. 4.6B). This may reflect centromeric depletion of H3K9me2, a histone modification that maintains DNA methylation in non-CG contexts (Stroud et al. 2014, Naish et al. 2021). Fine scale mapping of these epigenetic features along *CEN178* repeats was performed by Dr Andrew Tock. Repeats that contained the

most CENH3, were also the ones with the least divergence relative to the *CEN178* chromosome consensus and had the highest CG DNA methylation (Fig. 4.6C). HORs counts per repeat do not show a consistent decrease with decreasing CENH3 occupancy, possibly due to the relatively low HORs numbers on chromosome 5, affecting the distribution (Fig. 4.6C). CENH3 nucleosomes show a phased pattern of enrichment with the *CEN178* satellites, with relative depletion in spacer regions at the satellite edges (Fig. 4.6D). CENH3 spacer regions also associate with increased DNA methylation and *CEN178* sequence variants (Fig. 4.6D), consistent with the possibility that CENH3-nucleosomes influence epigenetic modification and satellite divergence. Increased *ATHILA* integrations at positions of the lowest CENH3 enrichment was also observed (Fig. 4.5B and Fig. 4.17D).

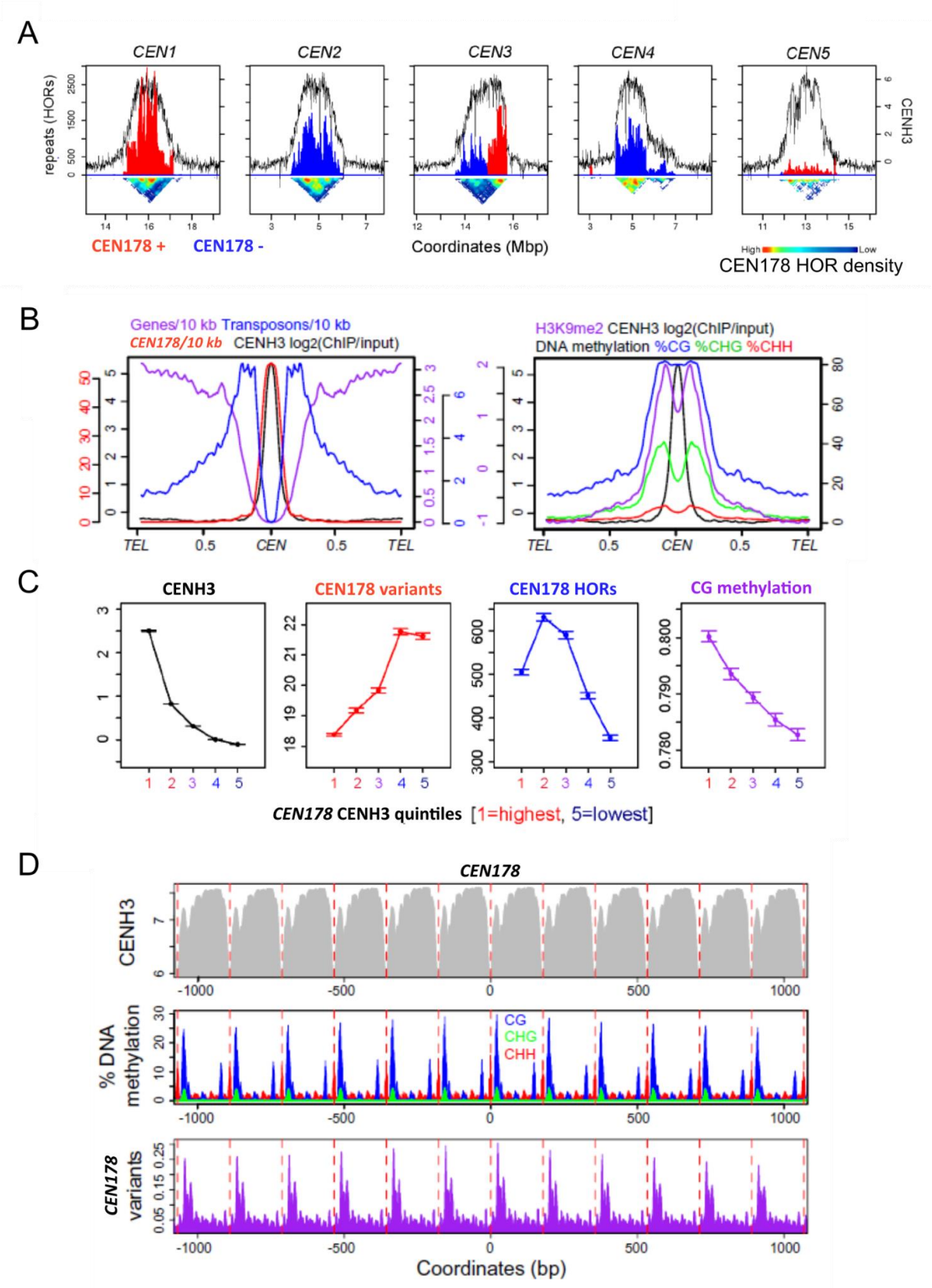


Figure 4.6. Epigenetic landscape of *A. thaliana* centromeres and *CEN178* repeats. **A**. *CEN178* repeat HOR counts per repeat (*CEN178* on the minus strand in blue, plus strand in red), CENH3 $\log_2$(ChIP/input) ratios in black and a *CEN178* sequence

similarity heatmap is shown below. **B**. Centromere to telomere averaged values of genes (purple), transposons (blue), *CEN178* (red) and CENH3 (black) on the left, and H3K9me2 (purple), CENH3 (black) and CG (blue), CHG (green), CHH (red) methylation on the right. **C**. *CEN178* repeats were divided into four quartiles based on their level of CENH3 occupancy. Average CENH3, *CEN178* repeat variants, HOR count and CG methylation per group were then plotted. **D**. CENH3, DNA methylation and sequence variant enrichment positions plotted over averaged *CEN178* arrays. This figure is adjusted from Naish et al. 2021.

Repeat homogenisation and HOR formation could be attributed to unequal crossover events, where a non-allelic template is used, resulting in contraction and expansion of the centromeric DNA. To investigate the role of the meiotic recombination in centromere evolution, 2,080 meiotic crossovers from Col×Ler F_2 sequencing data were mapped by Prof. Ian Henderson against the Col-CEN assembly, which were resolved, on average, to 1,047 bp (Naish et al, 2021). Crossovers were suppressed within and in proximity to the centromeres (Fig. 4.7). Therefore, unless only double (or another even number) crossover events occur that would not be identified due to the scarcity of SNPs. We conclude that centromeric crossover recombination is unlikely to be responsible for its evolution, although other recombination pathways including non-crossover repair could be active.

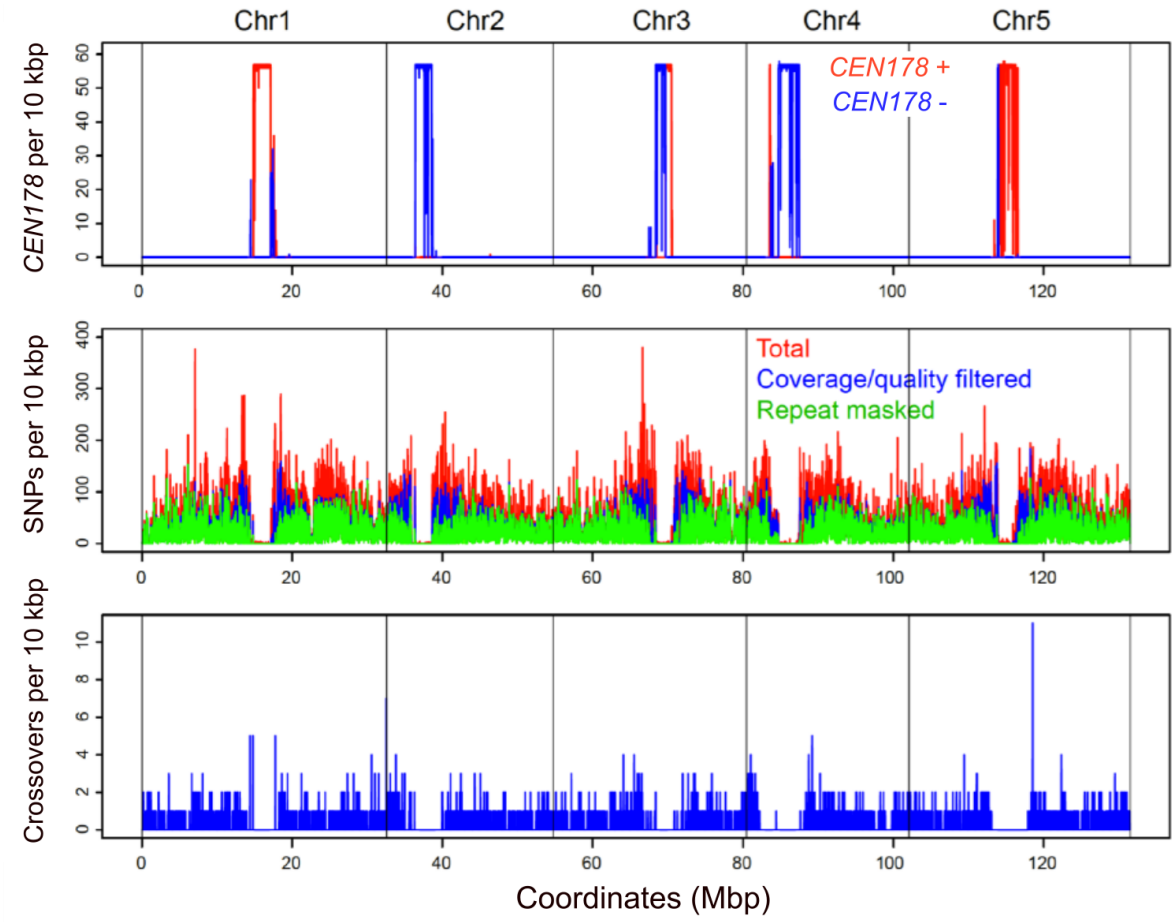


Figure 4.7 Meiotic crossover recombination suppression in the centromeric regions. **A**. *CEN178* repeats across the Col-CEN assembly. **B**. ColxLer SNPs frequency per 10 kbp (red), and the same set of SNPs filtered for quality (blue) and repeat masked (green). **C**. Crossovers per 10 kbp mapped against the Col-CEN assembly. This figure is adjusted from Naish et al. 2021.

4.1.2 Centromere satellite repeat evolution at a species level: analysis of 66 *Arabidopsis thaliana* assemblies

With centromeric repeat families being mostly restricted to individual species, due to their rapid evolution, a large set of same-species assemblies is required in order to perform centromere evolutionary analysis. Such an intra-species dataset was kindly provided by Dr Fernando Rabanal and Prof. Detlef Weigel from the Max Planck Institute for Biology in Tuebingen (65 assemblies), and Prof. Richard Durbin from the Department of Genetics at the University of Cambridge (1 assembly) (Table 4.2). The dataset consists of 66 *Arabidopsis thaliana* accessions assembled using PacBio HiFi

reads with a genome coverage range between 27-212×. 227 out of 330 chromosomes were gapless, meaning there was no ambiguity regarding the order, spacing or orientation of the underlying sequence. Phylogenetic tree of the pairwise sequence diversity based on the chromosome arm SNPs built using neighbour-joining method is presented in the Figure 4.8A. The pairwise sequence diversity has been calculated b Dr Robin Burns. Collection locations of the accessions represent a wide range of geographical positions throughout *Arabidopsis thaliana* native range (Fig. 4.8B) (Koornneef et al, 2004), while also including groups of geographically close accessions from the Iberian Peninsula and southern France (Fig. 4.8B). Principal component analysis (PCA) of single nucleotide polymorphisms (SNPs) in the chromosome arms performed by Dr Robin Burns confirmed the presence of four major genetic groups: Eurasian non-relicts, Iberian non-relicts, Iberian relicts, and non-Iberian relicts (Fig. 4.8C). The divergence time for these groups has been estimated at around tens to hundreds of thousands years ago (Hsu et al. 2019). It also revealed three pairs of accessions with nearly identical chromosome-arm SNPs, enabling the study of short-term centromere evolution (Table 4.2). This provides an unprecedented opportunity for studying centromere evolution in depth and at various relatedness levels.

Accession name	PCA group	Cen group chr1	Cen group chr2	Cen group chr3	Cen group chr4	Cen group chr5	CEN178 chr1	*CEN178* chr 2	*CEN178* chr 3	*CEN178* chr 4	*CEN178* chr 5
11C1	Eurasian	orphan	17	8	9	13	12191	14765	20569	14937	17401
Alo-0	Eurasian	orphan	19	23	9	13	18165	23495	10150	11930	18874
Alo-19	Relict	28	2	orphan	29	5	11176	15506	17556	23305	16819
ANGE-B-10	SW France	6	30	3	9	20	14655	19819	17306	15534	15559
ANGE-B-2	SW France	6	19	3	9	20	14772	18176	17373	15185	16042
AUZE-A-5	SW France	6	17	orphan	12	5	15950	9904	21564	8972	16708
BANI-C-1	SW France	6	14	8	12	20	15942	11160	11357	12295	16349
BANI-C-12	SW France	6	14	8	12	31	19333	12546	17281	9126	19568
BARA-C-3	SW France	24	14	32	12	20	20430	14043	14180	8965	12281
BARA-C-5	SW France	6	22	8	12	10	18557	16575	20351	9419	17144
BARC-A-12	SW France	24	17	15	9	20	22682	15171	9648	15563	16000
BARC-A-17	SW France	24	17	15	9	20	22683	15192	9600	15558	15998
BELC-C-10	SW France	6	33	23	26	20	10428	16565	13714	24279	15663
BELC-C-12	SW France	6	33	23	26	20	10401	16694	13740	24233	15625
Bon1	Eurasian	6	orphan	15	9	13	18220	16295	15311	19629	16770
BROU-A-10	SW France	34	14	15	12	20	16468	14225	12894	9231	8094
CAMA-C-2	SW France	24	11	35	9	20	13610	10137	14352	12058	15380
CAMA-C-9	SW France	24	11	35	9	20	13603	10022	14263	12061	15373

Cas-0	Eurasian	28	25	3	12	5	12967	19057	22132	9108	12242
Cas-6	Relict	28	25	3	4	13	19078	18721	19109	14920	18223
Cat-0	Relict	6	36	3	4	5	23602	12267	16192	25307	15255
Col-0	Eurasian	orphan	30	32	9	10	12991	12328	11912	15505	12245
Cvi-0	Atlantic	6	21	32	26	orphan	20708	13719	18775	21746	25755
ddAraThal4	Eurasian	6	11	23	9	10	15095	9601	21130	15746	17171
DraIII	Eurasian	16	11	8	9	10	21653	15008	14396	17703	18540
Etna-2	African	6	19	orphan	12	13	27134	13428	17527	19740	18984
Evs-0	Eurasian	6	19	3	12	13	14186	14272	13268	15545	17472
Evs-12	Eurasian	24	19	8	12	13	17827	17499	7438	15543	20088
Ey15-2	Eurasian	6	11	15	12	13	20402	9768	16864	12884	15288
FERR-A-12	SW France	6	22	15	9	20	15563	17627	16664	15225	24600
FERR-A-8	SW France	6	7	15	12	20	10538	15658	17591	14353	15164
GAIL-B-11	SW France	6	14	23	12	20	10381	14735	14388	11170	16110
Gel1	Eurasian	6	14	15	9	13	18422	8505	8647	11145	17888
Hom-0	Relict	28	2	8	4	5	27402	16244	18450	16147	11713
Hom-4	Eurasian	6	30	32	4	13	20715	9595	20643	17344	18075
HR-10	Eurasian	6	11	8	12	13	6413	9572	24307	11044	15148
Hum-2	Relict	1	7	3	4	13	14862	15349	20339	23890	16076
Hum-4	Eurasian	34	17	15	9	5	14580	15316	9253	16201	9997
Iasi-1	Eurasian	6	11	8	9	18	12972	6092	14907	17008	19515
IP-Bus-0	Eurasian	6	21	orphan	12	20	14549	10972	13807	11022	11224
IP-Fel-2	Eurasian	6	22	23	9	10	7025	16368	23125	11579	12346
IP-Ini-0	Eurasian	24	25	8	26	13	17894	20355	13028	28150	19489
IP-Per-0	Relict	27	2	3	4	5	18788	23676	17860	10590	19196
IP-Piq-0	Eurasian	27	19	23	9	5	20868	17343	11361	19114	12530
IP-Tri-0	Eurasian	6	orphan	15	9	5	16078	12195	18823	12297	16629
LACR-C-14	SW France	24	14	23	12	20	21822	14456	21320	8940	10282
Ler-0_110x	Eurasian	6	14	8	9	18	10841	8666	15889	14019	18539
Lor-16	Relict	28	36	3	4	5	11846	10161	16707	20843	24898
Mdc-14	Eurasian	27	25	3	9	5	22437	19496	22042	16606	16835
Med-0	Relict	28	2	32	4	orphan	17093	21432	17155	13901	21835
Med-3	Eurasian	orphan	30	8	9	13	20932	9963	14743	14452	18472
MERE-A-13	SW France	24	21	3	9	31	22379	11348	17182	16382	22663
Met-6	Relict	28	19	35	26	13	10317	13544	18424	27067	19773
MONTM-B-16	SW France	6	22	8	9	10	17290	17380	15433	11423	13252
MONTM-B-7	SW France	6	7	32	9	13	22394	15619	14783	11570	16508
Mos-5	Eurasian	6	19	8	9	10	11906	13183	13055	15501	17235
Mos-9	Eurasian	27	2	3	29	5	24208	16571	15003	22275	23191
PREI-A-14	SW France	6	7	23	12	20	18681	15702	13540	8995	15059
Rabacal-1	African	1	2	3	4	5	21514	22565	33712	16089	31306
Ru-2	Eurasian	16	14	3	9	20	21487	13788	18582	18614	10417
SALE-A-10	SW France	37	11	15	12	20	17724	9652	17719	11099	15631
SALE-A-17	SW France	37	11	15	12	20	17651	9684	17691	11061	15435
San-9	Relict	1	25	3	4	31	15247	18983	12155	23568	19293
Sln-22	Relict	6	2	orphan	4	5	15629	19639	14669	18496	21055
T850	Eurasian	6	7	8	9	10	18604	15990	16666	16525	17117
Tanz-1	African	orphan	orphan	3	orphan	orphan	34131	16279	22448	11050	35254

Table 4.2. *Arabidopsis thaliana* accessions with their chromosome arm PCA group, and centromere similarity group per chromosome and repeat number per chromosome.

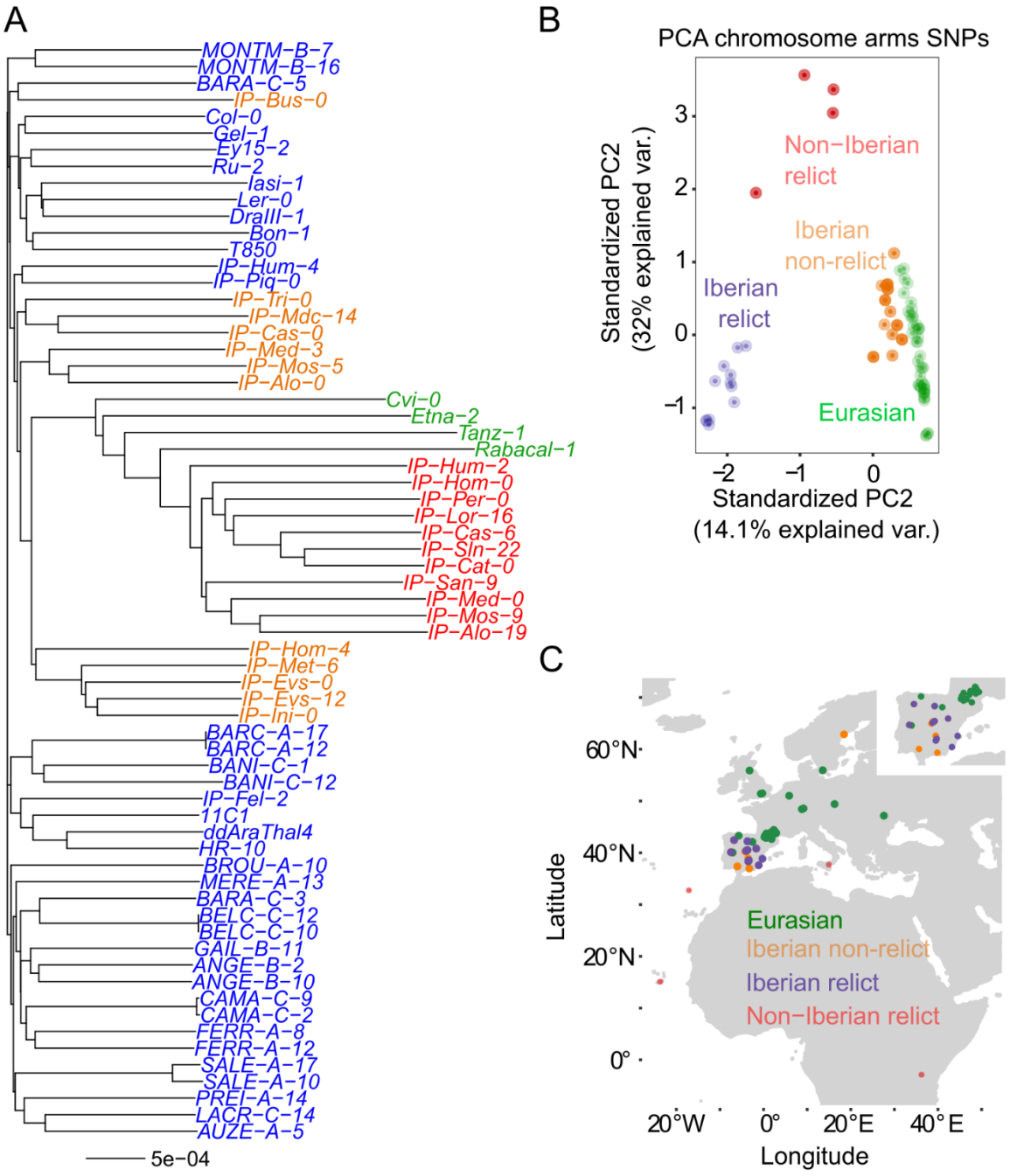


Figure 4.8. Chromosome arm based phylogeny and PCA analysis of the analysed accessions and their geographical distribution.

A. Neighbour-joining phylogenetic tree of the pairwise sequence diversity based on the chromosome arm SNPs. Accessions labels coloured according to the Principal Component Analysis (PCA) results shown in panel B. **B.** PCA of chromosome arm SNPs showing distinct Eurasian (green), Iberian non-relict (orange), non-Iberian relict

(red) and Iberian relict (purple) genetic groups. **C.** Geographic origin of the accessions coloured according to the PCA group membership. The figure was prepared by Dr. Robin Burns and modified for presentation here.

TRASH was run on all *Arabidopsis* accession assemblies, using a maximum repeat period of 1,000 bp. This window length is enough to capture the major tandem repeat arrays of *CEN178* (around 178 bp), *CEN159* (around 159 bp) and *5s* rDNA (around 500 bp). 7,710,759 tandem repeats were identified in total using sequence templates of *CEN178*, *CEN159*, and *5S* rDNA provided during the run. 5,810,923 (75.36%) of the tandem repeats were classified to one of these families (Fig. 4.9A). In addition, a large fraction of repeats were short tandem repeats under 10 bp (*n* = 968,890). Based on a manual inspection, they represent telomeric 7-bp repeats 5'-(TTTAGGG)(n)-3', and other short tandem repeats (STRs) with 1-6 bp units. These might be functionally significant, as variation in the STR copy numbers was found to have phenotypic effects in *Arabidopsis* (Sureshkumar 2009, Press et al., 2018). After accounting for these short repeats, 930,946 (12.07%) repeats were left with unknown characteristics. Classified repeats included 5,345,259 *CEN178*, 137,520 *CEN159* and 276,969 *5S* rDNA repeats (Fig. 4.9B-D). To examine the distribution of the classified repeats, they were first divided into individual arrays, where distance between consecutive repeats was lower than 1,000 bp. To account for the variable size of the individual chromosomes, these values were normalised by the length of the chromosome they occupy. Then, the middle point of each array was plotted along the x axis for each chromosome, with the tandem repeat array size on the y axis (Fig. 4.9E). With *Arabidopsis thaliana* chromosomes being monocentric, *CEN178* repeats strongly cluster in a similar position across the accessions (Fig. 4.9E). Much shorter *CEN159* arrays occur mostly outside of, and proximal to, the *CEN178* arrays across all chromosomes, and *5S* rDNA clusters can be found in relatively high copy on chromosomes 3, 4 and 5, in agreement with previously published data (Cloix et al 2002, Simon et al 2018). Interestingly, signatures of the *5S* rDNA repeats were also found on chromosomes 1 and 2 (Fig. 4.9E).

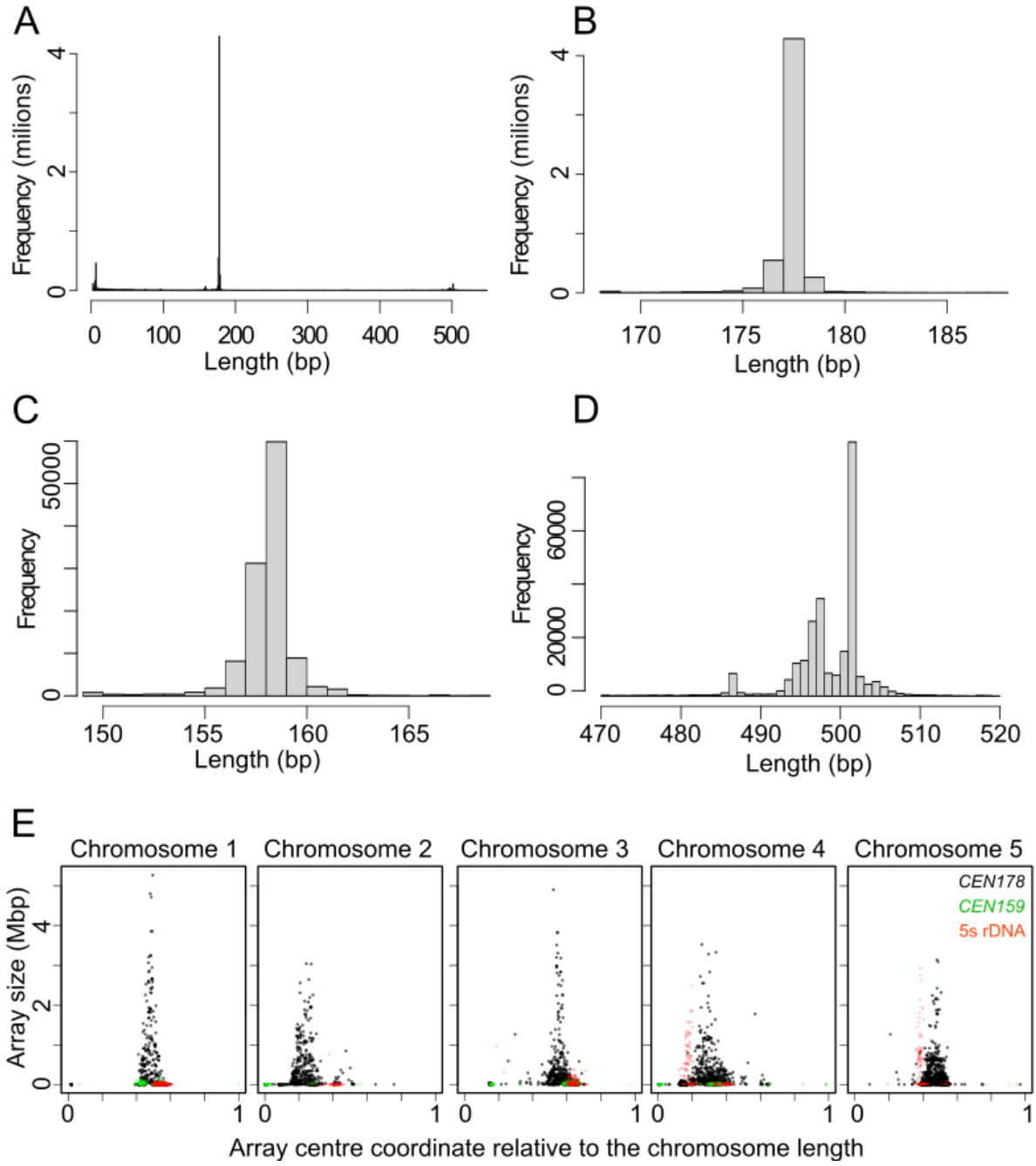


Figure 4.9. Identification of *Arabidopsis thaliana* repeats and tandem repeat arrays positions.

A. Histogram of all identified tandem repeat lengths (n=7,710,759). 4 visible peaks correspond to: 7 bp telomeric repeats, 159 bp *CEN159* repeats, 178 bp *CEN178* centromeric repeats and *5S* rDNA repeats. Repeats were classified to 3 families: *CEN159*, *CEN178* and *5S* rDNA. Their histograms are plotted in **B**-**D** respectively. Repeats from each family were divided into arrays, defined as consecutive repeat groups where the distance between each neighbouring repeat is under 1 kbp, their midpoint positions were normalised against length of the chromosome they are a part of and plotted against their sizes in Mbp (**E**).

Pairwise centromere similarity based on the number of shared identical *CEN178* repeats (shared repeat similarity, SRS) between each pair of accessions and each pair of chromosomes was calculated. The values were normalised by the number of repeats in both sets, so that they are in range of 0-100%, where 0% means no shared repeats and 100% means all repeats in the first set can be found in the second set and *vice versa*. The average SRS score between accessions was 1.86%, and between chromosomes the score was 1.84% (Fig. 4.10A). The majority of centromere similarity can be attributed to the similarity between same chromosomes across the accessions, where average similarity was 9.46% for chromosome 1, 5.19% for chromosome 2, 8.03% for chromosome 3, 8.90% for chromosome 4, and 12.52% for chromosome 5 (Fig. 4.10B-F). The average SRS between chromosomes within accessions was only 0.11% and between chromosomes between accessions it was similarly low, at 0.12%. This suggests that while *CEN178* arrays evolve rapidly, the mechanisms behind these processes are constrained to the individual chromosomes.

Discrete centromere similarity groups can be identified for individual chromosomes across all accessions. At a 10% SRS threshold, 37 groups were identified (Table 4.2). Only the same chromosomes formed groups, highlighting low cross-chromosome similarity. 8 *CEN1*, 12 *CEN2*, 6 *CEN3*, 5 *CEN4* and 6 *CEN5* groups were identified. (Table 4.2). 17 chromosomes were not included in any of these groups and were described as ‘orphan’ centromeres. It is expected that wider sampling of *Arabidopsis thaliana* accessions will identify chromosomes similar to the orphans, as none of the accessions contained orphan-only centromeres. Even one of the centromeres (*CEN3*) of the divergent Tanz-1 accession (Tanzania) was similar at the 10% threshold to *CEN3* of the Lor-16 (Spain) accession.

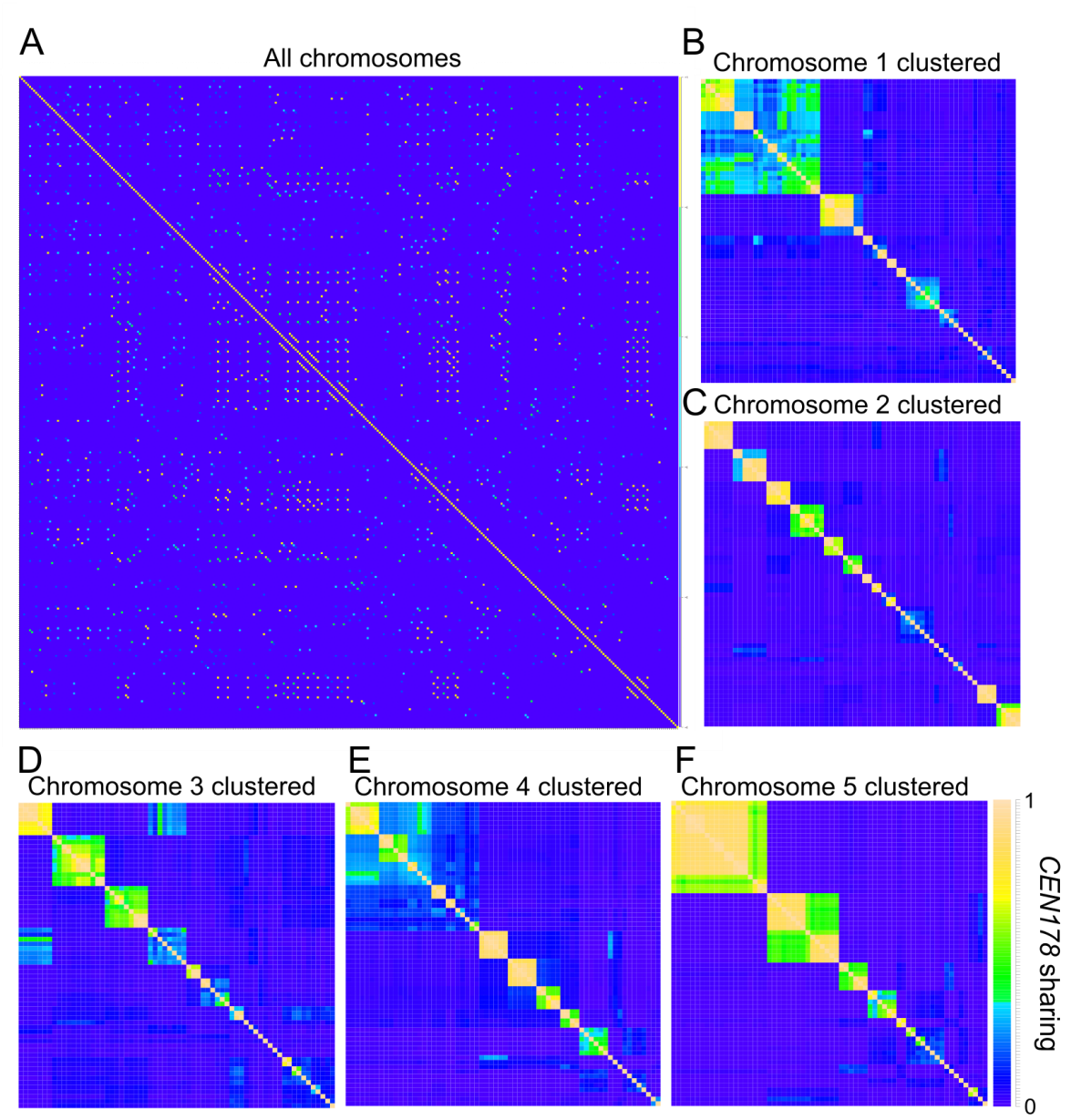


Figure 4.10. Pairwise centromere shared repeat similarity (SRS) scores and clustering per chromosome.
A. Centromere similarity was calculated as the fraction of identical repeat pairs between two chromosomes plotted for all pairs of individual chromosomes. (**B-F**) The same as A, but centromeres are clustered using a complete linkage method for each chromosome separately. Discrete clusters correspond to centromere similarity groups from Table 4.2 and Fig. 4.11.

Chromosome groups can be represented as graphs, where each vertex (or node) corresponds to a centromere, and the connecting edges are similarity scores, with only nodes above the 10% SRS threshold plotted (Fig. 4.11A-B). A high number of

edges per vertex in a graph (i.e., each chromosome has above threshold similarity with most or all other chromosomes) can indicate common ancestry for the chromosomes within the group, as they likely share an identical subset of repeats (Fig. 4.11A). Alternatively, a low number of edges per vertex (for example chromosome A shares repeats with chromosome B, but not chromosome C, while chromosome B is similar to chromosome C) suggest intermediate vertices being mixes of the surrounding ones, a situation that could arise due to recombination between the centromeres (Fig. 4.11B). In the first case, the number of edges would approach the theoretical maximum of:

where *V* is the number of vertices. Most graphs have the former characteristic, which would agree with the reported scarcity of crossover events in the centromeres that would lower the edge numbers (Fig. 4.11C) (Vincenten et al. 2015, Rowan et al. 2019).

Similarity groups were not overlapping between the chromosomes, only four accession pairs grouped identically across all five chromosomes, with three of these having near-identical chromosome-arm SNPs. Twenty accession pairs (out of 2,145 pairs total for 66 accessions) shared groups across four chromosomes (Fig. 4.11D). While overall centromere similarity is low between the accessions, it is highest in close geographical proximity and decays with increasing distance (Fig. 4.11E). Overall, this is consistent with *Arabidopsis thaliana* centromeres acting as non-recombining loci with respect to centromeres on other chromosomes.

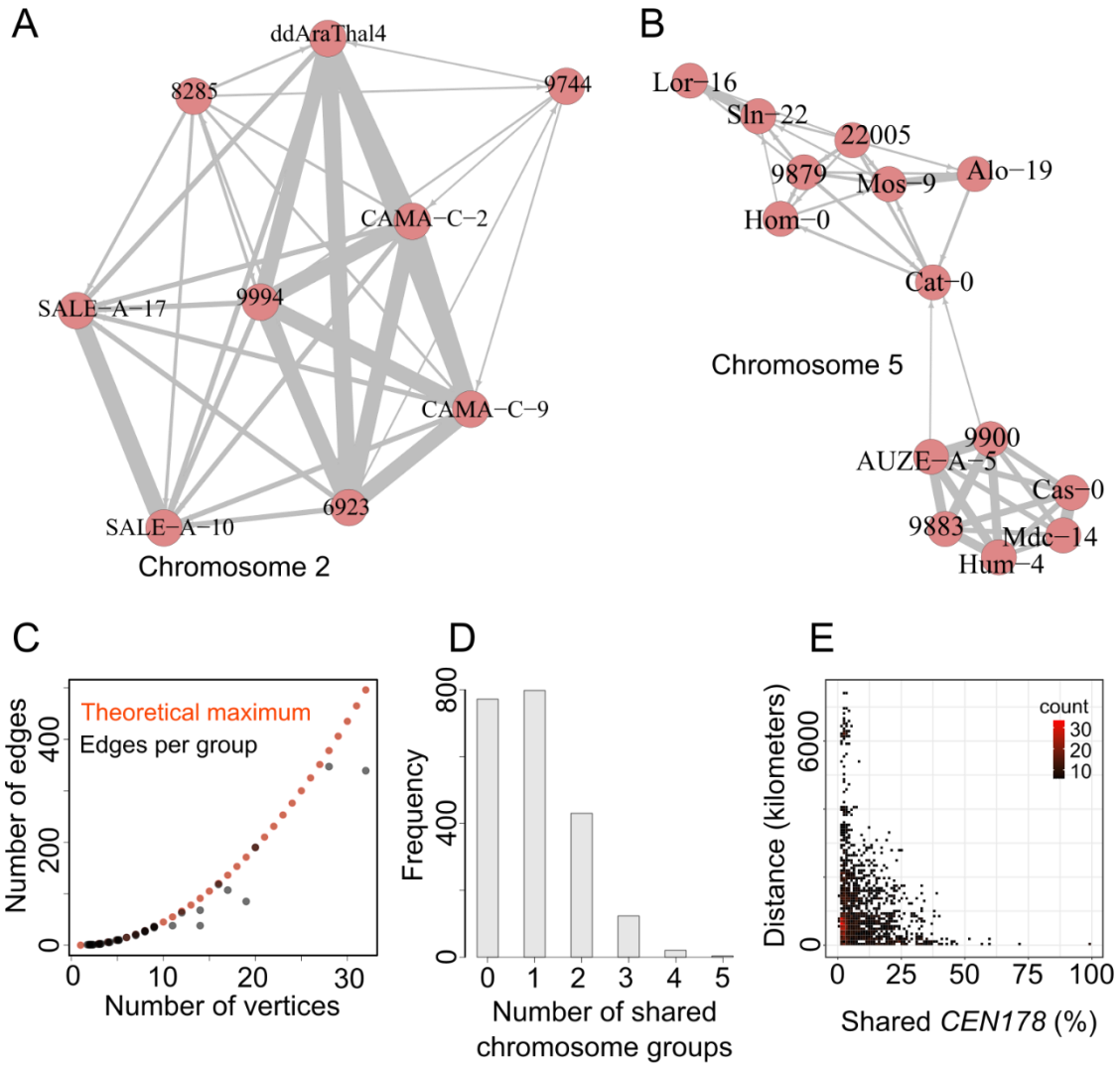


Figure 4.11. Centromere similarity groups characteristics.
A. A graph of one of the chromosome 2 centromere similarity groups. Edge widths correspond to the centromere similarity score between a pair of chromosomes. Almost all pairs are connected by an edge indicating high similarity within all pairs of this group. **B**. Chromosome 5 centromere similarity group example with two clusters connected with only one intermediate vertex. **C**. The number of edges versus number of vertices in the identified groups (black), or the theoretical maximum (red). **D**. Frequency of chromosome pairs that belong to a given number of chromosome groups. **E.** 2-D histogram of shared *CEN178* fractions between individual chromosomes versus geographical distance between the accession collection locations.

To analyse the sequence composition of the identified repeats, the consensus sequences of all classes were built using mafft (settings: -retree 1) alignment (Katoh

et al. 2013) (Fig. 4.12A-C). The base identity level for each position was calculated using the fraction of the consensus base in the whole alignment. Across the *CEN178* consensus sequence, identity levels remained mostly above 90%, with 14 positions having low identity under 75% (Fig. 4.12A). The non-normal distribution of sequence identity levels might suggest functional importance of these positions, especially when putative methylation sites are disrupted by cytosine or guanine replacement or represent polymorphisms across the population. In contrast, sequence identity levels of the *CEN159* consensus positions increase distally (Fig. 4.12B). One explanation might be that mutations occur preferentially at central positions, or that *CEN159* do not homogenise to the same extent as *CEN178*. The *5S* rDNA identity profile is similar to *CEN178*, possibly due to its conservation based on the functional importance coding 5S rRNA ribosome components, or similar mechanisms of homogenisation to *CEN178* (Fig. 4.12C). Two discrete identity levels at around 92% and 98% might be contributed to by the multi-modal distribution of the *5S* rDNA repeats lengths (Fig 4.12D)

Despite being the most numerous and having low centromere similarity levels across the accessions presented before, *CEN178* repeat average identity was the highest at 90.7%, with *CEN159* average identity of 68.0% and *5S* rDNA average identity of 88.0%. This agrees with the findings of high pairwise similarity within *Homo sapiens* centromeric alpha satellite repeats within active centromeres (Sullivan et al, 2017, Altemose et all. 2022) and suggests rapid repeat homogenisation that can be marked by the presence of higher order repeats (HORs) (Sujiwattanarat et al, 2015, Suzuki et al, 2020).

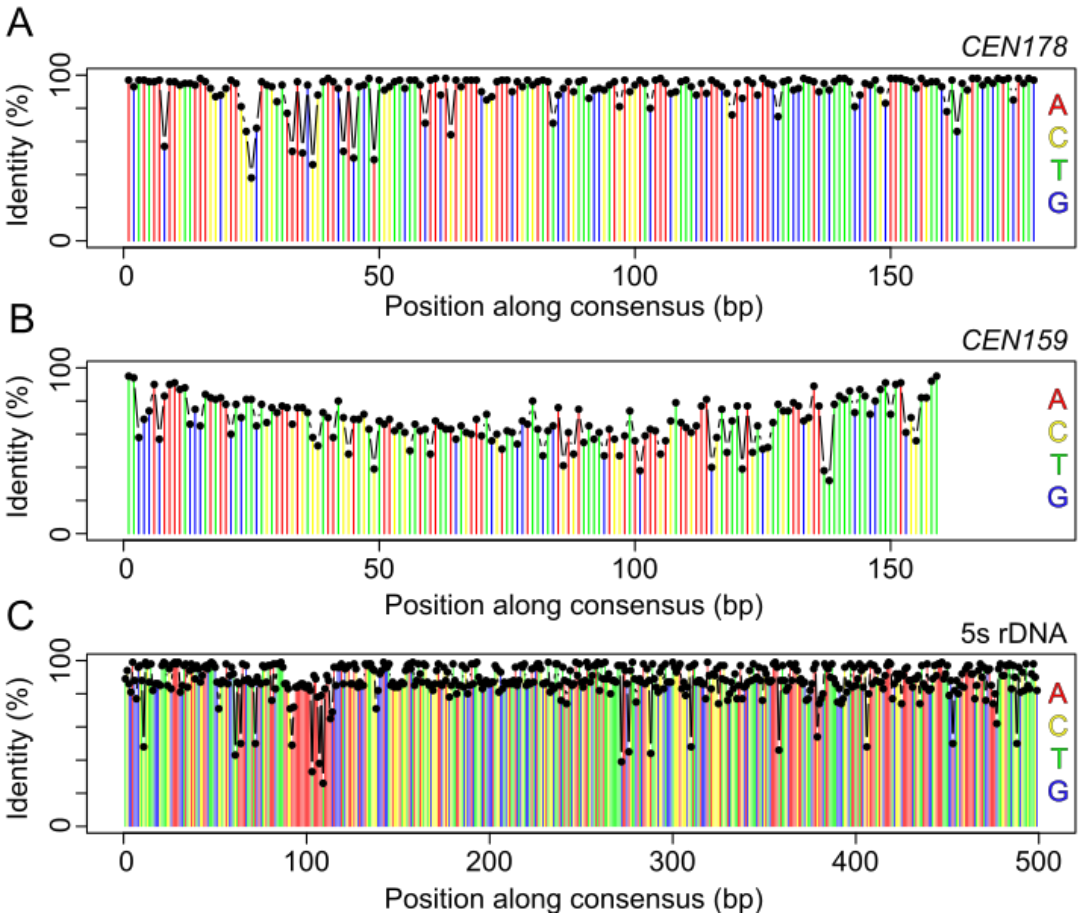


Figure 4.12. Sequence identity along consensus of *CEN178*, *CEN159* and *5S* rDNA repeats.
CEN178 (**A**), *CEN159* (**B**) and *5S* rDNA (**C**) repeat consensus sequences represented as coloured bars, where cytosine is in yellow, guanine in blue, adenine in red and thymine in green. Bars y value, highlighted with black dots, represents the fraction of the repeats that contain the consensus base at this position.

Higher order repeat analysis of *CEN178* (TRASH settings: --t 5 –c 3) identified 98,265,545 instances across all accessions. Analysis was restricted to within chromosomes. Most *CEN178* HOR blocks consisted of 3 monomers, the smallest possible value with the used settings. From there, *CEN178* HOR block sizes showed an exponential decay (Fig 4.13A). This indicates that either duplications that form HORs are inherently short, or they are initially long, but then become disrupted as the arrays accumulate mutations, and/or new HORs are inserted within them. The distance between HOR blocks was similarly short and the number of longer distances also showed an exponential decay (Fig. 4.13B). This was even more pronounced when only HORs with identical blocks were considered, strongly suggesting that the majority

of HORs are created locally (Fig. 4.13B). With no clear dominant HOR size, the distances between the closest identical repeats might provide insight into the duplication mechanisms. The HOR distribution peaks at around 1,780 bp, which corresponds to ~10 *CEN178* monomers (Fig. 4.13C). Overall, HORs appear over short distances with a small number of repeats being duplicated. This overlaps with the distance of observed *Arabidopsis* meiotic gene conversion tracts, which typically range between 2-2,000 bp in size (Yand et al. 2012, Wijnker et al. 2013). This is consistent with the molecular mechanism behind the HOR acquisition involving homologous recombination between sister chromatids or homologous chromosomes (in mitosis and meiosis), with possible non-allelic template choice, creating new HORs from ectopic locations in the centromere, but with proximal events preferred. The large number of HORs indicates repeat homogenisation through duplication of existing array elements. *CEN178* repeat size is heavily constrained to 178 bp (Fig. 4.9B), and only 12,032 gaps in the size range of 1 to 200 bp can be identified. This lack of truncated, or partial, repeats suggests a mechanism that relies on a homologous template to duplicate repeats ‘in frame’.

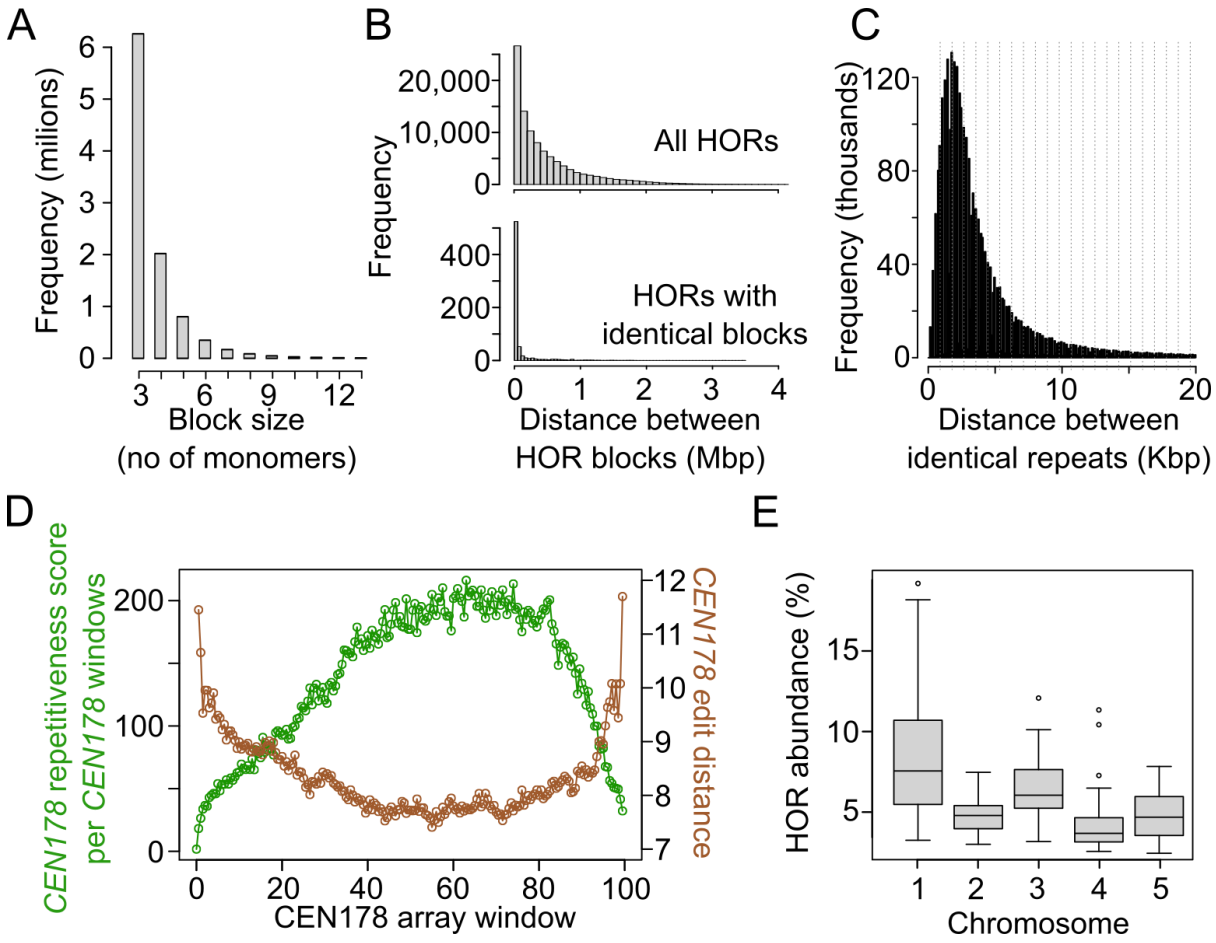


Figure 4.13. *CEN178* higher order repeats distribution and characteristics.
A. Histogram of *CEN178* HOR block sizes (bp). **B.** Histograms of distances between the HOR blocks, for all HORs (top), and only HORs with identical blocks (bottom). **C**. Histogram of distances between closest identical repeats. Dotted vertical lines indicate increments of 890, equivalent to ~5 *CEN178* repeats. **D**. The distribution of HOR-based *CEN178* repetitiveness scores and *CEN178* edit distance scores averaged across all centromeres and plotted on a scaled axis. **E.** Chromosome based HOR abundance score distributions plotted for each chromosome.

To measure the involvement of each *CEN178* repeat in HORs, their repetitiveness score was calculated as the sum of the lengths of monomers of all HORs that repeat is a part of. A measure of diversity across the repeats was calculated using Levenshtein edit distance, described as single character edits required to change a repeat into the consensus of all repeats from the same chromosome. Across all centromeres, the central *CEN178* arrays showed highest HOR activity and lowest edit distances (Fig. 4.13D), reminiscent of human alpha-satellite arrays (Miga 2021,

Altemose 2022). The average HOR abundance across all chromosomes was 5.67% with chromosomes 1 and 3 having the highest average abundance score (**Fig. 6.12E**).

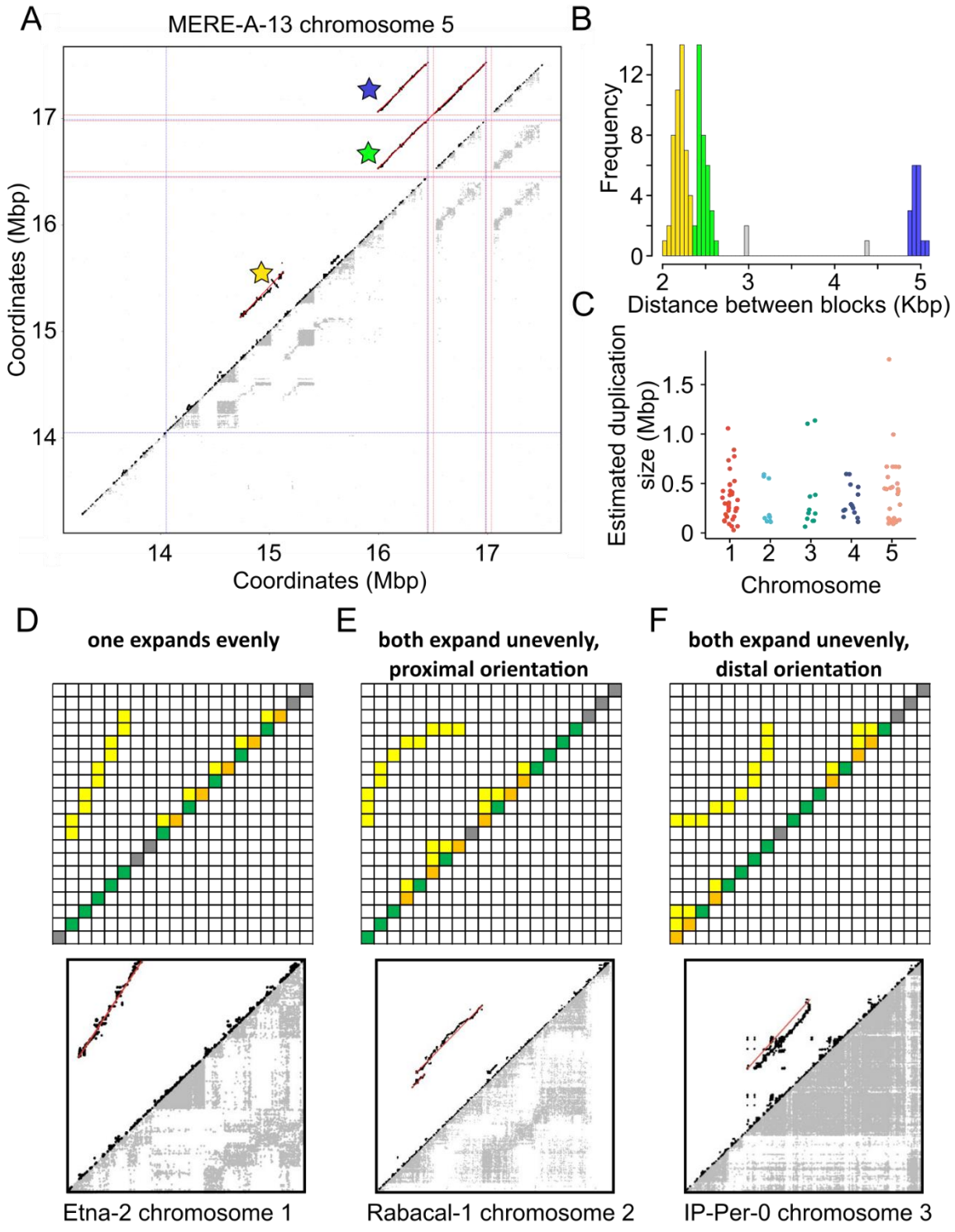


Figure 4.14. Identification of large duplications across the *CEN178* arrays.
A. Dot-plot of identified *CEN178* HORs in MERE-A-13 centromere 5. Dots represent HORs identified with less stringent settings (--t 5 –c 3, grey), versus more stringent settings (--t 0 –c 5, black). Identified duplications are highlighted in red and marked by

stars with colours corresponding to the histogram peaks. Horizontal and vertical dotted lines are *ATHILA* transposable element positions, with red lines being full length elements and blue lines representing solo-LTRs. **B**. Histogram of distances between the blocks of *CEN178* HORs identified with stringent settings across chromosome 5 of MERE-A-13. Colours correspond to marked duplications on the dot-plot. **C**. Distribution of *CEN178* duplication sizes across the 5 chromosomes. Figures **D** to **F** show theoretical HOR expansion patterns following a duplication, with representative examples from the *Arabidopsis* dataset. In the upper panels, coloured boxes on the diagonal represent *CEN178* repeat blocks. Green blocks are repeats that formed the duplication. Following that, local HOR expansions can occur on one of the duplicated segments; including (**D**) Both segments in proximal positions, (**E**) and both segments in distal positions (**F**) and are marked in orange. Grey boxes represent other interspersed repeat blocks. Similar blocks after the HOR expansions are represented as yellow boxes in a dot-plot manner. Since these correspond to individual HORs, the underlying expansion patterns can be inferred from the arrangement of the HOR block start positions, visualised on a dot-plot. Lower panels are examples of the identified duplications confirming to the theoretical scenarios presented above, with plot elements as in **A**. Other possible scenarios following a duplication would adhere to these rules.

Plotting the HOR block start coordinates revealed patterns of large duplications present within some of the *CEN178* arrays. For example, chromosome 5 of MERE-A-13 accession contains distant HORs indicative of three large duplications, two of them in tandem (Fig. 4.14A). HOR identification using more stringent settings (--t 0 –c 5), effectively only identifying 5 repeats long HORs with identical blocks, was found to sensitively detect these large duplications (Fig. 4.14A). Since a large duplication produces a series of HORs that are around the same region, and have similar distances between two blocks, an identification method was developed that recognised duplication patterns based on the histograms of these distances (Fig. 4.14B). A total of 94 duplications were found with a mean length of 367 kbp (Fig. 4.14C). Considering that HORs tend to be short and local, long duplications might arise from a different mechanism to the HOR duplications.

Observed duplications unveiled patterns of differential HOR accumulations post duplication (Fig 4.14D-F). This manifests as deviations from a straight, parallel to the diagonal line (representing self-similarity) on a dot-plot and can inform about the relative position of HOR expansion in the duplicated region since its formation. For

example, after a duplication on chromosome 1 of Etna-2 accession, more HORs accumulated in the downstream duplicated array, which caused the duplication line on the HOR dot-plot to rotate by around -12º relative to the diagonal (Fig. 4.14D). Other examples include Rabacal-1 chromosome 2 and IP-Per-0 chromosome 3 that also show deviation of the duplication line due to uneven HOR accumulation (Fig. 4.14E and F).

Since the developed duplication identification method proved to be robust within individual chromosomes, it was applied to analyse regional similarity across the chromosomes. To make this process computationally viable (analysing 54,285 chromosome pairs would be too computationally expensive), only chromosome pairs that had an SRS score (described before) of at least 5% (3,400 pairs) were considered. When two centromeres are identical, the HORs that are identified between them should overlap fully with the ones identified within each of them individually (Fig. 4.15A). Therefore, identified *CEN178* HORs were used to quantify similarity between regions of a pair of centromeres by measuring the ratio of the number of identified HORs between chromosomes against the number of HORs within a chromosome using a sliding window (Fig. 4.15B). These centromeric synteny-based similarity (SBS) scores were calculated individually for both chromosomes from each pair (Fig. 4.15B). To investigate this on a larger scale, SBS scores were averaged across centromere lengths and plotted (Fig. 4.15C). These scores were found to rise with increasing distance from the centre of the *CEN178* arrays, suggesting that distal regions do not accumulate novel HORs and maintain more ancestral similarity, which might be the reason for them having higher edit distances from the consensus (Fig 4.13C). Despite that, the very ends of the chromosomes show drops in the SBS scores, likely due to these regions containing the most diverged repeats where accumulated mutations disallow accurate identification (Fig 4.13C).

Identified HORs between the chromosomes can be also plotted on a 2-dimensional dot plot, or as connecting points between linear representations of the chromosomes. For example, Met-6 vs CAMA-C-9 chromosome 3 shows that an expanded region in the first accession also contained the highest levels of HORs within that chromosome (Fig 4.15D). Also, two regions present on the CAMA-C-9 centromere are missing from

the former (Fig. 4.15D). They do not appear to be novel array expansions since there is no visible increase in HORs in CAMA-C-9. Potentially, the expansion of the MET-6 central array caused compensatory deletions to maintain centromere repeat array size. It would be interesting to quantify the CENH3 profiles in these accessions to measure whether the expansion on Met-6 is following a change in CENH3 enrichment within the tandem repeat arrays.

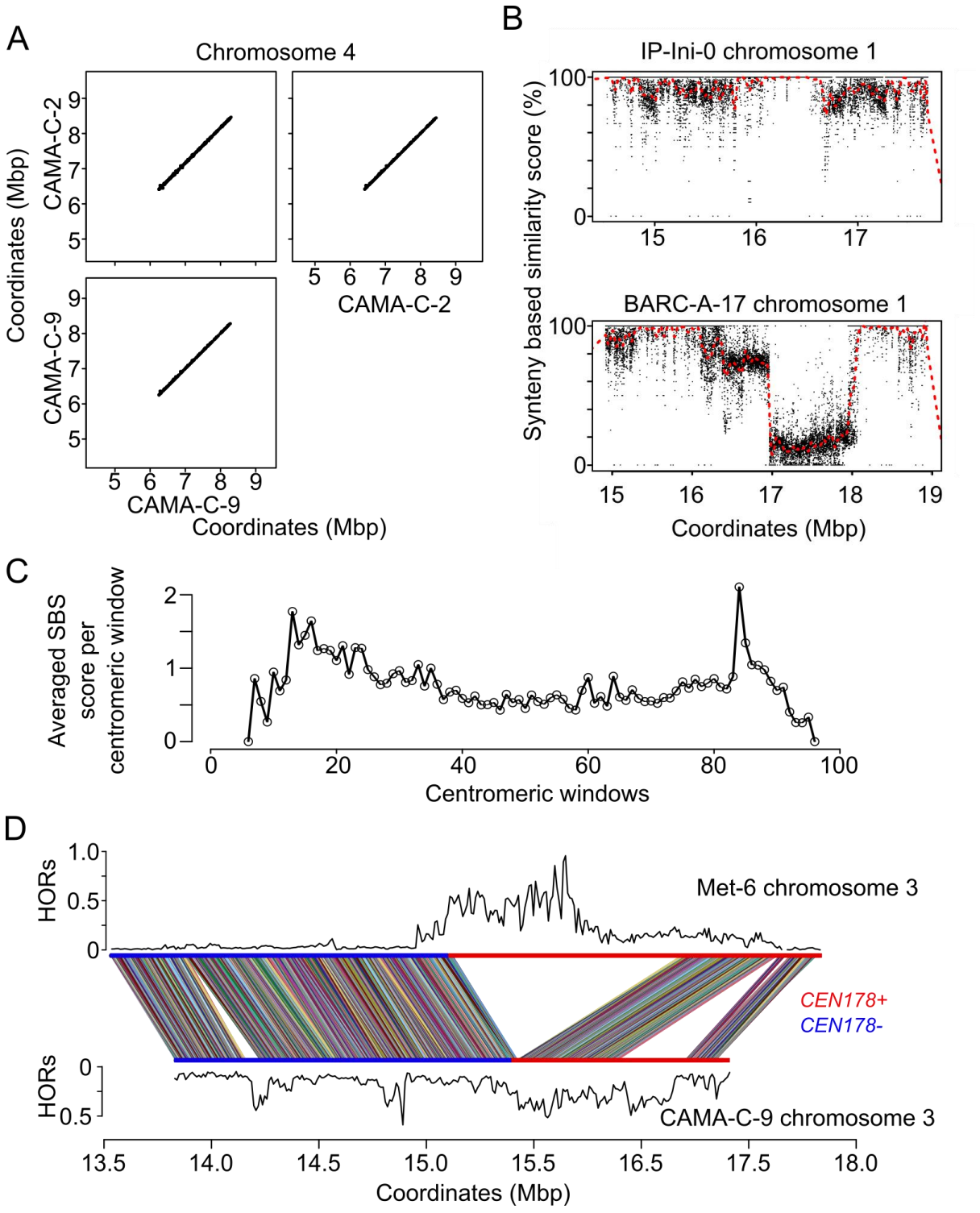


Figure 4.15. Large duplications of *CEN178* segments and centromere synteny-based scores.
A. *CEN178* HORs identified within chromosomes 4 of CAMA-C-2 and CAMA-C-9 and between them, using stringent (--t 0 –c 6) settings are shown. The number of identified *CEN178* HORs between the chromosomes is almost identical to the number of HORs identified within individual chromosomes. **B**. Pairwise comparison of chromosome 1 from IP-Ini-0 and BARC-A-17. The x axis provides information on repeat start site and the y axis shows the ratio of HORs a repeat is involved in, when comparing inter-chromosomally to intra-chromosomally. The red dotted line is a moving average of the y axis values. Consecutive values close to 1 indicate that the region can be found within the other chromosome, while those close to 0 indicate private repeats relative to the other chromosome. **C**. The distribution of the SBS scores averaged across all centromeres in 100 windows. Due to the centromere region definition including flanking non-repetitive sequence, not all 100 windows had any synteny score assigned. **D**. A synteny plot of two centromere 3 from the Met-6 and CAMA-C-9 accessions. Repeats and their strand are plotted along the x axis for both chromosomes (red for plus and blue for minus strand). *CEN178* HORs identified with stringent settings connect the chromosome maps according to the start positions of their duplication blocks. Above and below in black are repetitiveness plots to indicate regions of active HOR expansion.

While *CEN178* are the functional sequence elements of the *Arabidopsis* centromeres, as they are the CENH3 deposition sites, the function of *CEN159* repeats is unknown since their discovery (Simoens et al,1988, Bauwens et al, 1991). *CEN159* HOR abundance was low compared to the *CEN178* arrays (3.45% vs 5.67%, on average), with a total number of 47,674 identified, compared to a total of 98,265,545 *CEN178* HORs (Fig. 4.16A). 85 out of 294 chromosomes containing *CEN159* arrays did not contain any *CEN159* HORs. Two chromosomes had higher scores than any chromosome using *CEN178* HORs (above 20%), but they only contained 13 and 4 repeats respectively, meaning single identified HORs increased their abundance score significantly. The distribution of HORs across the centromeres show distal maxima, as expected from the localisation of *CEN159* repeats outside of the *CEN178* arrays (Fig. 4.16B). Fine scale analysis of HOR positions is hindered by the small number of *CEN159* repeats and their HORs.

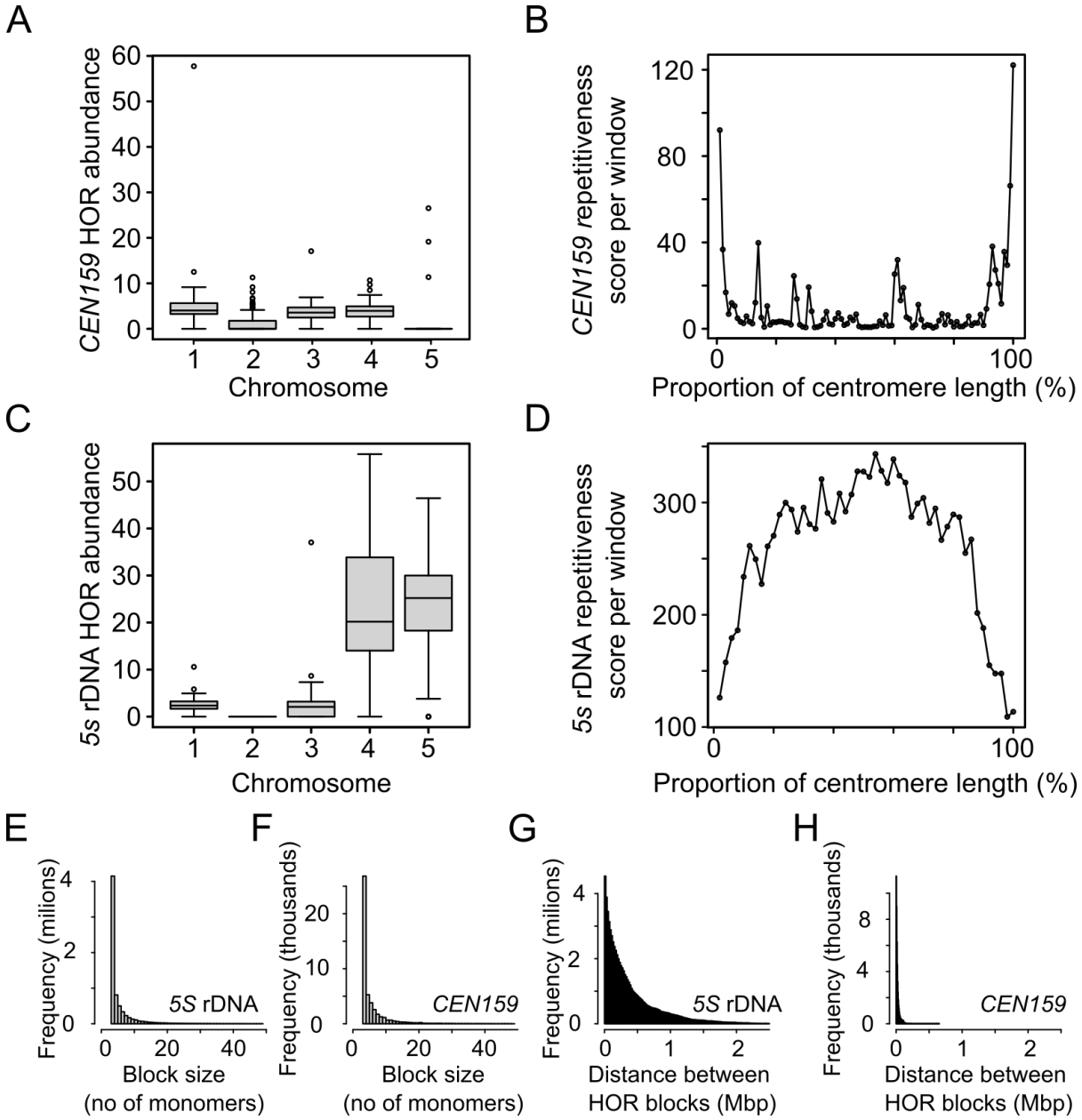


Figure 4.16. HOR analysis of *CEN159* and *5S* rDNA tandem repeats.
A. *CEN159* HOR abundance score distribution for all chromosomes. **B**. *CEN159* repetitiveness distribution across the centromeres (windows identical to those on Fig. 4.13D). Due to the small number of *CEN159* repeats and their scattered nature, averaging across individual arrays would not be informative. **C.** *5S* rDNA HOR abundance score distribution for all chromosomes. **D.** *5S* rDNA repetitiveness distribution averaged across all individual arrays that contained at least 100 individual repeats and did not have larger gaps between them than 1 kbp (*n* = 78). **E**. Histogram of the *5s* rDNA block sizes (in monomers). **F**. Histogram of the *CEN159* block sizes (in monomers). **G**. Histogram of the *5s* rDNA distances between HOR blocks (in Mbp). **H**. Histogram of the *CEN159* distances between HOR blocks (in Mbp).

An equivalent analysis was performed for the *5S* rDNA repeats. These repeats can provide a valuable comparison to the centromeric *CEN178,* since they have biologically defined functionality and form tandem arrays across higher eukaryotes and can be subject to copy number fluctuation between individuals and even between cells (Cloix et al, 2002, Xu B et al, 2017, Ding et al, 2022). 6,992,084 *5S* rDNA HORs were identified with mean chromosome HOR abundance of 16.05% per chromosome that contained any repeats, which indicates the highest homogeneity of all analysed repeats (Fig. 4.16C and D). The distribution of HORs across the *5S* rDNA arrays was similar to those seen for *CEN178*, with maximum values being reached in the central parts of *5S* rDNA arrays (Fig 4.16D). *5s* rDNA HORs also showed a higher number of instances with high distance between the blocks, highlighting their uniformity (Fig 4.16G).

Outside of the *CEN178* repeats, the only major repeat element in the *Arabidopsis thaliana* centromeric sequences are *ATHILA* transposable elements (Naish et al, 2021). Their positions, including the positions of their Target Site Duplications (TSDs), and their LTR percentage identity scores (PID), were identified and calculated by Dr Alexandros Bousios from the University of Sussex. A total of 9,250 intact and 13,556 soloLTRs across the chromosomes were identified, with 1,357 intact and 549 soloLTRs interspersed in the centromeric repeat arrays. The ratio of intact to soloLTR elements inside the centromeres was higher than outside (2.5 vs 0.6), which indicates increased centromeric *ATHILA* integration, reduced soloLTR formation, or more efficient removal of the soloLTRs inside the centromeres. Centromeric *ATHILA* were also significantly younger than those in the chromosome arms, based on their internal LTR PID (98.97% vs 94.35%) (Wilcoxon test $P<1.57\times10^{-8}$). A significant negative correlation between the number of *ATHILA* per centromere vs *CEN178* HOR abundance score was found (Wilcoxon test P<2.2×10-16) (Fig. 4.17A), suggesting that repeat homogenisation processes remove the transposons from the centromeres.

Distribution of *ATHILA* and HOR abundance across averaged centromeric bins was plotted and *ATHILA* number anticorrelated with averaged HOR abundances (Wilcoxon test $P<2.2\times10^{-16}$) (Fig. 4.17B). This is expected, as a high number of TEs have been

reported in the pericentromeric regions (Wright, Agrawal and Bureau 2003, Chang et al. 2022), and distal parts of the analysed regions contain less repeats than central ones (Fig. 4.17B). To analyse the distribution within *CEN178* arrays, only *ATHILA* surrounded by repeats were considered and the anticorrelation found previously was no longer true. For individual chromosomes, the distribution of centromeric *ATHILA* significantly anticorrelated with *CEN178* HOR abundance scores only for chromosome 3 (Wilcoxon test $P=4.70\times10^{-5}$) (Fig. 4.17C-F). Overall, despite chromosomes with more HORs having less *ATHILA* elements, the distribution of centromeric *ATHILA* does not create obvious patterns when averaged across the centromeres. *ATHILA* integration might not be spatially correlated with HORs, but on the population scale, chromosomes with more HORs are populated by less *ATHILA*, which might be caused by the purging effect of the *CEN178* expansion.

The integration site of centromeric *ATHILA* was mapped along the *CEN178* consensus (Fig. 4.17B). 20 bp sequences upstream and downstream of each *ATHILA* element were extracted and mapped to the *CEN178* consensus sequence using BBmap with settings: maxindel=16,000 ambiguous=random k=8 saa=f vslow=t settings (Bushnell sourceforge.net/projects/bbmap/). 677 elements were excluded when one or both sequences were not able to map to the consensus. *CEN178*-relative insertion positions were calculated in both upstream and downstream configurations. Out of the remaining 1,229 *ATHILA*, 1,154 of them had a 4 bp overlap in the mapped region, indicating the presence of the TSD sequences (Fig 4.17C). This indicated that the mapped insertion sites had single-base pair precision. *ATHILA* insertions were observed throughout the *CEN178* consensus, with some local stacking of values.

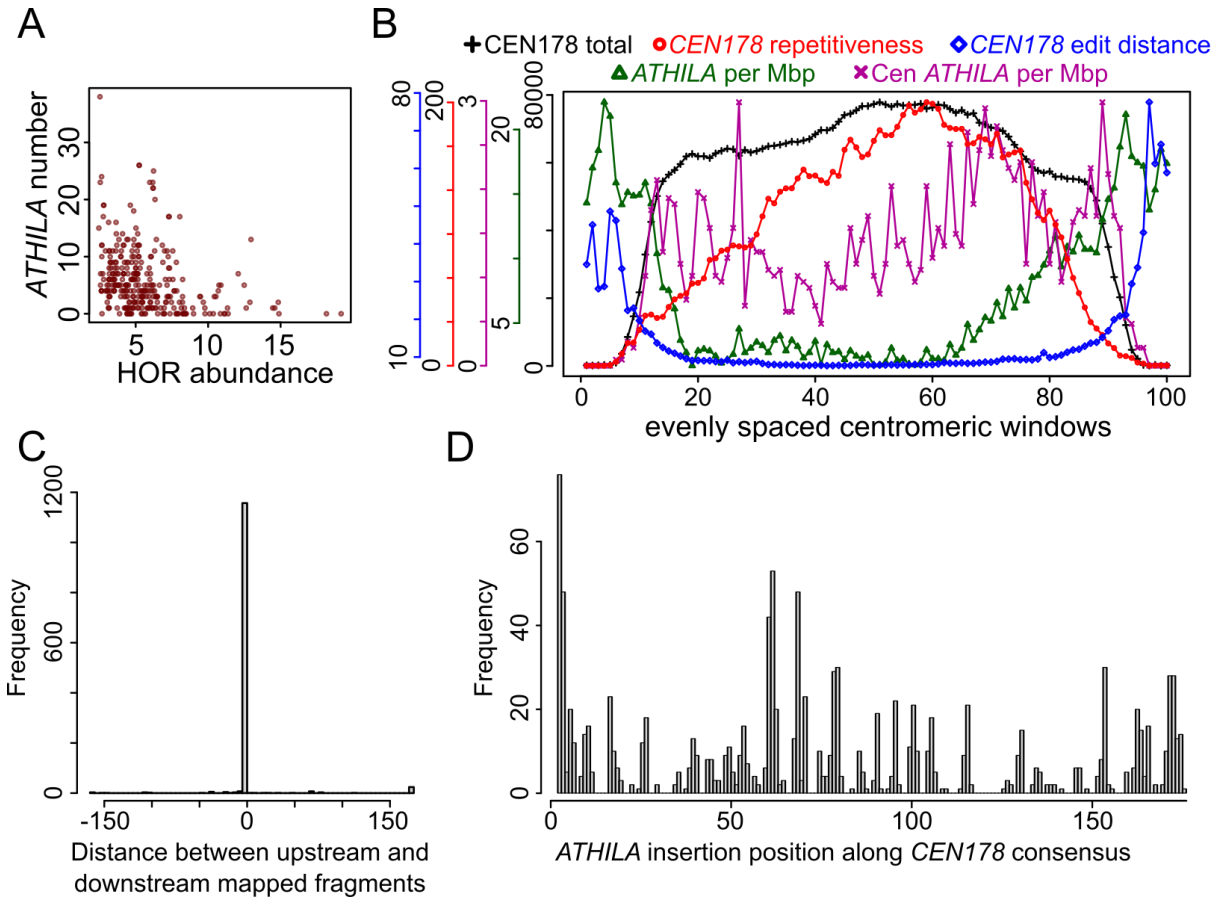


Figure 4.17. *ATHILA* insertion sites within *CEN178* satellite repeats.
A. A histogram of distances between mapped TSD-containing *ATHILA* insertions. The -4 bp shift corresponds to the DNA duplication that forms during TSD formation at integration. **B.** *ATHILA* insertion site locations across the *CEN178* consensus. **C**. Frequency of distances between the 20-bp ATHILA flanking regions after mapping to the CEN178 consensus sequence. -4 bp means a 4 bp overlap between the two alignments. This is equivalent to the overlap created by the TSD, confirming accurate insertion mapping for these elements. **D**. Insertion sites of *ATHILA* along the *CEN178* consensus.

The main function of the centromeres is to serve as a site for the deposition of CENH3/CENP-A histone variants, which are necessary for kinetochore assembly and binding to spindle microtubules during cell division (Hori et al. 2012). When analysed, the distribution of CENH3/CENPA within centromere regions is not uniform (Bodor et al. 2014). Neither is it necessarily constrained to the centromeric repeats in species that contain them (Cappelletti et al. 2019). There is no consensus whether the underlying DNA sequence determines the CENH3 deposition, or not (Talbert and Henikoff 2020). Neocentromere formation has been described in various species (for

example human: Murillo-Pineda et al. 2021, zebra: Cappelletti et al. 2022, *Drosophila*: Williams et al. 1998, maize: Yu et al. 1997), where CENP-A enrichment can be observed over repetitive or non-repetitive genomic regions when the endogenous satellite repeats are deleted.

CENH3 ChIP-seq experiments in the *Arabidopsis thaliana* Col-0, Cvi-0, Ler-0 and Tanz-1 accessions were performed by Dr Matthew Naish and compared to the distribution of *CEN178* HORs across the centromeres (Fig. 4.18). While CENH3 levels seem consistent in terms of their distribution across the chromosomes and overall abundance, *CEN178* HORs tend to vary in their counts, suggesting an uneven rate of centromere expansion in the history of the centromere. HOR peaks overlap with CENH3 enriched regions, apart from the Tanz chromosome 5, where a HOR-rich array, located around 12.5 Mbp, is positioned away from the CENH3-enriched array that is centred around 16 Mbp (Fig 4.18). Whether CENH3 is recruited to satellite arrays that are undergoing active expansion, or whether the expansion is a result of CENH3 deposition is unknown.

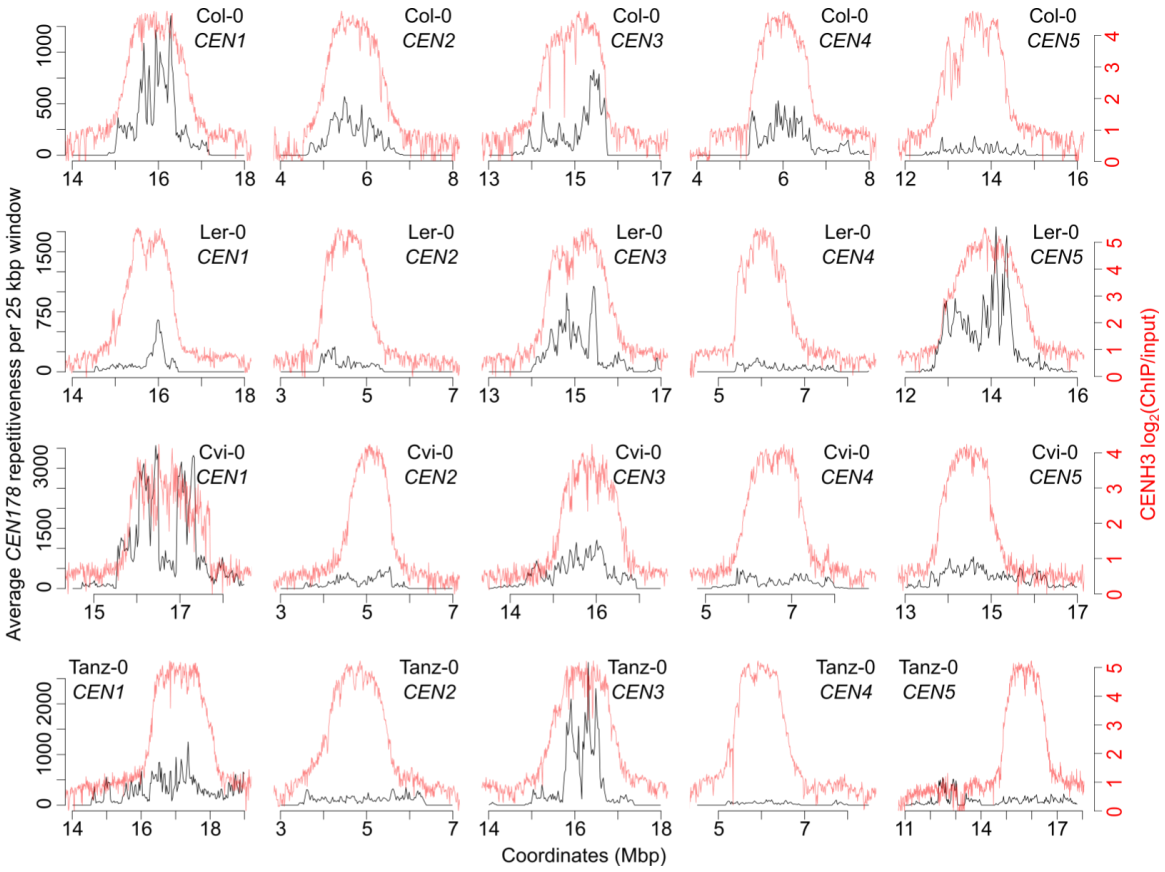


Figure 4.18. Centromere satellite higher order repeats in relation to CENH3 enrichment in Col-0, Ler-0, Cvi-0, and Tanz-1.
CEN178 HOR repetitiveness (black) and CENH3 $\log_2$(ChIP/control) ratios (red) averaged over 10 kbp regions across five centromeres of *A. thaliana* accessions Col-0, Ler-0, Cvi-0, and Tanz-1.

4.1.3 Two closely related repeat families in *Arabidopsis lyrata* centromeres

Arabidopsis lyrata is a close relative to *Arabidopsis thaliana*, which is estimated to have diverged ~5 million years ago and has been used to study genome evolution (Hu et al. 2011). Both Arabidopsis genomes are syntenic with the majority of gene pairs being in highly conserved collinear arrangements (Hu et al. 2011) (Fig. 4.19A). The pericentromeric regions of 8 *A. lyrata* chromosomes can be found across the *A. thaliana* genome, but the centromeres themselves do not contain the same families of repeats (Berr et al. 2006). Gapless genomic assemblies of two *Arabidopsis lyrata* accessions were kindly provided by Dr. Polina Yu. Novikova from The Max Planck Institute for Plant Breeding Research. The MN47 accession from North America, and

NT1 from Siberia represent subspecific lineages that diverged ~130,000 years ago and can be used to assess intra- and inter-specific centromere divergence (Mable et al. 2005, Foxe et al. 2010, Hu et al. 2011). Two main centromeric repeat families were identified to occupy centromeres of both accessions, 168 bp *CEN168* and 179 bp *CEN179* (Fig. 4.19B and C). Different centromeres were predominantly composed of a single repeat family: *CEN168* dominated *CEN2*, *CEN5*, *CEN6* and *CEN8*, while *CEN179* dominated *CEN1*, *CEN3*, *CEN4* and *CEN7* (Fig. 4.19B and C). Both repeats were similarly enriched across the accessions but displayed variation in their copy number across individual chromosomes (Fig. 4.19D-F). Sequence similarity analysis of *A. thaliana CEN159* and *CEN178*, and *A. lyrata CEN168* and *CEN179* revealed significant similarity between both *A. lyrata* repeat families and *CEN178*, but not *CEN159* (permutation test, $P<0.01$, Table 4.3). CEN168 and CEN179 difference in size can be mainly attributed to a 10-bp insertion/deletion found around position 80 (Fig. 4.19G). Similarity between *A. thaliana CEN178* and both *A. lyrata* repeat families together with colinearity of centromeres found in both species suggests their common ancestry. Two possible scenarios would have occurred: (i) *CEN168* and *CEN179* diverged before *A. thaliana* and *A. lyrata* speciated and *A. thaliana* centromeres removed (likely) *CEN168* from its centromeres, or (ii) *CEN168* and *CEN179* diverged after the speciation within the *A. lyrata* lineage. A deeper study of *A. lyrata* accessions and other *Arabidopsis* genus species could probably answer this question.

Both of the *A. lyrata* centromeric repeat families occupy their respective centromeres with varying numbers (Fig. 4.19D) and levels of HOR accumulation (Fig. 4.19H and I). Additionally, MN47 and NT1 accessions were collected in distant geographic locations: North America and Siberia, respectively. Based on the results from *A. thaliana*, these features are depictive of more divergent centromeres, which suggests that the existence of both repeat families is not transitory. This raises questions about the maintenance of two different centromeric repeats that seem to be established.

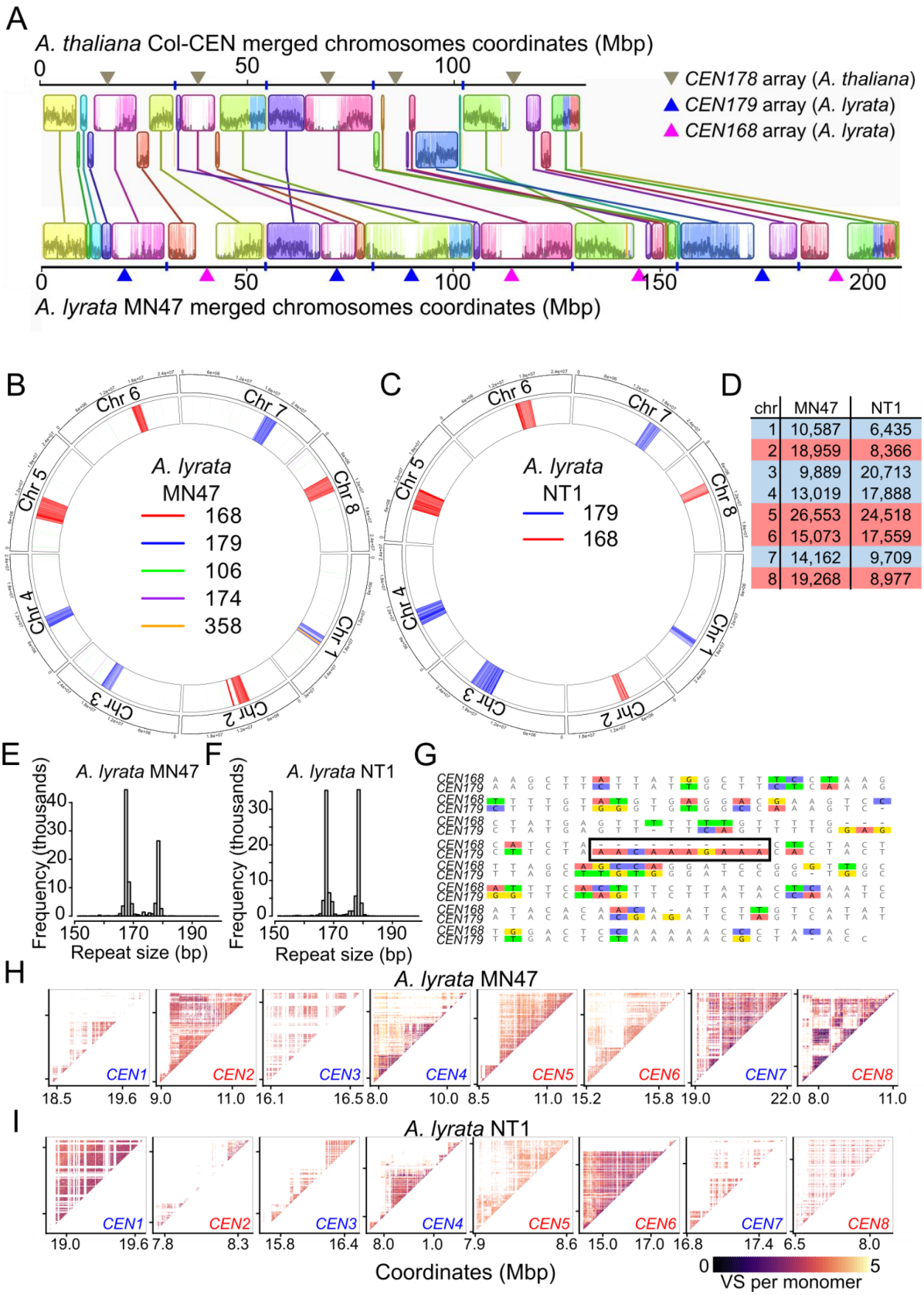


Figure 4.19. *Arabidopsis lyrata* centromeric structure

A. Synteny analysis of *A. lyrata* MN47 and *A. thaliana* Col-CEN. All chromosomes of Col-CEN and MN47 merged and aligned using Mauve (Darling et al. 2004). Blue ticks

along coordinate axes represent transitions between individual chromosomes. Triangles are positions of centromeric arrays. Coloured boxes connected by lines are synteny blocks found on both sequences. Coloured triangles represent centromeric repeat arrays. **B** and **C.** Circos plot of the tandem repeats found across two *A. lyrata* accessions. **D** Tally of the *CEN179* (blue fill) and *CEN168* (red fill) repeats across the 2 accessions. **E** and **F**. Histograms of *A. lyrata* repeat sizes in MN47 and NT1 accessions. Two peaks correspond to the *CEN168* and *CEN179* repeat families. **G**. Pairwise alignment between CEN168 and CEN179 with disagreements highlighted and the 10-bp insertion/deletion marked by a black box. **H** and **I**. HOR plots of both accessions' centromeric regions, points representing HOR blocks start positions are coloured according to the variant score (VS) per monomer.

4.1.4 Centromeric satellite repeats expansions in two *Brassica oleracea* accessions

The Brassicaceae family includes approximately 3,800 species, including commercially important vegetable, fodder, oilseed, and ornamental crops (He at al. 2021). The *Brassica* genus-wide whole genome triplication event occurred approximately 22.5 million years ago, and its descendants contain one of the three subgenomes: A, B, C or a combination of these, for example: *B. rapa* (AA), *B. nigra* (BB), *B. oleracea* (CC), *B. juncea* (AABB) (Nagaharu 1935, Paritosh et al. 2021, He at al. 2021). The collinearity of the three sub genomes was recently defined using orthologous genes analysis, which allowed reconstructing the 7-chromosome ancestral genome (He et al. 2021). However, the evolution and linearity of the centromeric regions is poorly understood, as analysis of the centromeric sequence is restricted to individual species (Lim 2007, Zhang 2018, Rosseau-Gueutin 2020). Centromeric repeat families across the *Brassica* species include 176 bp long *CentBr1* and *CentBr2* from *B. rapa* (Lim et al 2005, Lim et al 2006), 177 bp *CentBo1* and *CentBo2* from *B. oleracea* (Waminal 2021) and an unnamed, 176 bp repeat from B. napus (Chen 2021). *CentBr1* and *CentBr2* probes were used in a FISH experiment using species representing A, B and C subgenomes and were found to hybridise to at least some of the chromosomes in all species, suggesting partial similarity between these centromeric repeat families (Koo 2011).

Mapping and comparison of the *Brassica* centromeric repeats has the potential to uncover structural variation, facilitate ancestral genome reconstruction, and highlight breeding potential stemming from hybrid centromere instability (Metcalfe et al, 2007, He et al, 2021, Ning Guo et al, 2021, Boideau et al, 2022). To analyse *Brassica* centromere structures and to compare with Arabidopsis species, two genomic assemblies of *Brassica oleracea* (CC genome) ssp. *Alboglabra* and *Brassica oleracea* ssp. *Italica* were kindly provided by Prof. Jose Gutierrez-Marcos from the University of Warwick.

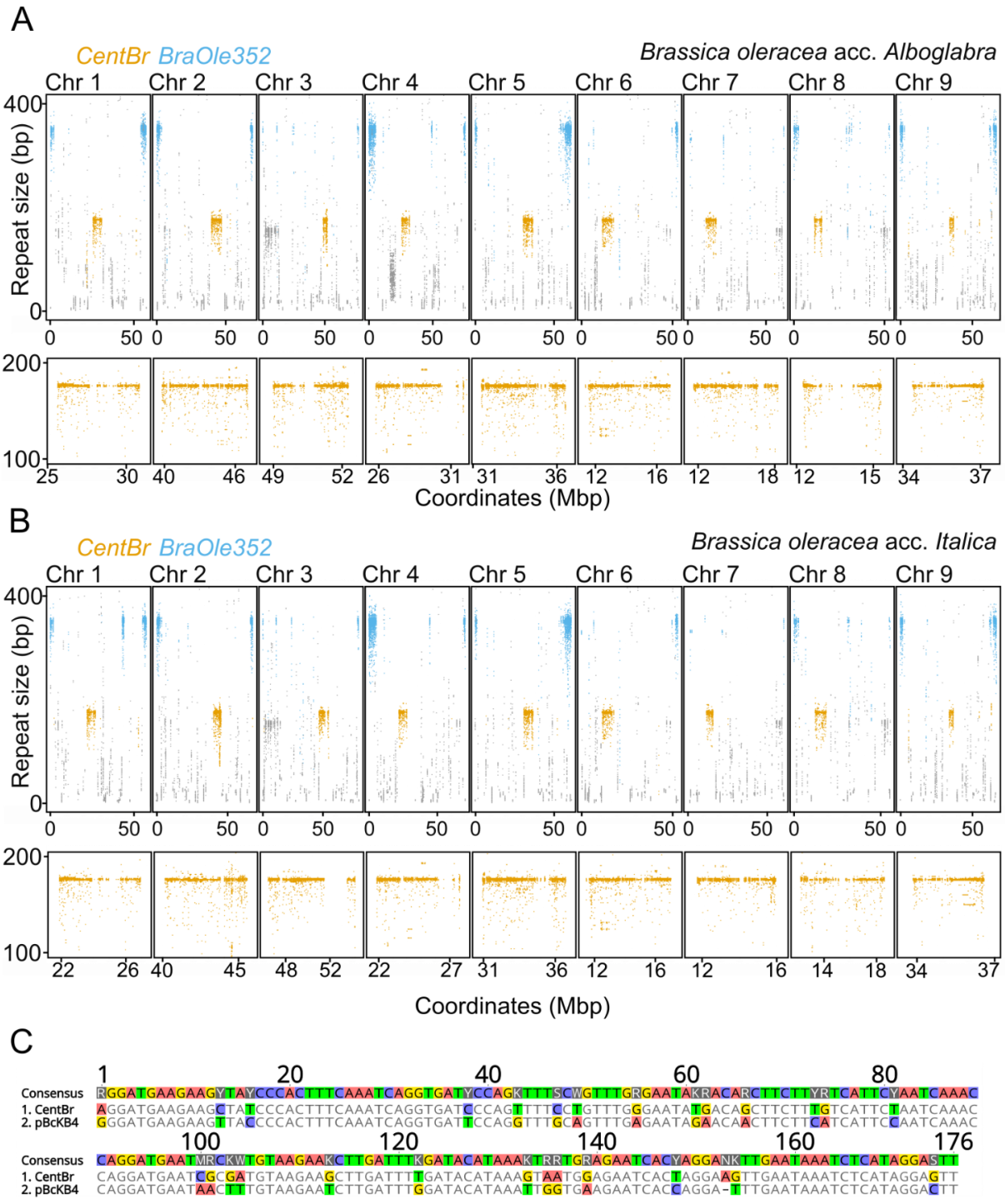


Figure 4.20. Tandem repeats of two *Brassica oleracea* subspecies: *Alboglabra* and *Italica*.

A. All *Brassica oleracea* ssp. *Alboglabra* identified tandem repeat sizes are plotted against their start coordinates for all 9 chromosomes (above), with centromeric *CentBr* arrays zoomed in below. *CentBr* repeats are in orange and *BraOle352* repeats are in blue. **B**. Same as **A**, but for *Brassica oleracea* ssp. *Italica*. **C**. Sequence alignment between the *CentBr* consensus identified by TRASH and previously published and

cytologically confirmed *CentBr* sequence derived from pBcKB4 BAC (PID = 84.1%) (Lim et al. 2005).

TRASH analysis using default settings identified two main tandem repeat families of 176 (155,720 in *Alboglabra* and 165,388 in *Italica*) and 352 bp (17,947 in *Alboglabra* and 18,211 in *Italica*) in the *B. oleracea* genome assemblies (Fig. 4.20A-B, Table 4.3). The 176 base pair tandem repeat family was confirmed to correspond to the previously described centromere satellite *CentBr* family (Fig. 4.20C) (Lim et al. 2005). At the sequence level, *CentBr* and *CEN178* alignment had a percentage identity (PID) score of 51.46%, which despite its significance compared with 500 permutations at $P<0.01$, indicates high divergence of these repeats (Table 4.3).

		A. thaliana		*B. Oleracea*		*A. lyrata*	
		CEN159	*CEN178*	*CentBr*	*BraOle352*	*CEN179*	*CEN168*
A. thaliana	*CEN159*	100.0	49.6	49.6	48.3	50.0	49.1
	CEN178	50.2	100.0	51.5	49.8	80.8	65.7
B. Oleracea	*CentBr*	50.6	49.7	100.0	48.4	47.9	48.8
	BraOle352	51.5	50.1	50.8	100.0	47.7	48.9
A. lyrata	*CEN179*	49.7	49.1	49.9	49.3	100.0	73.0
	CEN168	49.9	48.7	49.6	49.2	48.5	100.0

Table 4.3. Percentage identity (PID) levels of tandem repeats identified in the Brassicaceae.
Bottom-left part (grey shading) are 99th percentile scores coming from 500 permutations of alignments between the query sequence (row) and randomly shuffled subject sequences (column). The PID scores are shown in the top-right part of the matrix, with scores above the 99th percentile coloured green.

CentBr repeats have been reported to consist of two subfamilies, named *CentBr1* and *CentBr2*, with *CentBr1* occupying all chromosomes in *Brassica rapa* and *CentBr2* partially occupying 5 of the 9 chromosomes (Lim et al, 2005). To evaluate the similarity within *B. oleracea* repeats and to find putative sub-families, I extracted 1,000 *CentBr* repeats from both assemblies, aligned them using mafft (--auto) (Katoh et al, 2013) and created phylogenetic trees using Geneious (Jukes-Cantor model, Neighbour-Joining method, 10 times Bootstrap, Geneious 2022.2.1) (Fig 4.21A). Both trees contain a substantial cluster of 614 (*Alboglabra*) and 627 (*Italica*) repeats

(coloured red on Fig 4.21A), which after extraction have higher average pairwise identity (94.1% and 94.0%) than the remaining repeats (87.6% and 88.6%). The PID score between the consensus of the clustered repeats and the published *CentBr1* probe sequence (GenBank: CW978699.1) was 97.2%, and the PID score with *CentBr2* (GenBank: CW978837.1) was 88.8%. The PID of the consensus of the remaining repeats to the *CentBr1* consensus was 87.7% and to the CentBr2 consensus was 94.2%. Thus, the first cluster corresponds to *CentBr1* and the second cluster to *CentBr2*. The chromosome distribution of the *CentBr1* repeats is distinct from the remaining repeats, as they occupy the central regions of *CEN1*, *CEN2*, *CEN4*, *CEN5*, *CEN6* and *CEN7* in *Alboglabra* and *CEN1*, *CEN2*, *CEN3*, *CEN4*, *CEN5*, *CEN6* and *CEN7* in *Italica* (Fig. 4.21B). Only *CEN3* was different between the two subspecies with *Alboglabra* consisting mostly of *CentBr2* and *Italica* of *CentBr1*, with some *CentBr2* located towards the edges of the centromere (Fig. 4.21B).

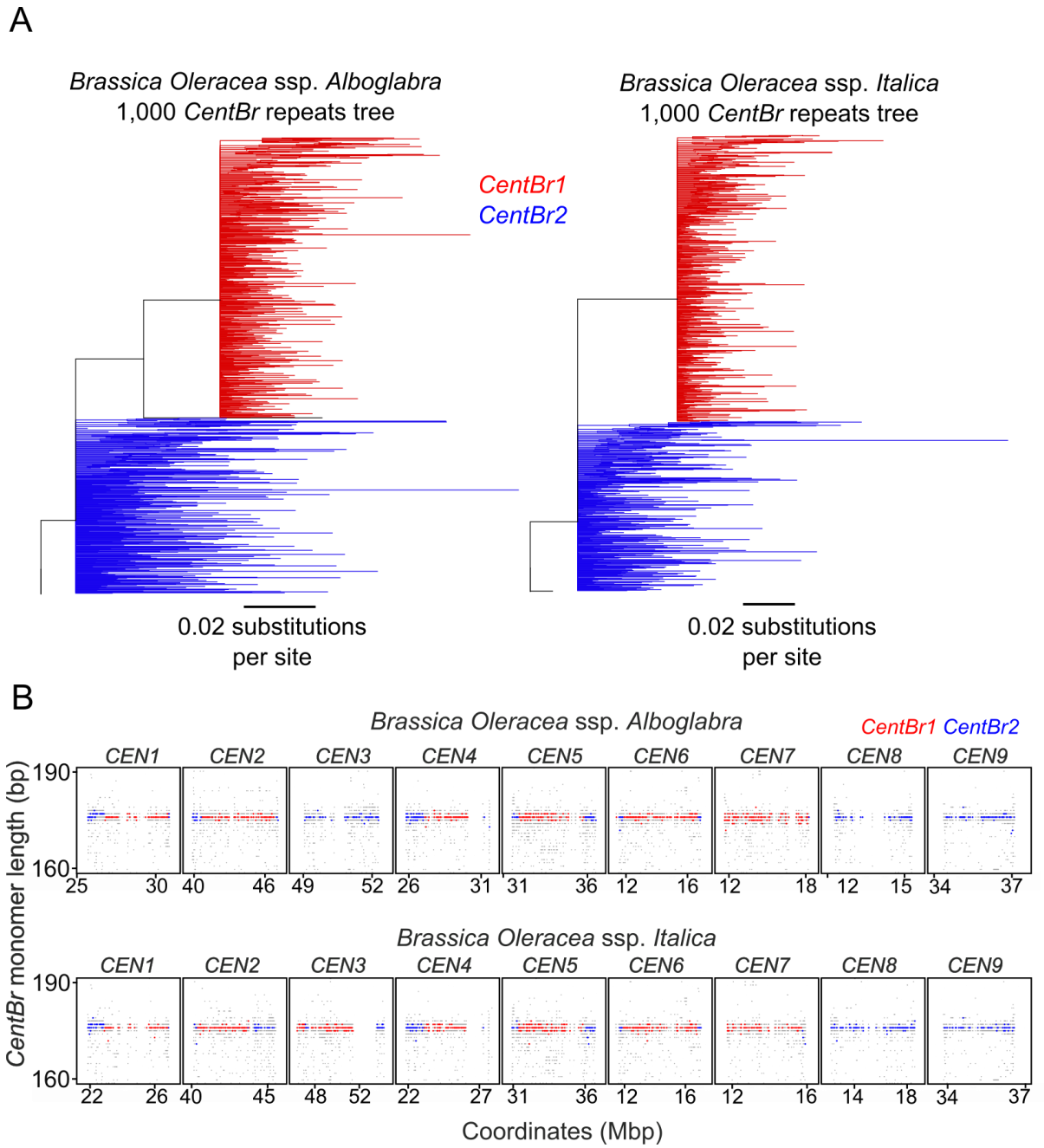


Figure 4.21. Distinct clusters of *CentBr* repeats occupy central parts of most centromeres.

A. Phylogenetic trees of 1,000 random *CentBr* repeats of *B. oleracea* ssp. *Alboglabra* (left) and *Italica* (right). In red, are branches of *CentBr1* repeats clustering with 94.1% and 94.0% average PID, compared to 87.6% and 88.6% for whole trees. In blue, are the *CentBr2* repeats. **B**. Plots of the centromeric regions of *Alboglabra* (above) and *Italica* (below) with repeats between 160 and 190 bp presented (grey). Repeats that were extracted for the phylogenetic tree constructions are coloured as for the tree with *CentBr1* in red and *CentBr2* in blue.

To estimate the similarity between the centromeres of both subspecies, the number of identical *CentBr* repeats from each chromosome found in all other chromosomes was calculated (Fig. 4.22A). The number of shared repeats between chromosomes of individual genomes was low, although higher than the same metric in *Arabidopsis thaliana* (Table 4.1). Interestingly, only chromosomes that contained *CentBr1* repeats shared more than 33 repeats between them, while the remaining chromosomes showed essentially private repeat libraries (Fig. 4.22A). Chromosome pairs of both subspecies shared over 99% of their repeats in *CEN4* and *CEN9* pairs, and these centromeres also appeared identical in dot plot sequence similarity analysis (Fig. 4.22B). While most respective chromosomes within both subspecies show similar patterns of similarity, *CEN3* in *Alboglabra* matched only a few repeats in *CEN3* in *Italica*, while the latter had higher levels of similarity with *CEN1-CEN7* within the subspecies (Fig. 4.22A). The relative lack of shared repeats with other chromosomes within the subspecies can be also observed for *CEN8* and *CEN9*. Together with *Alboglabra CEN3* these centromeres also do not contain *CentBr1* (Fig. 4.21B). Cytological data from *B. oleracea* ssp. *TO1000* using *CentBr1* and *CentBr2* probes suggested that centromeres 3, 4, 8 and 9 are devoid of *CentBr1* (Xiong and Pires 2011). Since *CentBr1* is actually present on *CEN4* of *Alboglabra*, and *CEN3* and *CEN4* of Italica, intra-species differences in centromere architecture may be one explanation. All chromosomes contained *CentBr2* probes (Xiong and Pires 2011), although they tended to be located on the centromere periphery, in agreement to the *CentBr2* map (Fig. 4.21B).

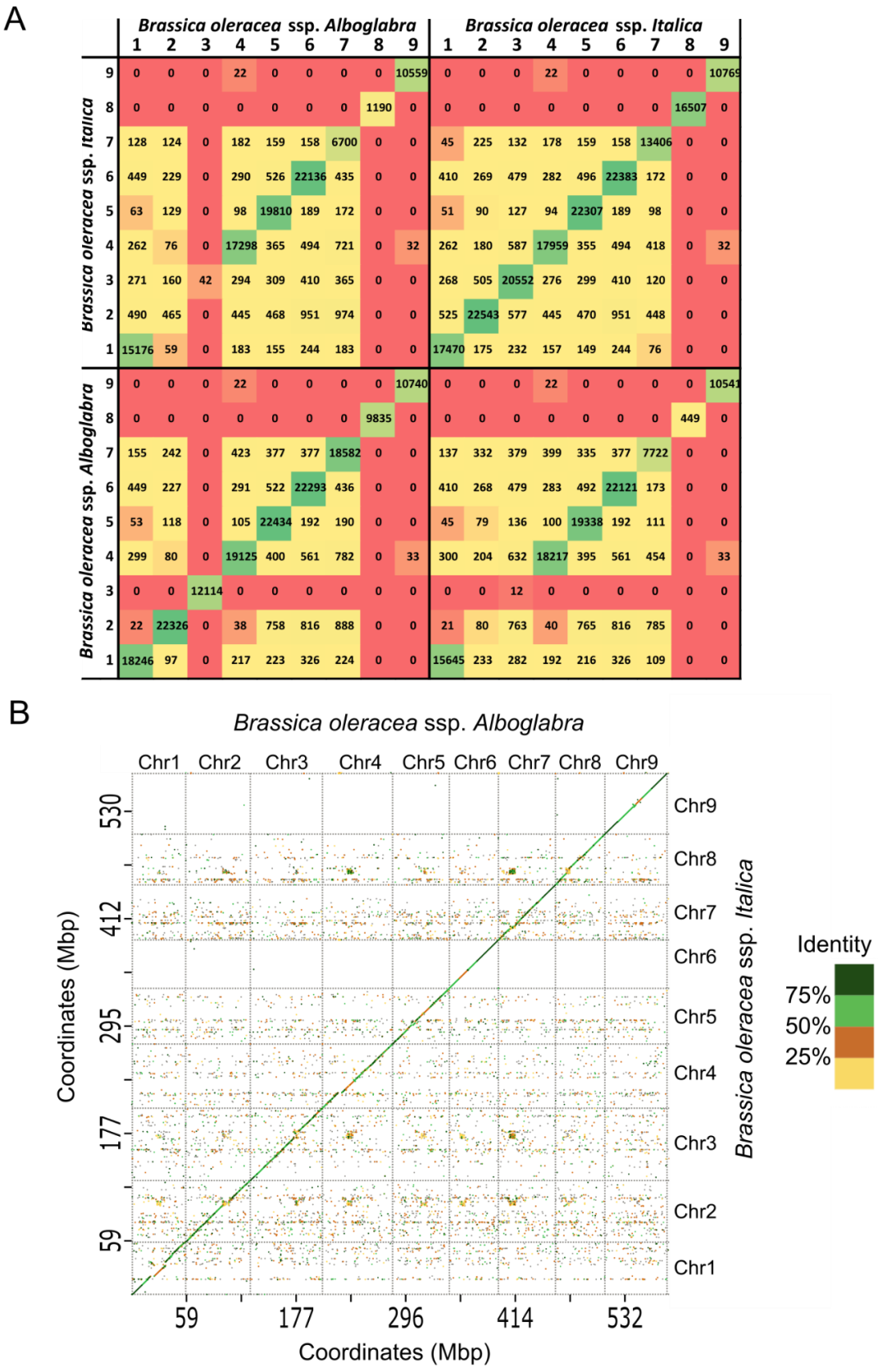


Figure 4.22. Shared *CentBr* tandem repeat counts between chromosomes of *B. oleracea* and dot plot of the genomic DNA.

A. The number of *CentBr* repeats from the column-described chromosome found within the row-described chromosome. Since all repeats from a chromosome can be found within itself, numbers on the diagonal represent the total number of repeats for that chromosome. **B**. Sequence similarity dot plot of *Brassica oleracea* ssp. *Alboglabra* (x) and *Italica* (y) made using D-GENIES online tool (Cabanettes and Klopp 2018)

Centromeric region start, and end coordinates were defined for each chromosome as the 1st and 99th percentile of *CentBr* start coordinates. The centromeres defined in this way span a range between 2.69 and 7.70 Mbp, with an average of 5.21 Mbp, which is almost double the average size of *Arabidopsis thaliana* Col-CEN centromeres calculated with the same method (2.70 Mbp on average). Despite this, the number of *CentBr* repeats per chromosome (17,840) is more comparable with *CEN178* numbers (13,226). This means the centromeric arrays of *B. oleracea* are not as continuous as those of *A. thaliana*. Indeed, the fraction of centromeric regions that are not occupied by *CentBr* annotations is 41.78% and 41.27% for *Alboglabra* and *Italica* respectively, while the same fraction for *Arabidopsis thaliana* Col-CEN *CEN178* centromeres is 13.02%. The disturbed continuity of repeats can also be visualised by the histogram of gaps between consecutive *CentBr* repeats (Fig. 4.23A).

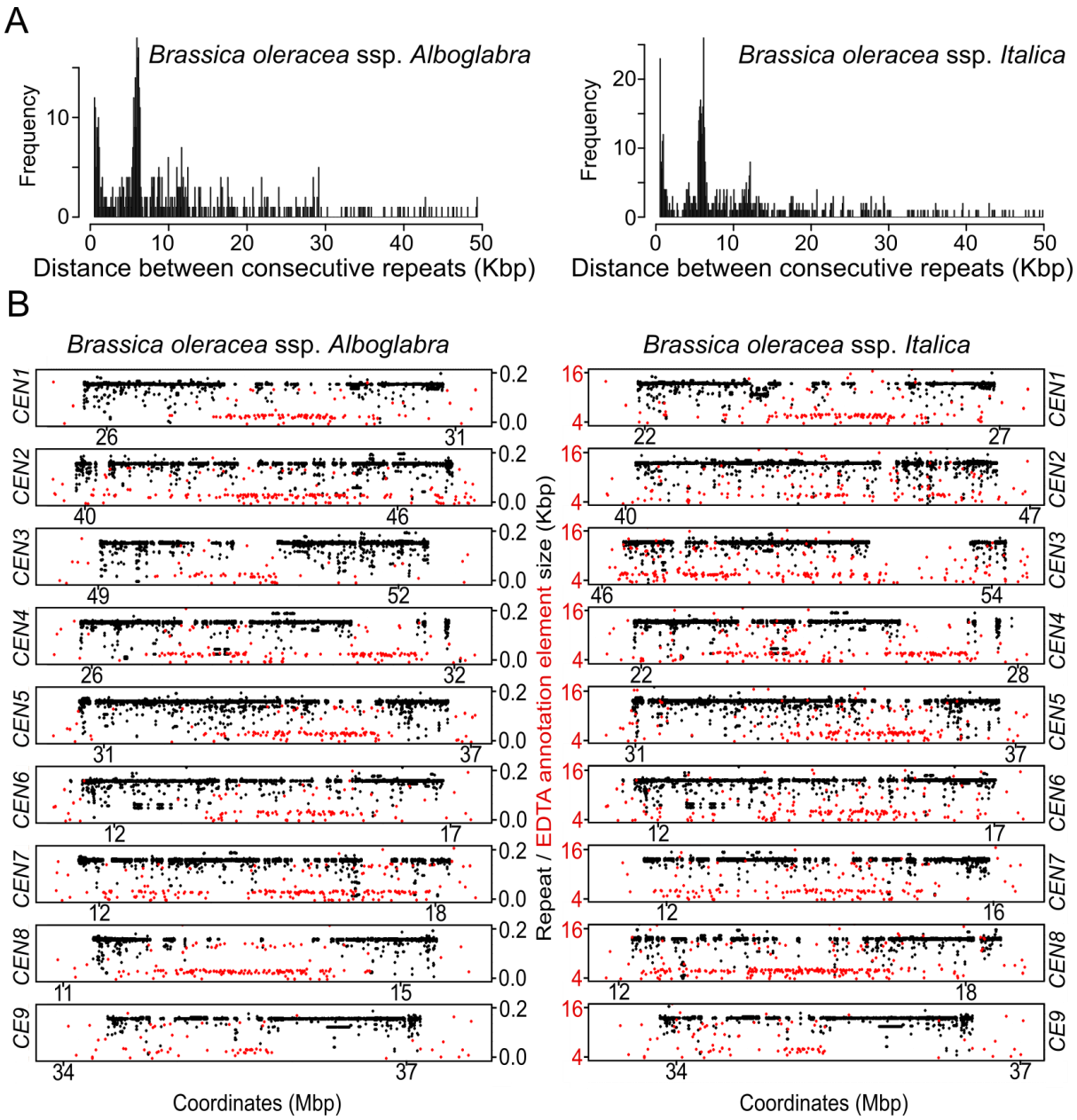


Figure 4.23. *CentBr* array gaps and centromeric transposable elements
A. Histogram of gaps between consecutive *CentBr* tandem repeats. **B**. The position and size of tandem repeats (black) and EDTA annotated TEs (red) in centromeric windows. Note, that black and red points have their own y scales. Short TEs were often overlapping with *CentBr* arrays, so annotation below 4 kbp was not shown. The 6 kbp peaks from **A** correspond to TEs that appear clustered within large centromeric gaps in **B**. and are likely belong to the the *'centromeric retrotransposon of Brassica'* (*CRB*) family (Ki-Byung Lim et al, 2007).

Since transposable elements almost exclusively occupy gaps in the centromeric regions of *Arabidopsis thaliana*, it can be expected that gaps in *Brassica* tandem repeat

arrays will also contain them. TEs were mapped using the EDTA annotation tool (Ou et al, 2019), and 1,020,605 and 670,441 TE instances were identified in *Alboglabra* and *Italica*, respectively. The EDTA TE annotations covered 94.85% of all centromeric coordinates not occupied by *CentBr* repeats. 90.8% of all the annotations were under 1,000 bp and were mostly assigned to DNA family elements. Additionally, the most common TE annotation length in *Alboglabra* was 176 bp, suggesting that the short EDTA annotations likely contain *CentBr* annotations. Because of that, only TEs over 4,000 bp were considered for further analysis, to focus on elements occupying the large gaps of the centromeric regions (Fig 4.23A). The ontology description of centromeric EDTA annotations was predominantly "LTR/Copia" (69.78%), "LTR/Gypsy" (9.41%) and "LTR/unknown" (20.06%) (Table 4.4). A large fraction of centromeric EDTA annotations within the 4-6 kbp range can be observed within internal centromeric regions not occupied by *CentBr* repeats (Fig. 4.23B). These are likely to correspond to the *centromeric retrotransposon of Brassica* (*CRB*) family which contains a single gene encoding a Ty1/Copia-like polyprotein (Ki-Byung Lim et al, 2007). Wang et al, 2011 described the cytological colocalization of *Brassica rapa* (AA) and *Brassica napus* (AACC) CENH3 with *CentBr* and *CRB* elements, confirming this observation.

Source	Name	*Alboglabra*		*Italica*	
		centromeric (TEs >4 kbp)	total (TEs >4 kbp)	centromeric (TEs >4 kbp)	total (TEs >4 kbp)
TRASH	CentBr	155,416	155,695	163,505	65,427
TRASH	BraOle356	-	17,927	-	18,221
EDTA	LTR/Gypsy	214	3,631	238	3,677
EDTA	LTR/unknown	456	1,905	367	1,676
EDTA	LTR/Copia	1,586	6,631	2,375	6,969
EDTA	DNA/Helitron	4	742	3	1,182
EDTA	DNA/En-Spm	3	30	2	26
EDTA	Unknown	2	451	-	67
EDTA	DNA/DTC	1	191	2	202
EDTA	DNA/DTM	6	232	6	226
EDTA	DNA/DTH	1	76	1	70

EDTA	DNA/DTT	-	17	-	16
EDTA	DNA/DTA	-	19	-	21
EDTA	TIR/MuDR_Mutator	-	6	-	-
EDTA	DNA	-	7	-	3
EDTA	RC/Helitron	-	12	-	12
EDTA	pararetrovirus	-	3	-	7

Table 4.4. *Brassica oleracea* ssp. *Alboglabra* and *Italica* tandem repeats and transposable elements annotation.

CentBr repeat edit distance scores from the chromosome consensus and repetitiveness scores were averaged across centromeric regions, with each of them divided into 100 equal windows (Fig. 4.24A). The trends are similar to those of *Arabidopsis thaliana*, with a central decrease in edit distances and increase in repetitiveness, although the trends are less clear, which can be attributed to the less continuous arrangement of the *CentBr* repeats. I also performed HOR analysis of all the *CentBr* repeats, including both sub-families, which identified 7,200,000 instances. Interestingly, *Alboglabra CEN3*, the most isolated chromosome in terms of shared repeats (Fig. 4.22B), contained the lowest amount of *CentBr* HORs (Fig. 4.24B). Chromosomes that contained *CentBr2* are not visibly different from those that contained mostly *CentBr1* (Fig. 4.24B).

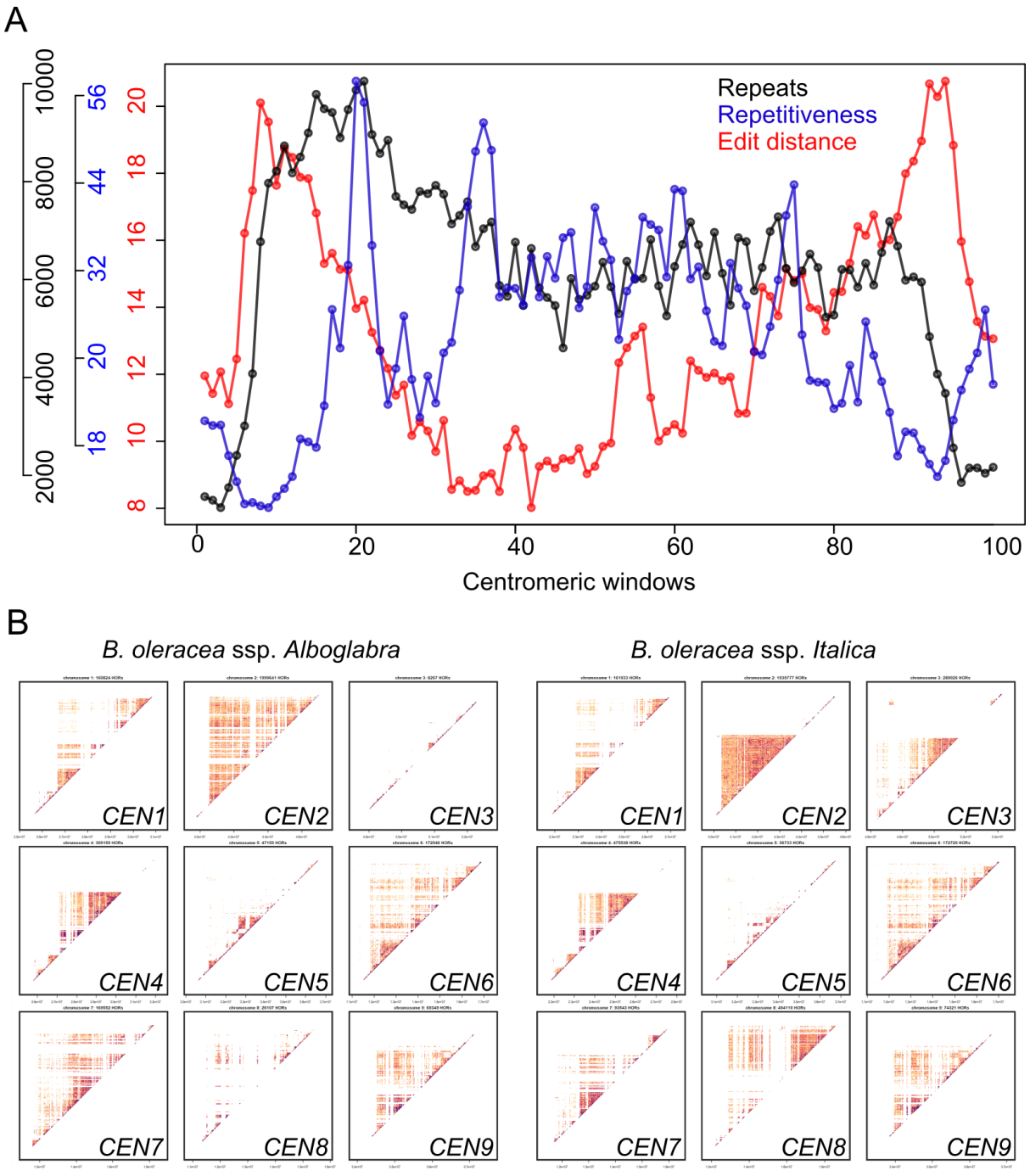


Figure 4.24. *CentBr* tandem repeat edit distance and higher order repeats
A. *CentBr* edit distance (red) and repetitiveness (blue) of repeats (black) averaged over 100 centromeric windows of *Brassica oleracea* chromosomes. **B.** *CentBr* HOR dot plots of *B. oleracea* ssp. *Alboglabra* and *Italica*.

4.1.5 Repeats in the holocentric genomes of *Rhynchospora* genus

Holocentric chromosomes contain multiple loci spread throughout the chromosomes that assemble the functional centromere, specifically kinetochore loading, during cell division (Kratka 2021, Wang, Wu and Yuen 2022). As in monocentric species, centromeric loci of holocentric species can consist of tandemly arranged repeats, and/or transposable elements (Marques et al. 2015, Kratka 2021). For example, chromosomes show this organisation in the genus *Rhynchospora*, where dispersed, short (kbp) tandem arrays of the *Tyba* repeats colocalize with CENH3 signal during meiosis, that is not constricted to a single location, like in monocentric chromosomes (Marques et al. 2015, 2016). Moreover, both female and male meiosis are asymmetric, meaning centromere drive can act in both germlines, in contrast to species such as *Arabidopsis* where only female meiosis is asymmetric (Furness Rudal 2011, Rocha 2016, Kratka 2021). Other than being holocentric, species of this genus are characterised by their large genome size, high levels of structural variation and chromosome rearrangements (Burchardt 2020).

To answer the question of holocentromere evolutionary advantage, Bures and Zedek (2014) proposed the holokinetic drive model which argues that chromosomal fusion and/or repeat proliferation when larger chromosomes are preferred, or chromosomal fission and repeat removal when smaller chromosomes are preferred, parallel to the repeat array expansions and contractions in the centromere drive model.

To investigate tandem repeats, TRASH was run on three *Rhynchospora* genomes: *R. breviuscula*, *R. pubera* and *R. tenuis*, all kindly provided by Dr André Marques from the Max Planck Institute for Plant Breeding Research, Cologne, Germany. 127,952, 293,321 and 52,907 tandem repeats were identified in these assemblies, respectively. The main identified tandem repeat family was 172 bp long and corresponded to 72.5% of all identified tandem repeats (343,577/474,160) (Fig. 4.25A). The consensus of the 172 bp repeat was confirmed to be the *Tyba* family, with little variation between consensus of *R. pubera* and *R. breviscula* (Fig. 4.25B).

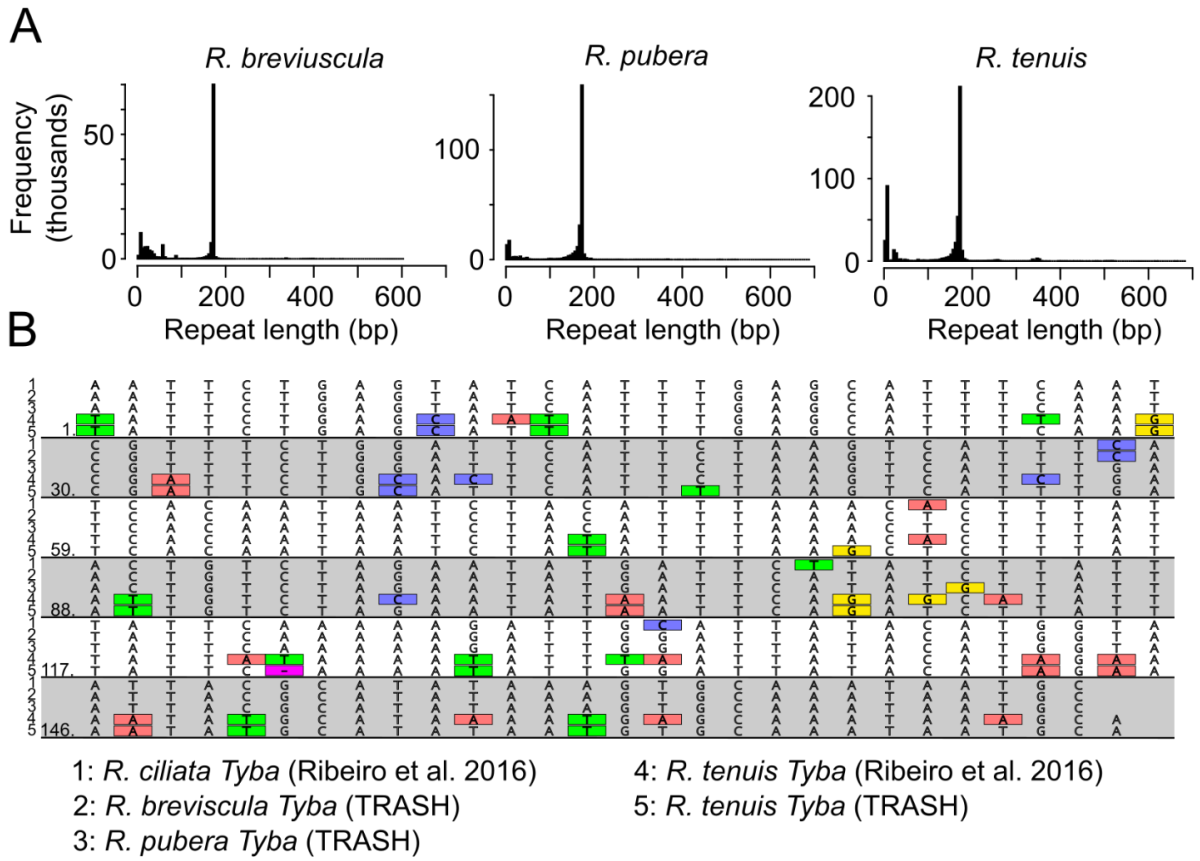


Figure 4.25. *Rhynchospora* tandem repeats and *Tyba* family identification
A. Histograms of all repeats identified by TRASH in the *Rhynchospora* genomes. **B**. Alignment of the *Tyba* consensus sequences from the three genomes (sequences 2, 3 and 5) with previously published *R. ciliata* (sequence 1) and *R. tenuis* (sequence 4) sequences. Minority sequence variants on positions with disagreements are highlighted.

Individual *Tyba* tandem repeat arrays were identified by dividing the repeats based on the distance between them and divided into arrays if that distance was greater than 550 bp (Fig. 4.26A). This value represents three individual Tyba repeats (~516 bp) that can be potentially not mapped, or other short insertions within otherwise continuous arrays (Fig. 4.26B). Average tandem repeat array lengths are 13.2 kbp in *R. breviuscula*, 9.5 kbp in *R. pubera* and 6.8 kbp in *R. tenuis* (Fig. 4.26C). Chromosome length correlates positively with the number of *Tyba* arrays (Pearsons's P=4.3x10^{-10}, r=0.99, Fig. 4.26D) and the total number of *Tyba* repeats (Pearsons's P= 6.7x10^{-6}, r=0.94, Fig. 4.26E), but not the average size of arrays (Pearsons's P=0.03, r =-0.63, Fig. 4.26F) across all chromosomes of the three analysed species. Therefore, it is unlikely that

there is a predetermined number of centromeric arrays in these holocentric species. Instead, the number of tandem repeat arrays appears to fluctuate with the variable genome size.

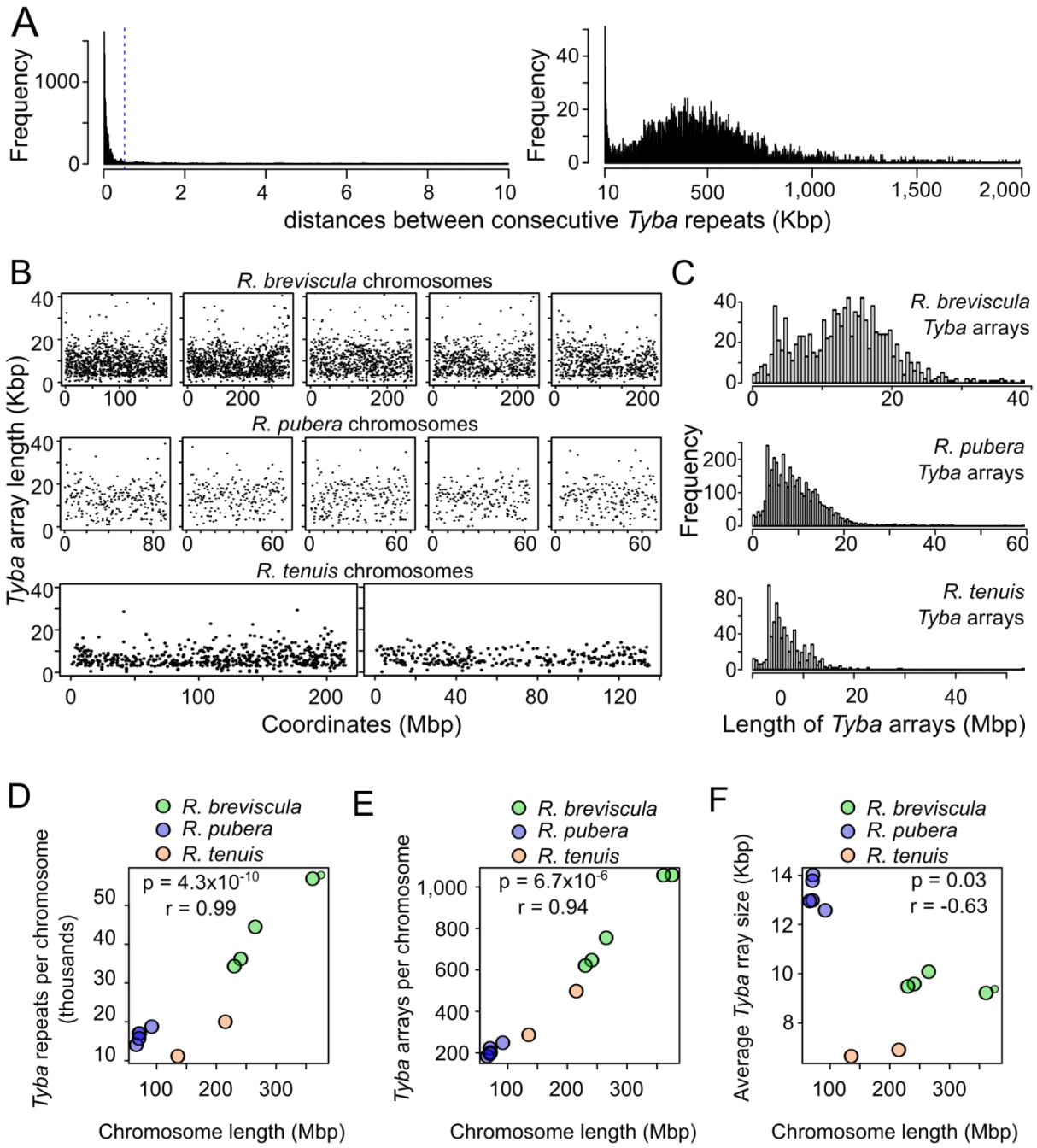


Figure 4.26. Identification of *Tyba* arrays and their characterisation.
A. Histograms of distances between consecutive *Tyba* repeats. On the left, distances under 10 kbp are represented, with 550 bp used for array filtering, marked by the blue horizontal dotted line. On the right, distances equal or higher than 10 kbp are represented. **B**. *Tyba* arrays sizes plotted across the chromosomes. **C**. Histograms of the *Tyba* repeat array sizes in *R. breviuscula*, *R. pubera* and *R. tenuis*. **D**. Relationship

between the number of *Tyba* repeats per chromosome versus chromosome length. **E**. Relationship between the number of *Tyba* arrays per chromosome versus chromosome length. **F**. Relationship between the averaged sizes of *Tyba* arrays per chromosome versus chromosome length.

To measure variability within the *Tyba* repeats, 10,000 repeats were sampled from each species and aligned with mafft (--retree 1, Katoh et al. 2013) to calculate a phylogenetic tree using FastTree (--auto, Price et al. 2009) (Fig. 4.27A). Some clustering can be observed, but no clear distinctions could be made as seen with *Brassica CentBr1* and *CentBr2* (Fig 4.21B). To visualise the positions of the aligned repeats, they were assigned a colour code according to their position within the tree and plotted across the chromosomes (Fig. 4.27). On the chromosome scale, the continuity from phylogeny cannot be identified, although when individual arrays are considered, it can be observed that repeats that clustered together occupy individual islands (Fig. 4.27B).

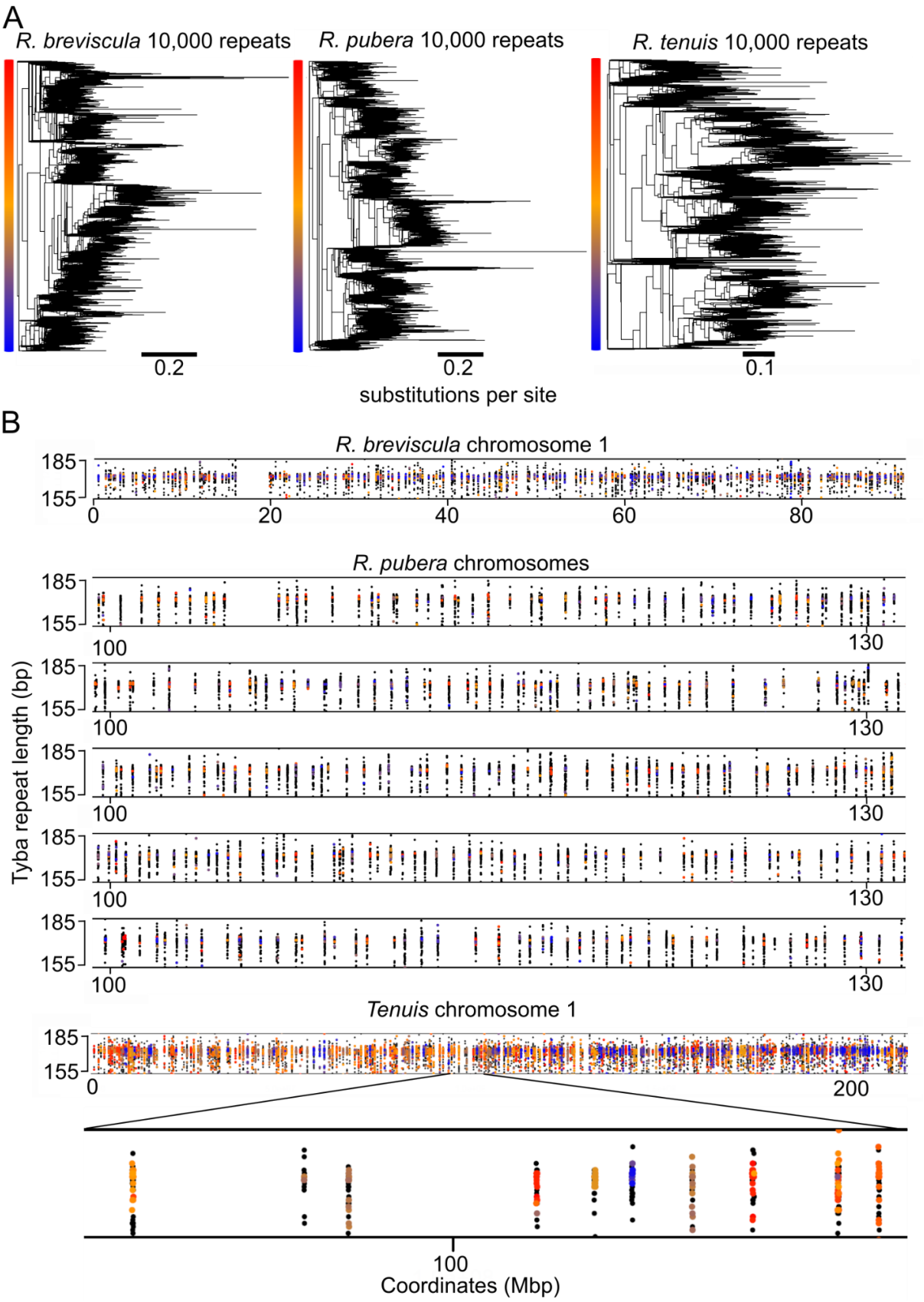


Figure 4.27. *Tyba* repeats phylogeny and clustering across arrays.

A. Phylogenetic trees of *Tyba* repeats made with 10,000 sampled repeats using mafft (--retree 1, Katoh et al. 2013) and FastTree (--auto, Price et al. 2009). Next to them, colour scales that correspond to the repeat colouring on the plots below. **B**. Location of the repeats included in the tree construction, with colour scale according to the position along the tree. From the top, example of a whole chromosome of *R. breviscula*, 32 Mbp extractions of 5 chromosomes of *R. pubera*, whole chromosome 1 of *R. tenuis* and a zoomed in region of it.

18,840 *Tyba* HOR instances were identified across all species, which equates to 0.032% average chromosome HOR abundance (compared to for example 5.67% on average across *Arabidopsis* accessions) (Fig. 4.28A-C). As in previously analysed species, *Tyba* HOR block sizes in monomers and distances between blocks were mostly short, with over half of them being less than 4 monomers long (Fig. 4.28D and 4.28F). Previous studies reported that *R. pubera* chromosomes are the result of end-to-end fusions of ancestral chromosomes (Hofstatter 2022, ref). A recent event would significantly increase the number of identified HORs between duplicated repeat arrays. This can be observed within chromosome 3 of *R. pubera*, where plotted HORs (n=5,997) form a diagonal structure indicating a head-to-head fusion (Fig. 4.28B). Excluding HORs from chromosome 3 of *R. pubera*, only 6.2% of all higher-order repeats were identified between *Tyba* arrays. Within *Tyba* arrays, HORs had less variants per monomer compared to HORs between arrays (2.38 vs 2.85, Wilcoxon *P*<2.2*10^-16) (Fig. 4.28E). This indicates a strong preference for repeat recombination within the arrays. The average distance between *Tyba* HOR blocks for those that were not contained within the same array was 18 kbp, which is lower than average distance between consecutive arrays (400 kbp) (Fig. 4.28F). This could suggest that only proximal arrays can form HORs, or alternatively HORs are almost exclusively formed within an array which is later split by a transposable element or recombination event.

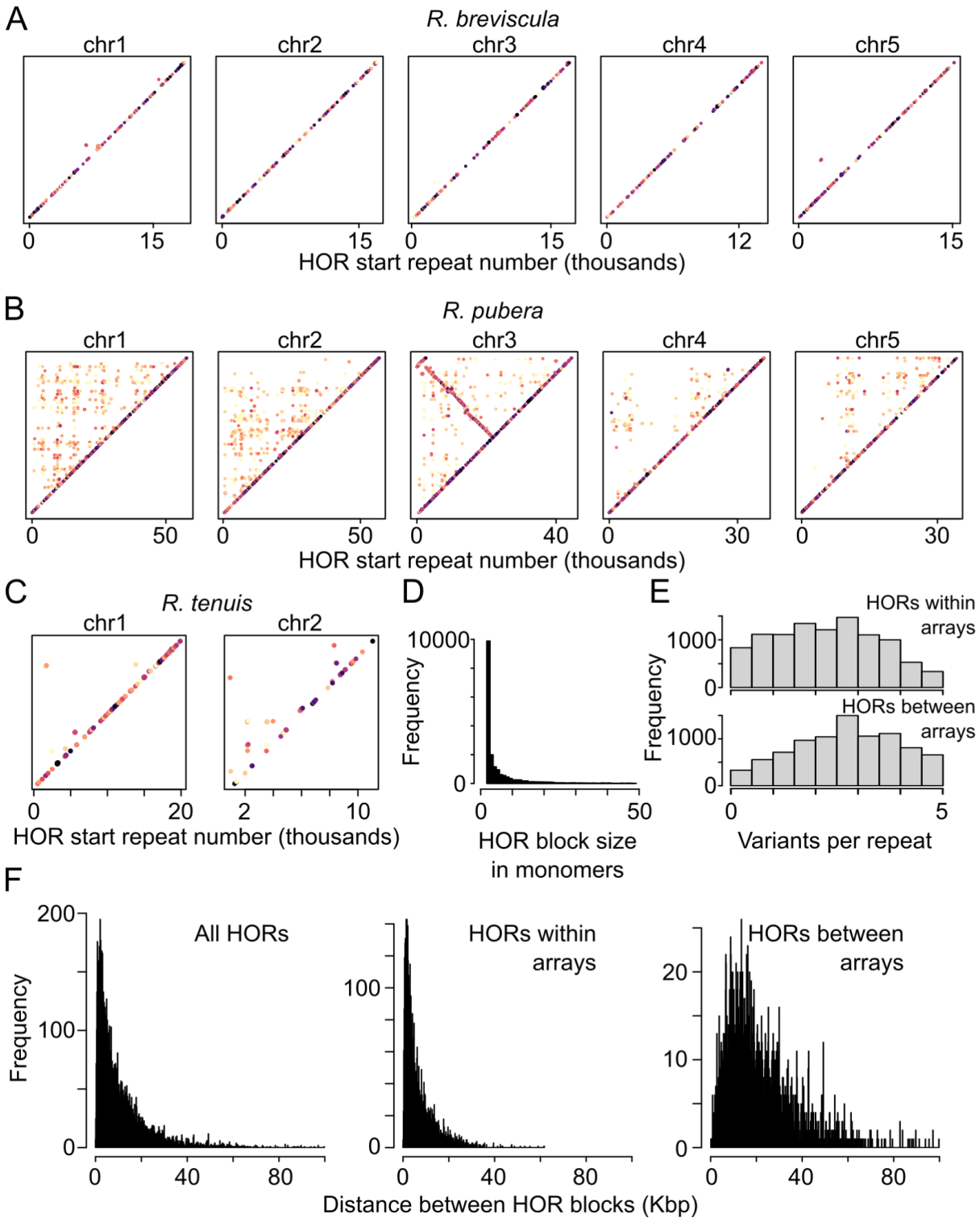


Figure 4.28. *Tyba* HORs across chromosomes of *Rhynchospora* accessions.
A, **B** and **C**. Plots of *Tyba* HORs, with the start monomer numbers for all analysed *Rhynchospora* chromosomes. Due to the low number of HORs, plotting across genomic coordinates is not informative, and monomer numbers are used. **D**. Histogram of *Tyba* HOR block sizes in monomers. **E**. Histogram of variants per monomer of *Tyba* HORs positioned within arrays (above), and between arrays (bottom). A maximum value of five results from the TRASH setting used (-t 5). **F**.

Histogram of distances between *Tyba* HOR blocks. From the left: all *Tyba* HORs, *Tyba* HORs within arrays and *Tyba* HORs between arrays.

In monocentric species, centromeric repeat diversity (measured as the edit distance from chromosome consensus) rises at the edges of the centromere satellite arrays (Naish et al. 2021, Altemose et al. 2021). In contrast, this trend was not found on chromosome level, nor within individual tandem repeat array levels in the *Rhynchospora* genomes (Fig. 4.29A and 4.29B). *CEN178* HOR abundance, a measure of the higher order repeat (HOR) involvement per repeat, decreases towards the centromere edges in *Arabidopsis* (Naish 2021), however this can be seen at the array level in *Rhynchospora*, but not on chromosome level (Fig. 4.16B). Overall, this suggests that tandem repeat islands in holocentric *Rhynchospora* species behave like individual centromeres in *Arabidopsis*, in terms of the distribution of HORs, but not distribution of edit distances.

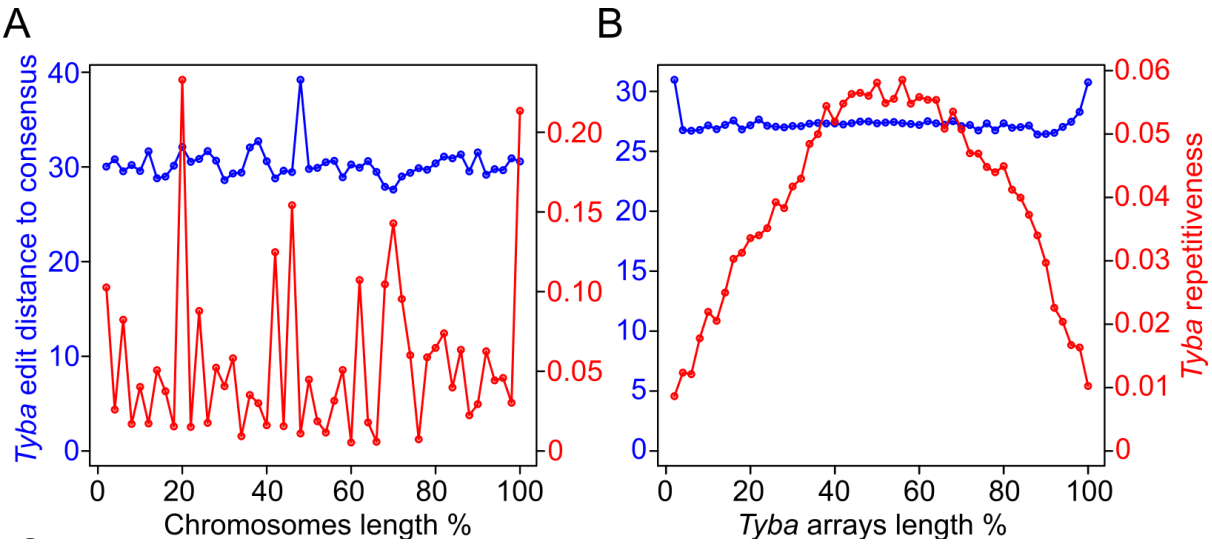


Figure 4.29 Edit distance and HOR abundance across chromosomes and islands of *Rhynchospora*.
A. *Tyba* edit distance from chromosome consensus and HOR abundance across all *Rhynchospora* chromosomes divided into 50 even bins. **B** Same as (A), but individual *Tyba* arrays are considered.

4.2 Discussion

Long-read based genome assemblies provide new insights into the structural organisation of eukaryotic centromeres. Using TRASH and analysis based on its results, I was able to identify and characterise tandem repeats and their families from *Arabidopsis*, *Brassica* and *Rhynchospora* genera.

The extreme diversity of satellite repeats across closely related species is likely to prove problematic in establishing ancestral centromere reconstruction over diverged species but can be used in experimental design to better understand these species. For example, genomic engineering can facilitate the study of diverse satellite arrays to target them with CRISPR-Cas9 system and induce genomic rearrangements that could unlink genes important for breeding, or centromere similarity can be used to infer potential hybrid compatibility.

During this analysis, I developed methods of centromeric sequence analysis, which are to be established within the community. For example, centromeric synteny analysis, which revealed unexpected similarities within the distal regions of centromeric arrays. I also proposed methods for analysis of differential expansions of duplicated tandem arrays, with theoretical and actual examples of their distributions.

Analysis of a large set of *Arabidopsis thaliana* accessions provided insight into the evolution of the centromeric arrays. The distribution of rearrangements within centromeres of the same similarity groups shows that centromere edges remain syntenic, despite accumulating mutations that increase their edit distance from the chromosome consensus. They also tend to be relatively uninterrupted compared to the non-syntenic regions, where no similarities are usually found. This paints a picture of distal centromeric regions harbouring the ancestral arrays that remain similar across the accessions, while HOR-active (and potentially CENH3 accumulating) central regions diverge rapidly. It would be interesting to explore this idea by mapping CENH3 in accessions that are in the same centromere similarity group, where synteny analysis shows expansion of just one region in one of the assemblies, for example Met-6 vs CAMA-C-9 chromosome 3. Additionally, the effect of centromere similarity/co-linearity following crossing can be measured and segregation of

centromeres in subsequent generations can provide information on the molecular basis of the centromere drive.

Questions of centromere evolution could be answered by comparative analysis of closely related species. Analysis of *Arabidopsis thaliana* and *Brassica Oleracea* centromeres, while informative and intriguing themselves, do not provide much information in connection with *Arabidopsis thaliana*, suggesting that even deeper analysis of individual species should be performed, while investigation of different species can inform on the general and likely shared phenomena within the centromeric regions.

4.3 Acknowledgements

Genome assemblies were accessed through public repositories, or kindly shared by: Dr Andre Marques (*Rhynchospora*), Prof. Jose Gutierrez-Marcos (*Brassica*), Dr Polina Novikova (*Arabidopsis lyrata*), Prof. Detlef Weigel, Dr Fernando Rabanal and Prof. Richard Durbin (*Arabidopsis thaliana*). *CEN178* cytology was performed by Dr Terezie Mandáková and Prof. Martin Lysák. Retrotransposons annotation was performed by Dr Alexandros Bousios from the University of Sussex, United Kingdom. CENH3 ChIP-seq was performed by Dr Matthew Naish and mapping of these epigenetic features along CEN178 repeats was performed by Dr Andrew Tock.

Chapter 5

Investigating *HEI10* dosage effects on meiotic crossover recombination in *Solanum lycopersicum* (tomato)

5.1 Introduction

Homologous chromosomes pair and recombine during meiosis, resulting in reciprocal exchange (crossover) (Kleckner, 1996). The process of meiosis reassorts mutations in populations having a profound impact on genetic variation patterns (Gerton et al. 2005, Bolcun-Filas and Handel 2018). Moreover, recombination is also used for domestication and breeding crops and animals, where it can reassort traits following inter- or intra-specific hybridization (Mercier et al. 2015). Meiotic E3 ligase HEI10 can be used to modulate the recombination levels by increasing its dosage in *Arabidopsis* (Ziolkowski et al 2017), but this effect is yet to be reproduced in crop species. In this chapter, identification of tomato *HEI10* homolog and its overexpression is described. Effects of overexpression are measured using genotyping markers and cytologically by DAPI-stained chiasmata counting. Additionally, *HEI10* dosage effect is further characterised in *Arabidopsis* by combining individual overexpressing lines to achieve very high levels of crossover recombination.

5.2 Results

5.2.1 Identification of *SlHEI10* in tomato and generation of transgenic overexpression plants

Solanum lycopersicum HEI10 (Solyc08g015770) was identified using BLAST searches by its homology with the Arabidopsis *HEI10* ortholog and verified by alignments of genomic DNA and amino acid sequences (Fig. 5.1A). Tomato *HEI10* (*SlHEI10*) is highly similar to its *Arabidopsis* homolog (amino acid pairwise identity of 73.9%), and as both plants are diploids within the dicotyledonous group, a similar overexpression effect of crossover increase was predicted in tomato. As no other work on *SlHEI10* genomic sequence was described, to generate an overexpression vector, a 7.2 kb region, including 3.5 kb upstream from the 5' CDS start and 1.5 kb downstream from the 3' CDS end, was amplified by PCR from M82 cultivar genomic DNA (Fig. 5.1B). The *HEI10* genomic fragment was subcloned into a pCAMBIA1300-derived binary vector, using Gibson assembly, and confirmed to be correct by restriction enzyme digestion and Sanger sequencing (Fig. 5.1C-D). A dwarf cultivar, Micro-Tom was chosen to be transformed, as it has shorter generation time than other cultivars (for example, Heinz or M82), and being smaller, needs less space to grow. Transformation was made with the protocol in use in the Baulcombe laboratory, received from Dr Matthew Smoker (John Innes Centre) (described in Methods). Briefly, tomato cotyledons were cut into explants from 5-day old seedlings (Fig 5.1E). Explants were submerged in *Agrobacterium* containing medium, and transferred to regeneration plates, where callus tissue develops, from which shoots can form (Fig 5.1F).

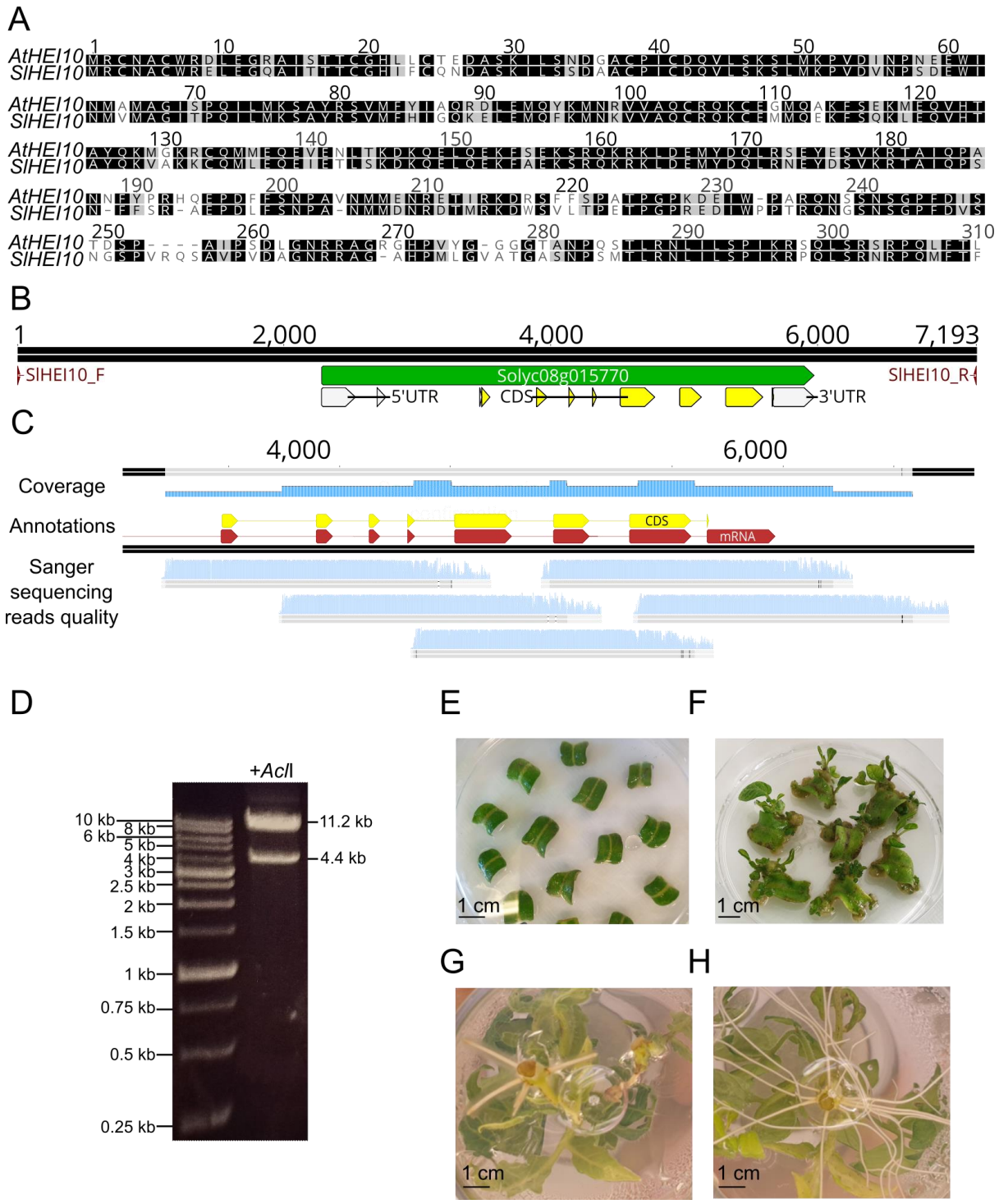


Figure 5.1. Generation of the *HEI10* overexpressing lines in tomato.

A. Identification of the tomato *HEI10* homolog: AtHEI10 and SlHEI10 (translated Solyc08g015770 coding sequence) protein sequence alignment. **B**. The *SlHEI10* nucleotide sequence, and surrounding regions, are shown with annotated primer positions (red) used for cloning. Green is Solyc08g015770 gene annotation, yellow is its coding sequence and white are 5' and 3' untranslated regions. Araport11 annotations were obtained from the TAIR website. **C**. Confirmation of binary vector construction using Sanger sequencing, and **D**. restriction enzyme digestion with *Acl*I. **E**. Tomato cotyledon cuttings after being placed on a regeneration medium. **F**. 4-week-

old explants, showing callus and shoot formation. **G**. Shoots cut and placed on a rooting medium, which are developing stubby roots, indicating a false positive (view from below). **H**. Root development in a true transformant.

Initial transformation of the Micro-Tom cultivar gave poor recovery of leaf formation following tissue culture of calli. Only seven explants showed callus and shoot induction, out of around 500 initial explants, giving a transformation rate of 1.4%. This is compared to around 30% achieved in previous experiments performed in the laboratory (unpublished observations). The protocol was therefore adjusted according to additional literature and personal communications to achieve higher transformation efficiencies. The changes included:

- Addition of cytokinin (8.9 μm BA) to the germination medium, as its presence was reported to increase transformation efficiency from 6% to 22% (Kumar Rai et al. 2011).
- A period of explant incubation was added, after cutting them from cotyledons, instead of immediate transfer onto recovery medium. This step is commonly used for tomato transformation, and its duration of at least three days was reported to be crucial for efficient transformation (Kumar Rai et al. 2011).
- Addition of 100 mM acetosyringone to the *A. tumefaciens* resuspension medium, as this induces *Agrobacterium vir* gene expression and has been used to increase transformation efficiencies in various plant species, including *Arabidopsis* (Sheikholeslam and Weeks 1987, Manfroi et al. 2015)
- A washing step to remove *Agrobacterium* from the plants after transformation. Without this washing, the *Agrobacterium* solution can inhibit the growth of explants, therefore a washing medium consisting of 0.33xMS was used (Kumar Rai et al. 2012).

As some protocols include making incisions on the adaxial site and some not, half of the explants were gently cut with a scalpel on their adaxial surface. None of the explants that were cut survived, as all of them developed necrosis starting from incision sites. Another variation was the addition of sucrose to the germination medium. Most protocols include sucrose, but it was suggested that tomatoes germinate better on no-sucrose medium and are less prone to infections (personal

communication with Dr Matthew Smoker). After supplementing with 3% w/v sucrose, by the time of cotyledon excision, only around 70% of seeds germinated, and they were not as synchronised as in sucrose-free germinating seeds. This synchronisation is important, as transformation efficiency is dependent on seedling age (Kumar Rai et al. 2012). Despite that, explants that originated from sucrose-containing media developed callus faster and maintained a green colour for longer during culture. Overall, to avoid potential contamination and ensure germination synchronisation, no-sucrose germination medium was used for the further transformations. Changes made to the protocol increased the transformation rate in terms of callus and shoot formation on the explants from 2/85 to 47/100.

Shoots formed on the regenerating calli over a period of several months. Multiple shoots can be formed on a single explant. Whenever a shoot reached 2 cm, it was carefully excised from the explant above the callus and transferred to the rooting medium. There, in the absence of cytokinins used for shoot induction, the natural ability of tomato plants to develop roots is taken advantage of. False positive transformants are known to occur during selection and they can be identified by different root morphologies compared to true transformants (personal communication with Dr Matthew Smoker) (Fig. 5.1G-H). False positive transformants develop thicker and shorter roots on phosphinothricin (PPT) containing medium compared to transgenic ones and were discarded. The rooting process takes several weeks and not all shoots were successfully rooted. After a sizable root system was developed, transgenic plants were carefully cleaned of the media and transferred to soil. Because of the lengthy and unsynchronised nature of transgenic plant generation, genotyping and expression assays were performed as soon as the material became available. The first eight transgenic T_0 plants were tested for transgene presence using genomic PCR genotyping and they were positive by this assay (Fig. 5.2A). For PCR positive transformants, *SlHEI10* mRNA expression was tested using a RT-qPCR assay (Fig. 5.2B), which showed over a 1.5-fold increase for *HEI10ox#1 SlHEI10* mRNA levels compared to a non-transgenic wild type Micro-Tom control, using double delta-Ct normalisation (Fig 5.2C).

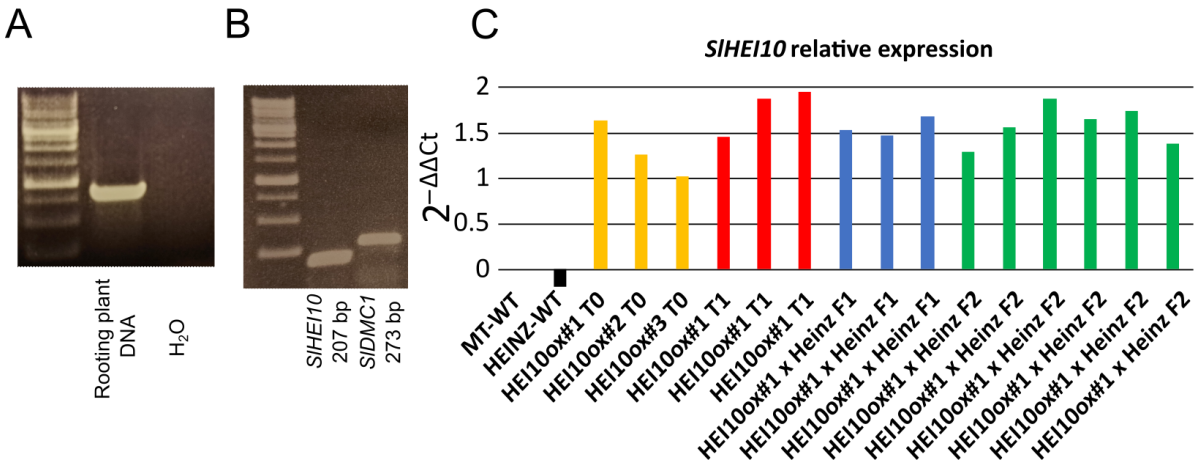


Figure 5.2. Relative SlHEI10 expression in the transgenic lines.
A. Gel electrophoresis of the products of the genotyping the vector presence in a rooting plant exhibiting regular root growth (**Fig. 4.1H**). Genotyping at an earlier stage is not recommended to not inhibit the regeneration (personal communication). Genomic DNA and water negative control shown. Expected amplicon size was 782 bp. **B**. Gel electrophoresis of the products of the qRT-PCR used to quantify *SlHEI10* expression. **B**. $2^{-\Delta\Delta Ct}$ expression measurements normalised against Micro-Tom wild type and *SlDMC1* expression were performed, as described in Methods. In black, wild type Heinz measurement; in orange, three independent *HEI10ox* T_0 transformants; in red, three F_1 plants from *HEI10ox#1* crossed with wild type Heinz and in green, six F_2 plants after propagation of one of the F_1 plants. Apart from wild type plants, only plants with the transgene were used for the expression assay, as confirmed by PCR genotyping. Tissue was collected for each plant, flash frozen in liquid nitrogen and stored to allow for simultaneous RNA extraction.

HEI10 T_0 plants were reciprocally crossed to the Heinz cultivar to generate F_1 hybrids and genotyped for the presence of the transgene. The *HEI10ox#1* plant was observed to have the highest mRNA overexpression and was crossed for the subsequent experiments (Fig. 5.2C). F_1 transgenic plants showed a significant increase of *SlHEI10* expression at levels similar to the T_0 generation (compared to the wild type plants, Student's *t* test $P = 0.0013$) (Fig. 5.2C). One of the *HEI10ox* Micro-Tom/Heinz F_1 plants was propagated into the F_2 generation and the transgene was found to be present in 81 out of 121 plants, showing approximate Mendelian segregation consistent with a single T-DNA locus. *SlHEI10* overexpression in *HEI10ox* F_2 plants remained stable, and the line was deemed suitable for further analysis (Fig. 5.2C).

5.2.2 Analysis of the crossover recombination increases in *SlHEI10* overexpressing tomato plants

Cytological analysis of chiasmata formation at metaphase I can inform on the crossovers formed in tomato (Strelinkova, Komakhin and Zhuchenko 2019). Bivalents that form two crossovers at opposite ends of the chromosomes typically form ring structures, while a single crossover results in a rod structure (López et al. 2012). Chiasmata counting of wild type and *SlHEI10* overexpressing tomato was performed by Dr Christophe Lambing. 40 nuclei were analysed for Micro-Tom wild type and *HEI10ox#1* plants (Table 5.1 and Fig 5.3A). The average number of chiasmata was slightly, although significantly, higher in the *SlHEI10* overexpressing plant compared to wild type (17.1 *HEI10ox* vs 16.1 wild type chasmata on average) (Wilcoxon test, *P*=0.0067). However, cytological analysis of chiasmata is limited in resolution and can underestimate closely spaced crossovers. Therefore, I next performed a genetic mapping experiment.

Micro-Tom wild type			*HEI10ox#1*		
ring	rod	chiasmata	ring	rod	chiasmata
4	8	16	6	6	18
1	11	13	2	10	14
4	8	16	2	10	14
1	11	13	4	8	16
4	8	16	5	7	17
3	9	15	4	8	16
6	6	18	5	7	17
4	8	16	6	6	18
4	8	16	5	7	17
3	9	15	6	6	18
4	8	16	5	7	17
5	7	17	5	7	17
5	7	17	7	5	19
3	9	15	7	5	19
5	7	17	5	7	17
6	6	18	7	5	19
3	9	15	5	7	17
3	9	15	12	0	24

5	7	17	1	11	13
6	6	18	5	7	17
5	7	17	3	9	15
2	10	14	4	8	16
5	7	17	6	6	18
3	9	15	4	8	16
5	7	17	7	5	19
5	7	17	7	5	19
3	9	15	3	9	15
4	8	16	1	11	13
4	8	16	6	6	18
4	8	16	8	4	20
3	9	15	2	10	14
4	8	16	7	5	19
3	9	15	6	6	18
9	3	21	5	7	17
4	8	16	5	7	17
6	6	18	8	4	20
4	8	16	3	9	15
5	7	17	5	7	17
4	8	16	4	8	16
3	9	15	6	6	18

Table 5.1 Proportion of rod and ring bivalents at metaphase I in DAPI spreads from transgenic *HEI10ox#1* and wild type Micro-Tom.
The twelve bivalents of metaphase I DAPI spreads were scored as either rod (1 chiasma) or ring (2+ chiasma), depending on chromosome morphology by Dr Christophe Lambing (Strelinkova, Komakhin and Zhuchenko 2019). Significant differences between the lines were observed (Wilcoxon test, *P*=0.0067).

To genetically identify crossovers, the *HEI10ox* lines and wild type controls were genotypes in the F_2 using simple short length polymorphism (SSLP) markers. An insertion/deletion list generated by short-read sequencing of Micro-Tom cultivar and mapping against the Heinz SL2.5 genomic assembly (Kobayashi et al. 2014) was used to design the SSLP markers. Chromosome 5 was chosen for analysis, as it contained the highest number of insertions and deletion in the data out of all chromosomes (37,923 compared to 15,290 genome average). In total, 32 SSLP primer pairs were designed with amplicon length differences ranging from 6 to 936 bp between Micro-

Tom and Heinz (Table 2.2). Of the 32 designed, 30 SSLP markers showed amplification with at least one amplicon of the correct size. Out of them, only 13 pairs showed the expected amplicons when tested using Micro-Tom, Heinz and F_1 hybrid template DNA (Fig 5.3B). When tested on 121 F_1 plants from a *HEI10ox#1/*Heinz F_1 parent, and 71 from wild type F_1 parent, three primer pairs had to be excluded due to lack of amplification in most plants. The average number of chromosome five crossover events per Mbp per F2 identified using the SSLP makers and calculated based on the genotype transitions in the marker results was 0.0195 for the *HEI10ox* parent, and 0.0163 for the wild type cross. The difference was not significant based on the unpaired Wilcoxon test (*P*=0.186), possibly due to mostly small values in the data, with 92% of chromosomes showing 0, 1 or 2 crossovers, leading to a large number of zero-differences. The number of chromosome 5 crossovers per Mbp per individual in the previously published data (Demirci et al. 2017) was 0.0242, which is higher than the measurements made here (0.0195 and 0.0163). This could be attributed to double-crossover events potentially unaccounted for, due to the relatively lower density of SSLP markers compared to the Demirci study (which used sequencing to identify crossovers). The distribution of crossovers along chromosome 5 was plotted for both data sets (Fig 5.3C). The distribution of the number of crossovers observed in equivalent chromosome regions was similar to previously published wild type data sets (Demirici et al. 2017) (Fig 5.3D).

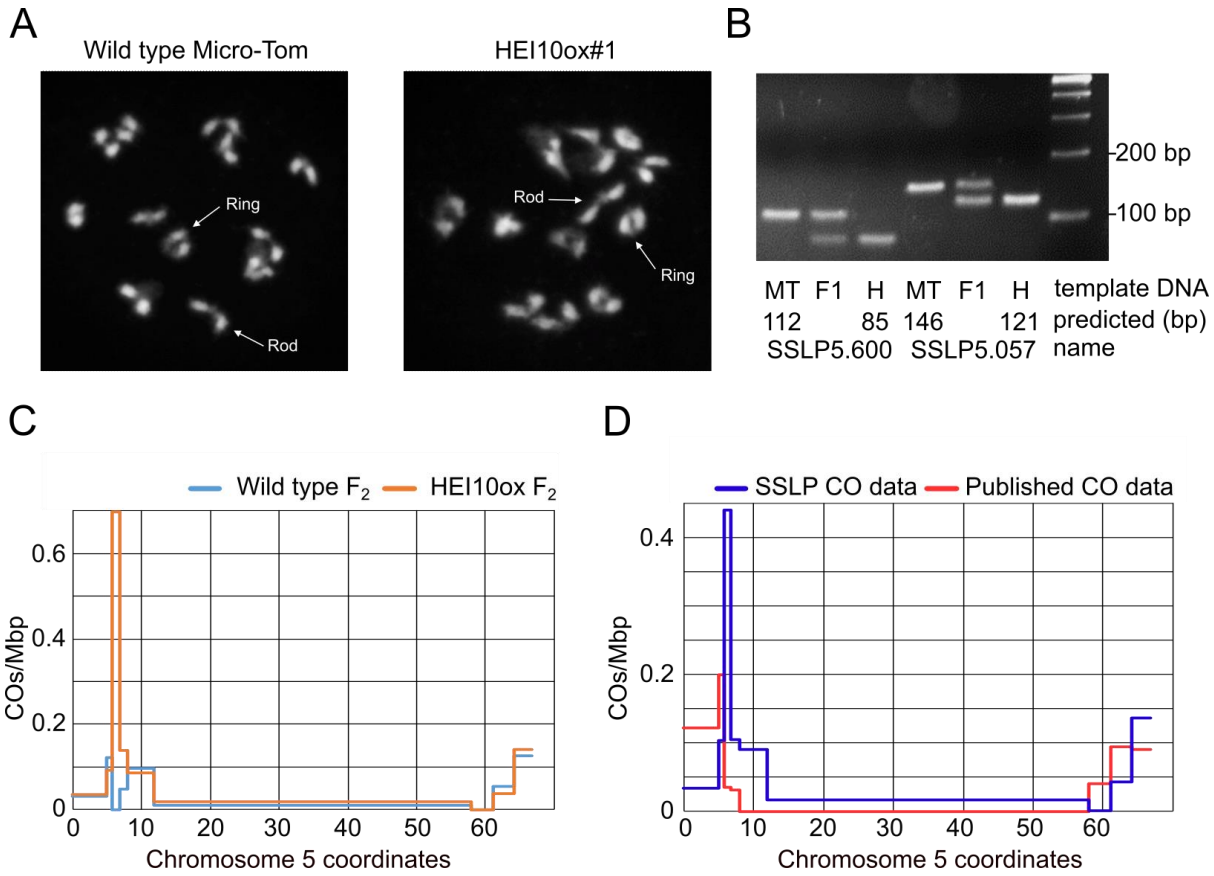


Figure 5.3 Tomato *SlHEI10* overexpression effect on detected crossover levels
A. Micrograph showing representative chiasmata formation at metaphase I in DAPI-stained spreads from a single wild type Micro-Tom and *HEI10ox#1* plant. Examples of ring and rod structures are marked by white arrows and labelled. Data were collected and analysed by Dr Christophe Lambing. **B**. An example of gel-based genotyping of SSLP markers. Products of PCR amplification using genomic DNA from Micro-Tom, Micro-Tom/Heinz F_1 hybrid and Heinz are presented for markers 5.600 and 5.057. The predicted amplicon sizes (bp) are annotated below. **C**. The distribution of crossovers identified in transgenic *HEI10ox#1*/Heinz F_2 plants (orange) vs wild type Micro-Tom□Heinz F_2 cross (blue). **D**. Chromosome five crossover distribution data from Demirici et al. 2017 fitted into chromosome bins according to the SSLP marker locations compared to all crossovers per Mbp identified in this study. Interval no 3 exhibited high increase in the *HEI10ox#1*/Heinz F_2 plants, while in the same interval in the wild type F_2 no crossovers were found.

5.2.3 Testing for *HEI10* dosage saturation effects on crossover frequency in *Arabidopsis thaliana*

As *HEI10* dosage modulates crossover numbers in *Arabidopsis thaliana (*Ziolkowski et al. 2017), it was interesting to analyse the kinetics of this effect and test if a saturation

point can be reached. For this experiment, a set of lines with known *HEI10* transgene location, copy-number and expression value is required, that can then be intercrossed to generate multi-transgenic lines with increasing dosage of *HEI10*. Previously, the Columbia accession carrying the *420* FTL reporter was transformed with a genomic *HEI10* transgene that uses the endogenous promoter and terminator (Ziolkowski et al. 2017). T_2 seeds from these experiments were obtained from Dr Piotr Ziolkowski and will be referred to as the PZH lines. The PZH lines (T_1) had previously measured *420* crossover measurements and can be analysed to single locus *HEI10* transgenics that could be combined by crossing (Fig. 5.4A). 13 T_2 *HEI10 420*/++ lines with T_1 crossover frequency rates higher than 30 cM were chosen (Fig. 5.4B). These lines had an unknown *HEI10* transgene copy number. Therefore, to isolate single locus lines I performed progeny testing, where T_2 seeds were sown on kanamycin containing plates and the number of sensitive versus resistant seedlings was scored (Fig. 5.4A and B). Lines with a segregation ratio of 3:1 (resistant to sensitive) (n=100) indicate lines with a single locus conferring resistance (Fig. 5.4B), although it is possible that multiple T-DNAs may be present at that locus, or T-DNA fragments could be present elsewhere in the genome that lack the resistance gene or are silenced.

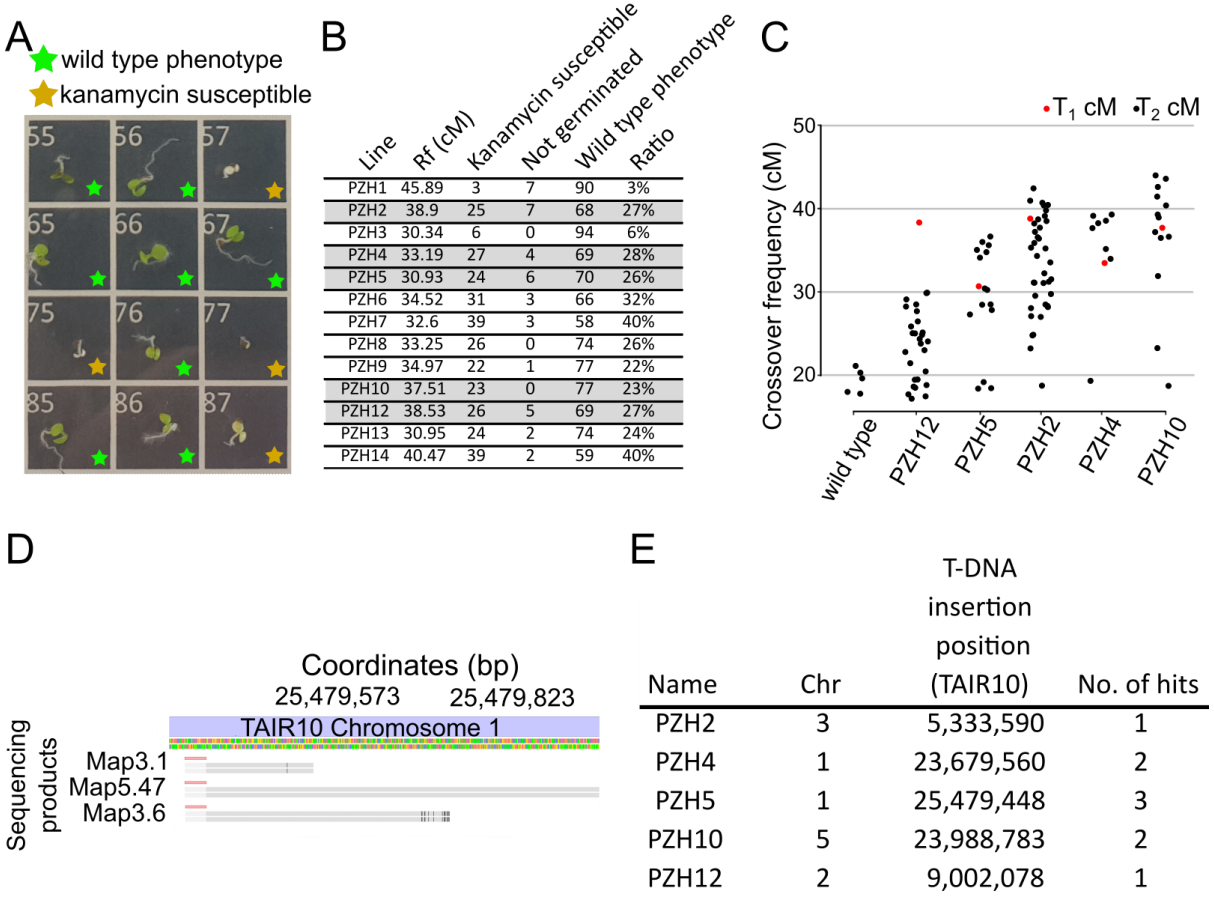


Figure 5.4 Identification of *Arabidopsis thaliana* single-locus *HEI10* overexpression lines.

A. Example of the progeny test seedlings exhibiting kanamycin (marked with orange stars) and resistance to kanamycin (marked with green stars) phenotypes. **B**. Segregation ratios of the initial *HEI10* overexpressor lines, based on the kanamycin resistance phenotypes. Highlighted are the five PZH lines chosen for further analysis **C**. *420* crossover frequency (cM) in the five individual PZH lines. Red dots represent parental cM (as in panel B). **D**. An example Tail-PCR analysis used for transgene mapping. Three overlapping Sanger reads mapped to the TAIR10 reference genome precisely identify the insertion site of the PZH5 *HEI10* line. Above is a representation of chromosome five (25,497,350:25,480,014) and below mapped positions are indicated by grey shading, with black shading showing cases of alignment disagreements. Pink regions at the beginning of the three reads are trimmed sequences complementary to the transgene cassette. **E**. The location of the mapped transgenes using TAIR10 as a reference genome. The number of independent Sanger sequencing reads that identified the position is indicated.

Eight PZH lines with a 3:1 resistance segregation ratio was identified. Five of these lines with T_1 *420* crossover frequency ranging from 30.9 to 38.5 cM were chosen for

further analysis (Fig. 5.4B). T_2 seeds were sown in order to make primary measurements of T_2 *420* crossover frequency for putative hemizygous and homozygous *HEI10* transgenic individuals (Fig. 5.4C). Crossover frequencies did not show clear division into three discrete levels, which would indicate single locus segregation (Fig. 5.4C). This suggests that the PZH lines might carry additional transgene copies that were not identified in the progeny tests. Nonetheless, transgene mapping using an updated TAIL-PCR protocol (Liu and Chen, 2018) was performed (Fig. 5.4D). T-DNA mapping efficiency was 13.2% (9 out of 64 sequenced amplicons mapped to the *Arabidopsis* genome), which was lower than over 90% achieved by Liu and Chen (2018). The majority of the sequenced amplicons mapped back to the plasmid, suggesting multiple tandemly arranged transgene insertions. Despite this, one genomic locus per PZH line was identified which allowed for the design of genotyping primers (Fig. 5.4E).

Quantitative PCR was performed as an alternative method of *HEI10* transgene copy-number definition. I focused analysis on one of the lines, PZH12, as it was the first one to acquire genotyping primers. Five plants with no transgene based on genotyping were used as a control group for ΔΔCt calculations and sample normalisation was achieved by *ACT11* amplification. 21 individuals (including controls) were analysed using qPCR and, after seeds were available, *420* crossover frequency was calculated. A significant correlation between the *HEI10* qPCR measurement and *420* crossover frequency was observed (Student's *t* test, $p = 7.58 \times 10^{-14}$) (Fig. 5.5A). To investigate whether *HEI10* expression is more strongly correlated with *420* crossover frequency than *HEI10* DNA copy number, floral bud tissue was collected from the same population for RNA extraction. To achieve meiotic tissue normalisation, qRT-PCR was performed using the meiosis-specific *DMC1* recombinase as a "housekeeping" gene (Klimyuk and Jones 2002). *HEI10* expression level was also positively correlated with *420* crossover frequency (Student's *t* test, $p < 2.2 \times 10^{-16}$), but in this case it formed more of a linear relationship ($R^2 = 0.675$) (Fig. 5.5B). This shows that despite PZH lines having uncertain genomic backgrounds in terms of the copy number of *HEI10* transgenes, *HEI10* expression was elevated and correlated with increased crossover frequency. Tissue collection, DNA extraction, cDNA synthesis, qPCR and seed scoring in PZH12

analysis experiment was performed with help of a summer student, Mr. Thomas Underwood.

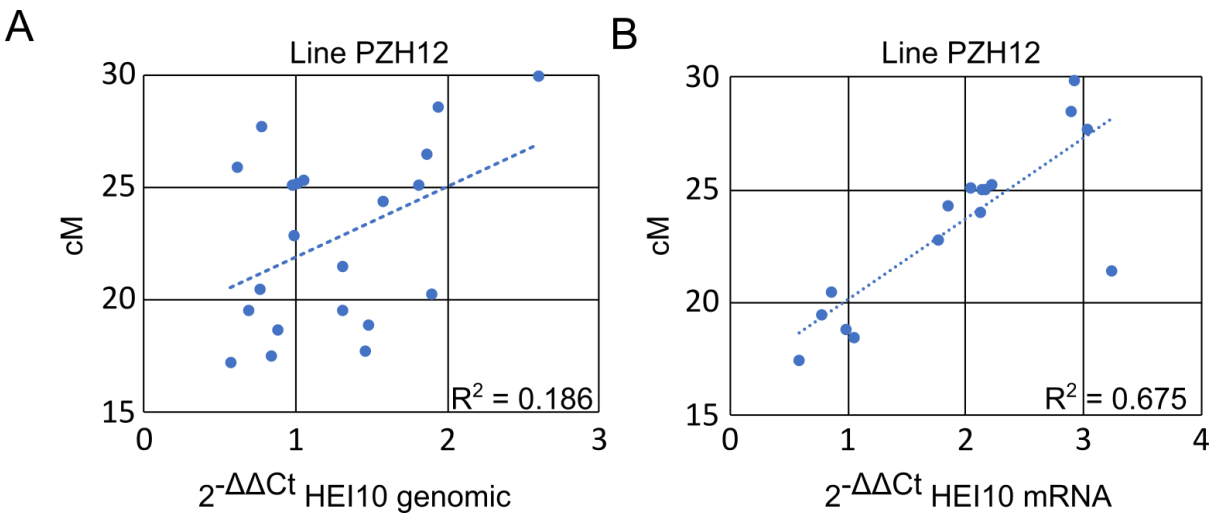


Figure 5.5 Correlation between *HEI10* copy number measurements and *HEI10* expression and *420* crossover frequency in PZH12 T_2 individuals.
A. *HEI10* copy number qPCR analysis of the *HEI10* abundance relative to *ACT11* plotted against *420* crossover frequency (cM). **B**. *HEI10* gene expression relative to *DMC1* calculated using the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001), plotted against *420* crossover frequency (cM).

In order to characterise the effect of very high levels of *HEI10* dosage on meiotic recombination, PZH lines were crossed reciprocally to combine transgenes into multi-copy lines. The pollen donor for the crosses was always a *420/420* double homozygous plant, so that a successful cross can be confirmed by the presence of the fluorescent protein transgenes. Unfortunately, not all crosses could be acquired in a timely manner and in some cases, crosses were failing consistently, F_1 seeds did not germinate, or despite successful crossing (based on the *420* genotyping) expected transgenes could not be identified. For these reasons, a limited number of PZH combining genotypes was acquired. *420* crossover frequency data which include T_2, T_3, F_1 and F_2 plants pooled together according to their genotype are presented in Figure 5.6. PZH10 results are not included as the designed genotyping primers often did not amplify. with an increasing number of *HEI10* transgenes, the crossover frequency is capable of reaching 50 cM, essentially unlinking the *420* loci. Some measurements suggested that eventually *HEI10* transgenes may undergo silencing.

For example, PZH5 line crossover frequency in a heterozygous configuration is around 32 cM, and PZH4 around 38 cM (Fig 5.6). In homozygous states, PZH5 reaches 38 cM, while PZH4 average frequency decreases, reaching both above 45 cM, and below 25 cM values, which suggests that transgene silencing becomes activated at elevated levels of *HEI10* expression. When PZH4 is crossed with PZH5 and both are in heterozygous states, average *420* crossover frequency is higher than any of the other configurations. It is worth noting that apparent *420* cM values might decrease due to increased frequency of double crossovers in high copy *HEI10* lines. Despite being a result of highly increased crossover levels, these events, when occurring within an interval, maintain their linkage and are missed during the fluorescent seed-based analysis. Further experiments that overcome the issues of multi-transgenic lines, like genotyping-by-sequencing (GBS) could explore these phenomena in greater detail.

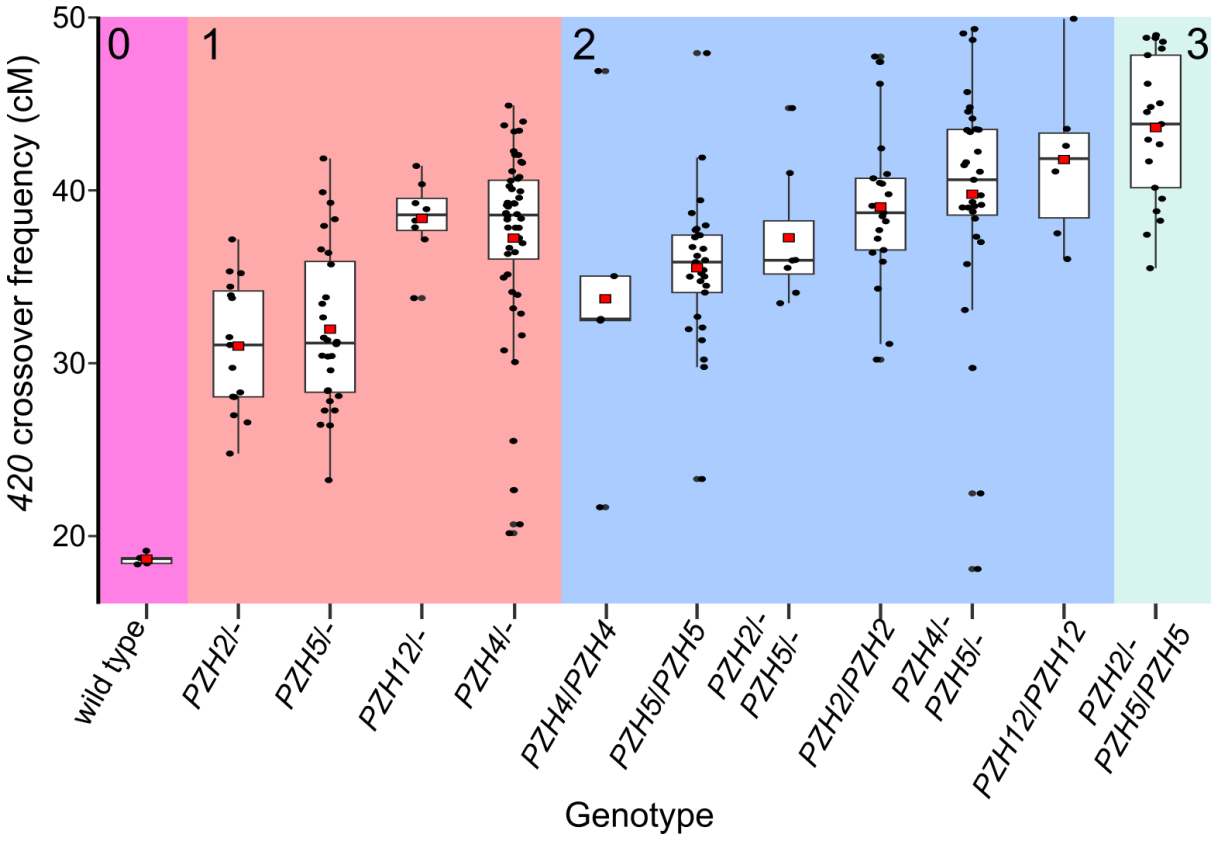


Figure 5.6 Crossover frequencies of the *HEI10* PZH lines after selfing and crossing
Genotype of each line is described under the plot, presence of the transgene is described using its italicised name and absence with "-", for example *PZH2/-* represents a PZH2 line in a heterozygous state. Sections of the plot are coloured according to the number of additional *HEI10* loci identified in this study: 0 in pink, 1 in orange, 2 in blue and 3 in green.

5.3 Discussion

HEI10 family proteins are thought to act as ubiquitin or SUMO ligases that modify recombination factors in order to promote crossover repair (Mercier et al. 2015). Here, I have identified a homolog of *HEI10* in tomato based on its sequence similarity to *Arabidopsis thaliana HEI10*(Ziolkowski et al. 2017). Over 2-fold overexpression (based on HEI10 in tomato was achieved and confirmed by qRT-PCR. The effect of *HEI10* overexpression on meiotic recombination was assessed using simple sequence length polymorphism based genotyping method, and cytologically by counting chiasmata on mitotic DAPI spreads. It is likely that increased recombination is localised within the chromosome distal regions, where it is found natively, based on results obtained in Arabidopsis (Rowan et al. 2015, Ziolkowski et al. 2017). Using the developed *HEI10ox* overexpressing lines and combining them with knockouts of anti-recombination pathway genes, such as helicases *RECQ4A* and *RECQ4B* could potentially further additively promote crossover frequency (Sierra et al. 2018).

I also established a number of highly recombining lines by combining *Arabidopsis thaliana HEI10* overexpressing lines, which were used to investigate the behaviour of crossover frequency with progressively increasing dosage of *HEI10*. In these lines, *420* interval crossover frequency was increased to its maximum 50 cM value, using a combination of just two PZH *HEI10* lines. A moderate decrease of crossover frequency was observed when lines of highest frequencies were crossed. This could be attributed to the properties of the used system, where an even number of crossovers within the *420* interval lead to reestablishment of the parental genotype in regards to the *420* loci and crossover underestimation. Interestingly, increased *HEI10* dosage has been connected with a decrease in crossover interference (Serra et al. 2018, Morgan et al. 2021), which could explain consistent acquisition of additional crossovers in the analysed *420* region, leading to a decrease in the measured recombination. Transgene silencing could also play a role in inhibition of the *HEI10* effect at transcriptional or post-transcriptional levels. These lines provide an opportunity for even higher increase in *HEI10*-mediated recombination and together with a high-throughput analysis method like genotyping-by-sequencing, can inform on the recombination landscape in this extreme scenario.

My *HEI10* experiments were stalled by the COVID pandemic which restricted access to the facilities. Most work during that time was focused on the bioinformatics project and it became more achievable in terms of the goals set in my PhD proposal, while experimental work encountered additional difficulties. Around 6 rounds of tomato transformation were performed before acquiring transformant plants, each of them took several days to complete and required over a month to estimate their success. Similarly, mapping of the PZH lines required an immense amount of gel extractions and sequencing to acquire just a few results, which later could not fully explain the whole crossover frequency phenotypes. For these reasons, I did not fully accomplish the goals set in my PhD proposal regarding *HEI10* work, which included genotyping-by-sequencing experiments for both of them.

5.4 Acknowledgements

HEI10 overexpressing lines were provided by Dr Piotr Ziolkowski. Chiasmata counting was performed by Dr Chris Lambing. Analysis of the PZH12 line was a collaborative effort between myself and Mr. Thomas Underwood.

Chapter 6

Discussion

The results described in this thesis address several aspects of plant genome organisation and analysis. In Chapter 3, I described the development of TRASH, a novel bioinformatic tool capable of accurately identifying repeats in tandemly arranged arrays *de novo* or using pre-defined templates. It is capable of a higher-order repeat annotation, providing insight regardless of analysed species and pre-existing knowledge of these structures, which is a limitation in all alternative software. In Chapter 4, I utilised TRASH and developed complementary analyses for centromeric repeat discovery in plant genomes. In *Arabidopsis thaliana*, I identified *CEN178* repeats in the Columbia accession and described inter-centromere differences in their organisation. Additionally, I used 65 other *Arabidopsis thaliana* accessions to find centromere positions and their types, identify rearrangements and syntenic regions, describe centromeric repeat characteristics over the centromeric regions and identify centromeric transposon insertion patterns. In further analysis, *Arabidopsis lyrata* centromeres, despite relatedness, shared only superficial similarity in their location and repeat sequence with *Arabidopsis thaliana*. I also described two related centromeric tandem repeat families of *A. lyrata* that occupy distinct chromosome sets. Based on two *Brassica oleracea* species, I was able to confirm the division of their *CentBr* repeats into two subfamilies, consistent with previous reports using cytological methods (Lim et al. 2005, Fujii and Ohmido 2011). Insights into inter-chromosomal similarity can be a beginning point for Brassica-wide analysis of the A, B and C genome centromere evolution and related speciation. I identified centromeric arrays in the holocentric species of *Rhynchospora breviscula*, *R. pubera* and *R. tenuis*,

and found that they resemble monocentric species in terms of higher-order repeats accumulation, but not sequence identity distribution. Aside from the *in-silico* analysis, I have generated a transgenic line with tomato *HEI10* overexpression and showed its effect on meiotic recombination, described in Chapter 5. I also used pre-existing *Arabidopsis thaliana HEI10*-overexpressor lines to test whether crossover frequency saturates with progressively greater levels of overexpression

These results will be discussed separately, highlighting (i) the potential of TRASH as an analysis tool for the community in the era of long-read sequencing, (ii) insights into the centromere structures and evolution of plant species, including a model for repeat array expansion and contraction, and (iii) the potential of *HEI10* overexpression on inducing meiotic crossovers in the context of crop improvement. Future developments will be outlined for each of these sections.

6.1 TRASH is a robust tool for tandem repeat annotation and analysis

Tandem Repeat Annotation and Structural Hierarchy software was developed to assist in tandem repeat array description, which includes mapping individual repeats, their classification into repeat families, identifying higher-order repeats formed within the families, calculating divergence and HOR involvement of each repeat, and plotting found structures. TRASH has been designed with consideration of the peculiar biology of centromeric tandem repeats, and as such its main advantage is accurate delimitation of individual repeats and their description allowing for efficient downstream analysis, without prior knowledge of the repeats involved. The ability to adjust the repeat shift independently, such that repeats of the same family in a *de novo* setting are identified with the same relative start position is crucial in analysis of such short sequences. The output files facilitate downstream analysis, providing information on all identified repeats, all tandem repeat regions and HORs (when queried). Each tandem repeat is characterised by its family (when family templates are provided), divergence from the chromosome-wide consensus, involvement in HORs and its coordinates. These output formats allow for tallying of the results and rapid identification of variation within the arrays and secondary structures.

TRASH has a quick and easy installation and can facilitate parallel processing, which accelerates the analysis. It can provide output on medium size genomes like *Brassica oleracea* within a few hours. TRASH doesn't require any pre-existing software, other than an R installation. Additionally, it can be run with nothing other than an input fasta file location from a command line, and default parameters will generate adequate results to inform on the analysed genome's repeats. On the other hand, when required, over 20 parameters can be adjusted to tailor its behaviour to the user's specifications. These include alternative work modes and specific parameters related to the analysis itself.

Numerous software has been developed to identify and describe genome repetitive elements. They are mostly designed for specific repeat structures and reflect the available data at the time of their creation. Since microsatellite arrays could have been analysed using short-read data, a number of short tandem repeats (STR) identification software were developed, including SRF (Sharma et al. 2004), E-TRA (Karaca et al. 2005), pSTR (Chun-I Lee et al 2016), GangSTR (Mousavi et al 2019), STRique (Giesselmann et al. 2019), STRique (Giesselmann et al. 2019) and OSTRFPD (Mathemba, Dondorp and Imwong 2019). TRASH is able to identify repetitive regions regardless of their size, including two-nucleotide repeats and STRs. However, after identification of repetitive regions and their periods, mapping and polishing steps can misrepresent the very short monomers, which was found during some test runs. This could be addressed in the future by introducing an STR mode to TRASH, although with the large number of specific software available it is not a priority.

Maintenance of bioinformatic tools is unfortunately an issue with older software. TROLL (Castelo. Martins and Gao 2002), SRF (Sharma et al. 2004), E-TRA (Karaca et al. 2005), INVERTER (Wirawan et al 2010), or an unnamed software by Domanic and Preparate (2007), are no longer accessible through their published links, limiting their accessibility. The distribution model and installation simplicity, together with version control introduced into TRASH installation can help to promote its longevity. Creating a self-contained package with all required dependencies, for example a Docker image, could increase accessibility and limit problems with updates to software that TRASH relies upon, although these were purposefully minimised.

Even after 23 years, Tandem Repeat Finder remains a widely used tool for repeat identification, with 10% of all its citations coming from 2022 alone. TRF is versatile, simple to use and fast (Benson 1999 and Chapter 3). Its initial search for locally repeated *k*-mers is similar to TRASH identification of repetitive regions and their periodicity. While TRF processes these early matches in a heuristic manner, producing results that have to be filtered and validated, TRASH systematically identifies individual repeats and describes them, providing immediate feedback in an accessible manner using plots and summary tables. TRASH will also not create overlapping annotations, removing the need for data filtering. A TRF feature that could be implemented in TRASH is a statistical significance measure for the identified repeats, so that even less stringent settings of repeat identification can be set by default, while potential false positives are identified by their low significance.

RepeatModeler is a powerful tool capable of *de novo* annotating, not only transposable elements interspersed throughout the genomes, but also tandem repeats (Hubley and Smit 2008, Flynn et al. 2020). Together with pre-defined repeat libraries, it can provide comprehensive identification of elements present in the analysed genome, and coupled with RepeatMasker, these elements can be fully annotated throughout the genome (Hubley and Smit 2008, Flynn et al. 2020). Despite this potential, the analysis is time consuming and requires significant user input, while identified tandem repeats are not always properly characterised (Chapter 3). However, RepeatModeler provides accurate annotation of other repeated elements, often found in the proximity to tandem arrays (Chapter 3). This could be facilitated by TRASH by implementing identification of tandem array gaps and providing their sequence to RepeatModeler for their description. In *Arabidopsis thaliana* for example, this would presumably result in rapid identification of the centromeric *ATHILA* elements.

Higher order repeat identification has been a problematic task for bioinformatics software. Many approaches have been developed based on the specific characteristics of the human centromeres (Kunyavskaya et al. 2022, Bzikadze and Pevzner 2020, Gao et al. 2022). Unfortunately, these approaches rely on prior mapping of repeats and/or monomer definitions, which can be algorithmically complex processes which can limit their applicability. Developers of only one of the tools,

HiCAT, have attempted to analyse non-human centromeres (Gao et al. 2022). Interestingly, HiCAT is capable of annotating HORs in the *Arabidopsis thaliana* centromere 2 region, which was previously presented in Naish et al. 2021, and identifies a satellite octamer. Despite this, most of the genome was occupied by monomic expansions as reported by HiCAT, and the HOR pattern from chromosome 2 was the only case with HORs with more than 4 monomers of length (Gao et al. 2022). When related to the results presented in Chapter 4 and in Naish et al. 2021, it can be argued that monomer based HOR identification can be problematic in non-human species and not adequately informative. At the same time, HiCAT results cannot be disregarded, as they find human HORs with high accuracy compared to alternative methods (Gao et al. 2022). Therefore, the HOR identification method described in Chapter 3 may be more suitable for general analysis of tandem repeat arrays organisation, as shown in Chapter 4.

The biggest limitation of TRASH lies in its runtime when the HOR module is enabled. While repeat identification analysis has linear complexity, HORs are calculated based on a 2-dimensional matrix, hence an increase in repeats used for the analysis exponentially increases the runtime and required memory. However, this has been only a problem during TRASH development and testing with individual species containing over 50,000 repeats per chromosome.

Apart from the mentioned future developments of TRASH, potential additional improvements include alternative repeat mapping methods, potentially integrating HMMER into the TRASH workflow, as it proved to be very accurate when provided with a sequence template. An automatic mode could be introduced that would dynamically adjust run settings. For example, default 12-mers in a 1 kbp window could start as 30-mers in a 10 kbp window, to allow for initial approximate identification of repetitive regions, followed by an increase in the stringency of settings for subsequent annotation. This could enable identification of long (several kilobases) tandem repeats without sacrificing on the runtime required for analysis using large windows. Additionally, the TRASH repeat classification module can be expanded to make use of existing databases, and in *de novo* run settings, to automatically identify repeat families based on the most common monomer.

6.2 Plant centromeric repeats form various structures related to their evolutionary origins

Analysis of *Arabidopsis* centromeres showed extreme differences in *CEN178* repeat structure and sequence within and between centromeres, despite them being occupied by a single family of repeats. Together this suggested that homogenisation of repeats is limited to single chromosomes within the species. Indeed, a limited number of HORs have been identified between centromeres within individual species. Additionally, these HORs were mostly short and local, but without distinguishable HOR n-mers. Instead, HORs appeared to form using a range of short sizes. It is also uncertain whether HORs in *Arabidopsis* form as multiples of 178-bp regions, since a limited number of truncated repeats was identified. Ectopic gene conversion could explain homogenisation of repeats within the chromosomes, since repeats could provide a similar enough template, even during non-allelic recombination, and the mean size of higher-order repeats (HORs, duplications of several repeats) is similar to observed gene conversion tracts (unidirectional transfer of genetic material between homologous sequences) in *Arabidopsis* (Talbert and Henikoff 2010, Wijinker et al 2013).

Clear distinctions between centromeric groups were identified across 66 *A. thaliana* accessions. Since centromere groups can be shared between some, but not all centromeres, this suggests ancestral centromere outcrossing events, where different centromeres are exchanged through independent chromosome segregation during hybrid meiosis. To explain centromere evolution in *Arabidopsis thaliana*, I propose the existence of distinct centromeres, called centrotypes, that can be found across accessions (Fig. 6.1). These centrotypes can segregate, but not reciprocally crossover in hybrids, although allelic or non-allelic non-crossover pathways may act and contribute to sequence change. Each centrotype may actively homogenise its repeats through DSB formation and homologous recombination, which may preserve the repeat library of a centrotype and increase similarity within centromere groups. However, this homogenisation appears to be more effective in the central parts of the chromosomes, while centromere edges remain unchanged, causing them to maintain more ancestral similarity and to become more divergent from the central repeats over

time. This layered organisation, with more homogenised repeats in the core, and more diverged repeats on the edges, has been described in human centromeres (Logsdon et al. 2021, Nurk et al. 2022, Altemose et al. 2022). Existence of a high number of "orphan" centromeres in *Arabidopsis*, or centrotypes with a single representative, might suggest that new centrotypes arise relatively frequently through structural rearrangements and HOR accumulation or that we have yet to saturate our sampling of *Arabidopsis* centromere diversity.

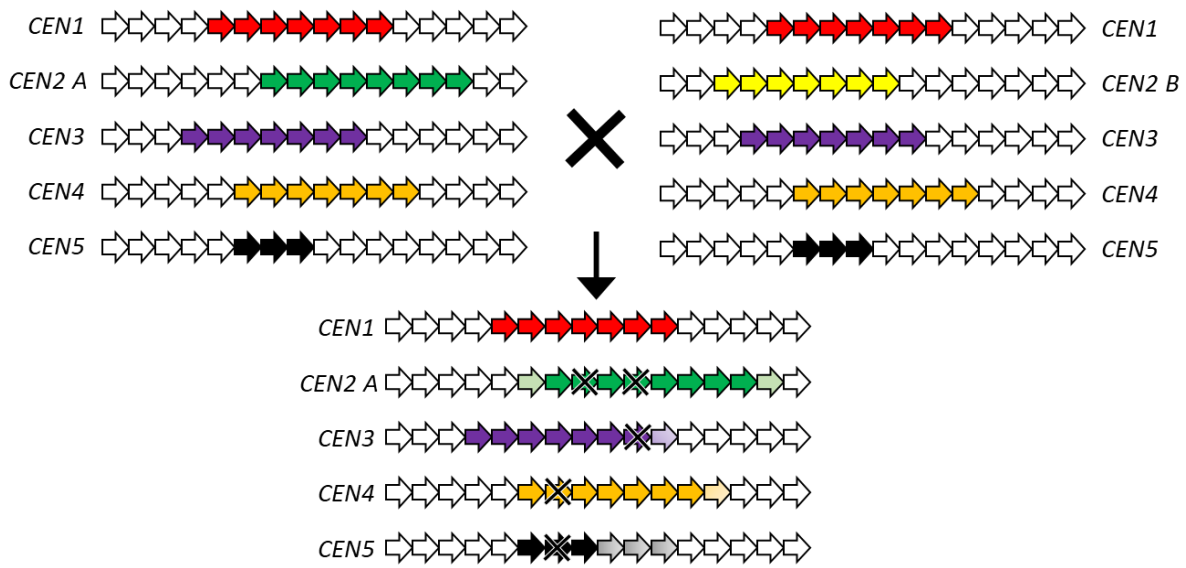


Figure 6.1 Model of centrotype evolution.
The five centromeres of Arabidopsis thaliana are represented as tandem repeat arrays. A small array is shown to represent the much larger arrays seen in the genome. Different centrotypes (A and B) of chromosome 2 in a hybrid setting coexist, but only one is preferentially transmitted due to centromere drive acting during asymmetric female meiosis. At the same time, active repeats that form HORs (coloured) contract and expand via DSB formation and homologous recombination repair pathways, which can eventually lead to creation of a distinct centrotype.

This rapid evolution of centromeric repeats might also be a direct response to a potentially destabilising threat of retrotransposon invasion. The presence of *ATHILA* transposable elements in regions evolving so rapidly, the high similarity between them, and between LTRs of single elements suggest that these mostly constitute recent insertions, further supported by the relatively low abundance of *ATHILA* in the

chromosome arms and their low LTR similarity (Nasih et al. 2021). At the same time, centromeres with the most HORs have the least *ATHILA* insertions, suggesting a transposon-purging effect of repeat homogenisation and ongoing antagonistic coevolution between *ATHILA* elements and their host centromere arrays. To perform further analysis of centromere strength, this could be evaluated by reciprocal crosses and measuring segregation distortion in the F_1 generation. Mouse experiments show that centromere strength depends upon the amount of deposited CENPA protein (Chamatal et al. 2014), therefore, to identify potential strong and weak centromeres, CENH3 ChIP-seq data could be obtained and used.

Available CENH3 ChIP-seq data shows that, like CENPA in other species, its deposition in Arabidopsis is constrained to certain parts of the centromeric repeats. To investigate neocentromere formation in Arabidopsis, information on the centromeric repeats sequence could be used to inform guide-RNA design for CRISPR-Cas9 centromere elimination studies. Based on neocentromere formation in other species and the diversity of Arabidopsis *CEN178* array sizes, removal of array parts correlated with CENH3 deposition could force centromere repositioning towards otherwise CENH3 depleted repeats. In the Col-0 accession, chromosome 4 could be used to remove the main CENH3-occupied centromeric array, Stable lines resulting from centromere knock-out would potentially contain neocentromeres, likely over the adjacent inactive array of chromosome 4. Such lines could then be back-crossed to wild type Col-0 to estimate the neocentromere strength against the native centromere by segregation distortion in subsequent generations.

The human kinetochore has been hypothesised to recruit recombination proteins that promote centromere satellite formation and recombination, termed the Kinetochore Associated Recombination Machine (Shepelev et al. 2009, Miga and Alexandrov 2021). A similar process could be present in *Arabidopsis* and promote *CEN178* recombination of the centromeres, contributing to centromere homogenization, and purging of ATHILA insertion. Interactions between *CEN178* HOR accumulating loci and CENH3 enrichment domains could be studied in the Tanz-0 accession, where greatest HOR abundance can be identified on the left side of the centromere 5, while CENH3 occupies the right side. It is conceivable that this is a result of a recent centromere

migration and if HORs accumulate over CENH3 enriched regions, propagation of Tanz-0 line over several generations and re-sequencing is expected to show accumulation of novel HORs in the right-side part of the centromere.

The existence of two closely related centromeric repeats in *Arabidopsis lyrata* (*CEN168* and *CEN179*) provides an opportunity to analyse the effect of repeat monomer sequences on centromere viability. *CEN168* and *CEN179* differ mostly by a single 11-bp domain that has been most likely lost in *CEN168*, considering *CEN179* and *CEN178* share sequence and length similarity. Analysis of natural variation in *Arabidopsis lyrata* centromeres could answer the question whether this is a transitional period in genome evolution, with one of the repeats being dominant, or it is a stable organisation and both repeats can coexist. Stable coexistence of these repeats is suggested by the fact that the diverged North American and Siberian lineages have maintained *CEN168* and *CEN179* distributions on the same chromosomes since they split.

Mapping of *Brassica oleracea* repeats established the localisation of the two subfamilies of *CentBr* repeats. Their distribution however is not identical between the two analysed subspecies. Particularly, centromere 3 in subspecies *Alboglabra* contains *CentBr2* repeats, while in subspecies *Italica* it is mostly composed of *CentBr1*. The existence of *CentBr2* repeats on most centromeres in their distal regions suggests that they could have occupied the whole centromeres in the past but have since been replaced by *CentBr1*. It is possible that with time, *CentBr1* would occupy all centromeres of *Brassica oleracea*, displacing the *CentBr1* repeats to the centromere edges, where they would resemble the *CEN159* repeats in *Arabidopsis thaliana*. With *Brassica* species being extensively studied, other genomic assemblies of A, B and C genomes are available (Rousseau-Gueutin et al. 2020, Paritosh et al. 2020, Paritosh, Pradhan and Pental 2020, Yang et al. 2022). This will allow for parallel study in closely related species and potentially will inform about centromere evolution in *Brassica* species. CENH3 ChIP-seq data would be also very valuable in the *Brassica oleracea* lines, to define binding preferences between the *CentBr* subfamilies and identify deposition sites along the centromeres, especially potential differences in CENH3 deposition on chromosome 3 in both subspecies. The CENH3 localisation would be

particularly interesting, since CENH3 and CENPA tend to occupy only a part of centromeric repeat arrays and most *CentBr* arrays also contain large retrotransposon insertions. The functional importance of this organisation could be analysed by specific deletion of the central transposon regions or parts of *CentBr* arrays using the CRISPR-Cas9 approaches described above.

Tyba repeats between the three analysed *Rhynchospora* species similar to one another, despite the observation that centromeric repeats of even closely related species are often very different - for example between *Arabidopsis thaliana* and *Arabidopsis lyrata*. This is also despite large differences in their genome size and organisation. Additionally, individual *Tyba* arrays do not form a layered organisation, like monocentric species analysed in Chapter 4, but retain the same level of similarity (as measured by edit distance to consensus) throughout individual arrays. This is regardless of the fact that HORs still can be predominantly found in the central parts of the arrays, although the total number of HORs is not as high as in the monocentric species. These differences in organisation could be a result of the achiasmatic meiosis, where homologous chromosomes do not have a chance to form physical connections, thus inhibiting any homolog exchange or recombination. Because of the similarities in centromeric repeats and peculiarities of *Rhynchospora* meiosis, it would be interesting to assess the viability of inter-species hybrids, when both species contain identical numbers of chromosomes of similar lengths.

6.3 HEI10 dosage effect on meiotic recombination can be translated into tomato

The gene coding *Arabidopsis thaliana* meiotic recombination dosage-sensitive regulator, *HEI10*, was identified in tomato and used to generate transgenic overexpressor lines to test its effect on a species with recombination largely constrained to distal regions of the chromosomes. Although much weaker than observed in *Arabidopsis*, a significant recombination increase was confirmed cytologically with chiasmata counting and genetic segregation using SSLP markers. The results from the SSLP experiment suggest that any increase in recombination is present in the distal regions and *HEI10* overexpression is not sufficient to overcome

heterochromatic suppression of meiotic recombination. Due to the failure of my experiments to map crossovers by sequencing, it has not been possible to test whether the closely spaced crossovers in the distal regions were increased. Genotyping-by-sequencing of tomato lines would additionally generate high-quality maps of recombination in the *HEI10*-overexpressing plants (Rowan et al. 2015, Ziolkowski et al. 2015). Potentially the level of HEI10 increase was not sufficient to create a significant increase in crossover numbers. Additionally, overexpression was only measured at the mRNA level and protein quantification might be required.

A possible saturation effect of *HEI10* overexpression was investigated by combining independent *Arabidopsis thaliana HEI10 transgenic lines*. I showed that recombination frequency in the *420* interval can reach 50 cM, theoretically unlinking two alleles from the same chromosome. This could be already achieved after crossing two lines with high cM values and propagating to the F_2 generation Additionally, decrease in recombination of plants reaching 50 cM can be attributed to additional crossovers leading to an increase in occurrence of double crossovers, which preserve linkage between the *420* markers, effectively decreasing cM values measured in this interval. If 50% of meiotic recombination events resulted in a single crossover and remaining events in double crossovers, the measured cM value would be ~25%, despite an average of 1.5 crossovers per meiosis. This decrease in recombination between two loci when additional crossovers are introduced can be mitigated by crossover interference, which was found to increase with increasing *HEI10* dosage (Morgan et al. 2021). In the HEI10 coarsening model proposed by Morgan et al. (2021), HEI10 is initially deposited throughout the chromosome lengths at recombination intermediate foci, which later compete for a limited quantity of HEI10, with stronger foci accumulating more protein at the cost of the weaker foci. This effect is proposed to happen mostly over short distances; hence crossover interference decreases with the distance from a crossover. A higher dosage of the HEI10 protein might lead to additional recombination intermediate foci to maturate into crossovers, thus explaining the dosage effect (Morgan et al. 2021).

6.4 Final comments

This study provides insight into the organisation of centromeric arrays of plant species and *HEI10*-mediated increase of recombination in tomato and *Arabidopsis*. Tandem Repeat Identification and Structural Hierarchy (TRASH) software was designed to bridge the gap between current long-read sequencing based genomic assemblies and our understanding of the tandem repeat arrays. It is anticipated that TRASH will be valuable to the community as its capabilities were demonstrated by results achieved in Chapter 4. Detailed analysis of *Arabidopsis thaliana* accessions is the first of its kind where pan-centromere of a single species has been described. Genomic landscapes of *Arabidopsis lyrata*, *Brassica oleracea* and *Rhynchospora* species additionally demonstrate unique peculiarities of their organisation, and it is expected that with each new species analysed, new details of plant centromere evolution will be uncovered. Therefore, a large study that combines all available high-quality genomes harbouring centromeric tandem arrays and systematically analyses them would be an appealing approach to uncover commonalities and differences in centromere organisation and to potentially identify conserved mechanisms leading to their evolution, despite widely varying structure. *HEI10* overexpression, while interesting from the point of recombination modulation, requires further studies that could couple it with interference abolishing mutations to be applicable for heterochromatin recombination increase.

Bibliography

Aftab, Tariq, and Khalid Rehman Hakeem, eds. 2020. Plant Micronutrients: Deficiency and Toxicity Management. Cham: Springer International Publishing. http://link.springer.com/10.1007/978-3-030-49856-6 (December 30, 2022).

Agarwal, Seema, and G.Shirleen Roeder. 2000. 'Zip3 Provides a Link between Recombination Enzymes and Synaptonemal Complex Proteins'. Cell 102(2): 245–55.

Ågren, J et al. 2014. 'Mating System Shifts and Transposable Element Evolution in the Plant Genus Capsella'. BMC Genomics 15(1): 602.

Ahmad, Kami, and Steven Henikoff. 2001. 'Centromeres Are Specialized Replication Domains in Heterochromatin'. Journal of Cell Biology 153(1): 101–10.

Akera, Takashi et al. 2017. 'Spindle Asymmetry Drives Non- Mendelian Chromosome Segregation'.

Akera, Takashi, Emily Trimm, and Michael A. Lampson. 2019. 'Molecular Strategies of Meiotic Cheating by Selfish Centromeres'. Cell 178(5): 1132-1144.e10.

Aldrup-MacDonald, Megan E. et al. 2016. 'Genomic Variation within Alpha Satellite DNA Influences Centromere Location on Human Chromosomes with Metastable Epialleles'. Genome Research 26(10): 1301–11.

Alexandrov, Ivan et al. 2001. 'Alpha-Satellite DNA of Primates: Old and New Families'. Chromosoma 110(4): 253–66.

Allipra, Sreejith et al. 2022. 'The Kinetochore Protein NNF1 Has a Moonlighting Role in the Vegetative Development of Arabidopsis Thaliana'. The Plant Journal 109(5): 1064–85.

Allshire, Robin C., and Gary H. Karpen. 2008. 'Epigenetic Regulation of Centromeric Chromatin: Old Dogs, New Tricks?' Nature Reviews Genetics 9(12): 923–37.

Altemose, Nicolas et al. 2022. 'Complete Genomic and Epigenetic Maps of Human Centromeres'. Science 376(6588): eabl4178.

Amarasinghe, Shanika L. et al. 2020. 'Opportunities and Challenges in Long-Read Sequencing Data Analysis'. Genome Biology 21(1): 30.

Anderson, Matthew Z., Gregory J. Thomson, Matthew P. Hirakawa, and Richard J. Bennett. 2019. 'A "Parameiosis" Drives Depolyploidization and Homologous Recombination in Candida Albicans'. Nature Communications 10(1): 4388.

Anderson, Sarah N. et al. 2019. 'Transposable Elements Contribute to Dynamic Genome Content in Maize'. The Plant Journal 100(5): 1052–65.

Arkhipova, Irina R. 2017. 'Using Bioinformatic and Phylogenetic Approaches to Classify Transposable Elements and Understand Their Complex Evolutionary Histories'. Mobile DNA 8(1): 19.

Armstrong, S. J., and G. H. Jones. 2003. 'Meiotic Cytology and Chromosome Behaviour in Wild-Type Arabidopsis Thaliana'. Journal of Experimental Botany 54(380): 1–10.

Armstrong, Susan J., Anthony P. Caryl, Gareth H. Jones, and F. Christopher H. Franklin. 2002. 'Asy1, a Protein Required for Meiotic Chromosome Synapsis, Localizes to Axis-Associated Chromatin in Arabidopsis and Brassica'. Journal of Cell Science 115(18): 3645–55.

Arunkumar, Ganesan, and Daniël P. Melters. 2020. 'Centromeric Transcription: A Conserved Swiss-Army Knife'. Genes 11(8): 911.

Aze, Antoine et al. 2016. 'Centromeric DNA Replication Reconstitution Reveals DNA Loops and ATR Checkpoint Suppression'. Nature Cell Biology 18(6): 684–91.

Bachrati, C. Z. 2006. 'Mobile D-Loops Are a Preferred Substrate for the Bloom's Syndrome Helicase'. Nucleic Acids Research 34(8): 2269–79.

Baduel, Pierre et al. 2021. 'Genetic and Environmental Modulation of Transposition Shapes the Evolutionary Potential of Arabidopsis Thaliana'. Genome Biology 22(1): 138.

Baduel, Pierre, and Vincent Colot. 2021. 'The Epiallelic Potential of Transposable Elements and Its Evolutionary Significance in Plants'. Philosophical Transactions of the Royal Society B: Biological Sciences 376(1826): 20200123.

Baduel, Pierre, and Leandro Quadrana. 2021. 'Jumpstarting Evolution: How Transposition Can Facilitate Adaptation to Rapid Environmental Changes'. Current Opinion in Plant Biology 61: 102043.

Barbosa-Cisneros, Olga, and Rafael Herrera-Esparza. 2002. 'CENP-B Is a Conserved Gene among Vegetal Species'. Genetics and Molecular Research.

Barnhart, Meghan C. et al. 2011. 'HJURP Is a CENP-A Chromatin Assembly Factor Sufficient to Form a Functional de Novo Kinetochore'. Journal of Cell Biology 194(2): 229–43.

Baum, M P, and L Clarke. 1990. 'Functional Analysis of a Centromere from Fission Yeast: A Role for Centromere-Specific Repeated DNA Sequences'. MOL. CELL. BIOL. 10.

Becker, Claude et al. 2011. 'Spontaneous Epigenetic Variation in the Arabidopsis Thaliana Methylome'. Nature 480(7376): 245–49.

Benson, G. 1999. 'Tandem Repeats Finder: A Program to Analyze DNA Sequences'. Nucleic Acids Research 27(2): 573–80.

Berchowitz, Luke E, Kirk E Francis, Alexandra L Bey, and Gregory P Copenhaver. 2007. 'The Role of AtMUS81 in Interference-Insensitive Crossovers in A. Thaliana' ed. Jonathan Pritchard. PLoS Genetics 3(8): e132.

Bergmann, Jan H. et al. 2012. 'HACking the Centromere Chromatin Code: Insights from Human Artificial Chromosomes'. Chromosome Research 20(5): 505–19.

Berry, Charles, Sridhar Hannenhalli, Jeremy Leipzig, and Frederic D. Bushman. 2006. 'Selection of Target Sites for Mobile DNA Integration in the Human Genome'. PLoS Computational Biology 2(11): e157.

Bestor, T H, and D Bourc'His. 'Transposon Silencing and Imprint Establishment in Mammalian Germ Cells'.

Bidau, C.J., and D.A. Martí. 2004. 'B Chromosomes and Robertsonian Fusions of Dichroplus Pratensis (Acrididae): Intraspecific Support for the Centromeric Drive Theory'. Cytogenetic and Genome Research 106(2–4): 347–50.

Bishop, K, and Nancy Kleckner. 'DMC1: A Meiosis-Specific Yeast Homolog of E. Coli RecA Required for Recombination, Synaptonemal Complex Formation, and Cell Cycle Progression'.

Black, Ben E., ed. 2017. 56 Centromeres and Kinetochores. Cham: Springer International Publishing. http://link.springer.com/10.1007/978-3-319-58592-5 (December 30, 2022).

Black, Ben E., and Don W. Cleveland. 2011. 'Epigenetic Centromere Propagation and the Nature of CENP-A Nucleosomes'. Cell 144(4): 471–79.

Blackwell, Alexander R et al. 2020. 'MSH 2 Shapes the Meiotic Crossover Landscape in Relation to Interhomolog Polymorphism in Arabidopsis'. The EMBO Journal 39(21). https://onlinelibrary.wiley.com/doi/10.15252/embj.2020104858 (December 30, 2022).

Blower, Michael D. 2016. 'Centromeric Transcription Regulates Aurora-B Localization and Activation'. Cell Reports 15(8): 1624–33.

Bobkov, Georg O.M., Nick Gilbert, and Patrick Heun. 2018. 'Centromere Transcription Allows CENP-A to Transit from Chromatin Association to Stable Incorporation'. Journal of Cell Biology 217(6): 1957–72.

Bodor, Dani L et al. 2014. 'The Quantitative Architecture of Centromeric Chromatin'. eLife 3: e02137.

Bolcun-Filas, Ewelina, and Mary Ann Handel. 2018. 'Meiosis: The Chromosomal Foundation of Reproduction'. Biology of Reproduction 99(1): 112–26.

Bourque, Guillaume et al. 2018. 'Ten Things You Should Know about Transposable Elements'. Genome Biology 19(1): 199.

Bouzinba-Segard, Haniaa, Adeline Guais, and Claire Francastel. ‘Accumulation of Small Murine Minor Satellite Transcripts Leads to Impaired Centromeric Architecture and Function’. CELL BIOLOGY.

Bowen, Nathan J. et al. 2003. ‘Retrotransposons and Their Recognition of Pol II Promoters: A Comprehensive Survey of the Transposable Elements From the Complete Genome Sequence of Schizosaccharomyces Pombe’. Genome Research 13(9): 1984–97.

Bozek, Monika et al. 2012. ‘Chromosome and Genome Size Variation in Luzula (Juncaceae), a Genus with Holocentric Chromosomes: Chromosome and C-Value Evolution in L Uzula’. Botanical Journal of the Linnean Society 170(4): 529–41.

Britton-Davidian, Janice et al. 2000. ‘Rapid Chromosomal Evolution in Island Mice’. Nature 403(6766): 158–158.

Britton-Davidian, Janice et al. 2005. ‘Chromosomal Phylogeny of Robertsonian Races of the House Mouse on the Island of Madeira: Testing between Alternative Mutational Processes’. Genetical Research 86(3): 171–83.

Bureš, Petr, and František Zedek. 2014. ‘HOLOKINETIC DRIVE: CENTROMERE DRIVE IN CHROMOSOMES WITHOUT CENTROMERES: BRIEF COMMUNICATION’. Evolution: n/a-n/a.

Burke, W D, C C Calalang, and T H Eickbush. ‘The Site-Specific Ribosomal Insertion Element Type II of Bombyx Mori (R2Bm) Contains the Coding Sequence for a Reverse Transcriptase-like Enzyme’. MOL. CELL. BIOL.

Bzikadze, Andrey V., and Pavel A. Pevzner. 2020. ‘Automated Assembly of Centromeres from Ultra-Long Error-Prone Reads’. Nature Biotechnology 38(11): 1309–16.

Cabral, Gabriela et al. 2014. ‘Chiasmatic and Achiasmatic Inverted Meiosis of Plants with Holocentric Chromosomes’. Nature Communications 5(1): 5070.

Camacho, Juan Pedro M., Timothy F. Sharbel, and Leo W. Beukeboom. 2000. ‘B-Chromosome Evolution’. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 355(1394): 163–78.

Capilla-Pérez, Laia et al. 2021. ‘The Synaptonemal Complex Imposes Crossover Interference and Heterochiasmy in Arabidopsis’. Proceedings of the National Academy of Sciences 118(12): e2023613118.

Casola, C., D. Hucks, and C. Feschotte. 2007. ‘Convergent Domestication of Pogo-like Transposases into Centromere-Binding Proteins in Fission Yeast and Mammals’. Molecular Biology and Evolution 25(1): 29–41.

Cesari, Michele et al. 2003. ‘Polymerase Chain Reaction Amplification of the Bag320 Satellite Family Reveals the Ancestral Library and Past Gene Conversion Events in Bacillus Rossius (Insecta Phasmatodea)’. Gene 312: 289–95.

Chan, F. Lyn et al. 2012. ‘Active Transcription and Essential Role of RNA Polymerase II at the Centromere during Mitosis’. Proceedings of the National Academy of Sciences 109(6): 1979–84.

Chang, Paul, and Ryoma Ohi, eds. 2016. 1413 The Mitotic Spindle: Methods and Protocols. New York, NY: Springer New York. http://link.springer.com/10.1007/978-1-4939-3542-0 (December 30, 2022).

Charlesworth, B, and C H Langley. 1986. ‘THE EVOLUTION OF SELF-REGULATED TRANSPOSITION OF TRANSPOSABLE ELEMENTS’. Genetics 112(2): 359–83.

Cheeseman, Iain M., and Arshad Desai. 2008. ‘Molecular Architecture of the Kinetochore–Microtubule Interface’. Nature Reviews Molecular Cell Biology 9(1): 33–46.

Chelysheva, L. et al. 2010. ‘An Easy Protocol for Studying Chromatin and Recombination Protein Dynamics during ArabidopsisThaliana Meiosis: Immunodetection of Cohesins, Histones and MLH1’. Cytogenetic and Genome Research 129(1–3): 143–53.

Chelysheva, Liudmila et al. 2007. ‘Zip4/Spo22 Is Required for Class I CO Formation but Not for Synapsis Completion in Arabidopsis Thaliana’ ed. R. Scott Hawley. PLoS Genetics 3(5): e83.

Chelysheva, Liudmila et al. 2012. ‘The Arabidopsis HEI10 Is a New ZMM Protein Related to Zip3’ ed. F. Chris H. Franklin. PLoS Genetics 8(7): e1002799.

Chmátal, Lukáš et al. 2014. ‘Centromere Strength Provides the Cell Biological Basis for Meiotic Drive and Karyotype Evolution in Mice’. Current Biology 24(19): 2295–2300.

Chmátal, Lukáš, Karren Yang, Richard M. Schultz, and Michael A. Lampson. 2015. 'Spatial Regulation of Kinetochore Microtubule Attachments by Destabilization at Spindle Poles in Meiosis I'. Current Biology 25(14): 1835–41.

Choi, Eun Shik et al. 2012. 'Factors That Promote H3 Chromatin Integrity during Transcription Prevent Promiscuous Deposition of CENP-ACnp1 in Fission Yeast' ed. Beth A. Sullivan. PLoS Genetics 8(9): e1002985.

Choi, Eun Shik, Youngseo Cheon, Keunsoo Kang, and Daeyoup Lee. 2017. 'The Ino80 Complex Mediates Epigenetic Centromere Propagation via Active Removal of Histone H3'. Nature Communications 8(1): 529.

Choi, Kyuha et al. 2013. 'Arabidopsis Meiotic Crossover Hot Spots Overlap with H2A.Z Nucleosomes at Gene Promoters'. Nature Genetics 45(11): 1327–36.

Choi, Kyuha et al. 2018. 'Nucleosomes and DNA Methylation Shape Meiotic DSB Frequency in Arabidopsis Thaliana Transposons and Gene Regulatory Regions'. Genome Research 28(4): 532–46.

Chowdhury, Reshmi et al. 2009. 'Genetic Analysis of Variation in Human Meiotic Recombination' ed. Gregory P. Copenhaver. PLoS Genetics 5(9): e1000648.

Cifuentes, Marta et al. 2013. 'Haploid Meiosis in Arabidopsis: Double-Strand Breaks Are Formed and Repaired but Without Synapsis and Crossovers' ed. Tai Wang. PLoS ONE 8(8): e72431.

Cole, Francesca, Scott Keeney, and Maria Jasin. 2010. 'Comprehensive, Fine-Scale Dissection of Homologous Recombination Outcomes at a Hot Spot in Mouse Meiosis'. Molecular Cell 39(5): 700–710.

Cole, Hope A., Bruce H. Howard, and David J. Clark. 2011. 'The Centromeric Nucleosome of Budding Yeast Is Perfectly Positioned and Covers the Entire Centromere'. Proceedings of the National Academy of Sciences 108(31): 12687–92.

Comai, Luca, Shamoni Maheshwari, and Mohan P A Marimuthu. 2017. 'Plant Centromeres'. Current Opinion in Plant Biology 36: 158–67.

Cooper, Jennifer L., and Steven Henikoff. 2004. 'Adaptive Evolution of the Histone Fold Domain in Centromeric Histones'. Molecular Biology and Evolution 21(9): 1712–18.

Copenhaver, G P, E A Housworth, and F W Stahl. 2002. 'Crossover Interference in Arabidopsis'. Genetics 160(4): 1631–39.

Copenhaver, Gregory P, and Daphne Preuss. 1999. 'Centromeres in the Genomic Era: Unraveling Paradoxes'. Current Opinion in Plant Biology 2(2): 104–8.

Cosby, Rachel L, Ni-Chen Chang, and Cédric Feschotte. 2019. 'Host–Transposon Interactions: Conflict, Cooperation, and Cooption'. Genes & Development 33(17–18): 1098–1116.

Crismani, Wayne et al. 2012. 'FANCM Limits Meiotic Crossovers'. Science 336(6088): 1588–90.

Da Ines, Olivier et al. 2012. 'Differing Requirements for RAD51 and DMC1 in Meiotic Pairing of Centromeres and Chromosome Arms in Arabidopsis Thaliana' ed. Wojciech P. Pawlowski. PLoS Genetics 8(4): e1002636.

Da Ines, Olivier, and Charles I. White. 2015. 'Centromere Associations in Meiotic Chromosome Pairing'. Annual Review of Genetics 49(1): 95–114.

Darling, Aaron C.E., Bob Mau, Frederick R. Blattner, and Nicole T. Perna. 2004. 'Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements'. Genome Research 14(7): 1394–1403.

De Muyt, Arnaud et al. 2007. 'AtPRD1 Is Required for Meiotic Double Strand Break Formation in Arabidopsis Thaliana'. The EMBO Journal 26(18): 4126–37.

De Muyt, Arnaud et al. 2009. 'A High Throughput Genetic Screen Identifies New Early Meiotic Recombination Functions in Arabidopsis Thaliana' ed. Gregory P. Copenhaver. PLoS Genetics 5(9): e1000654.

De Muyt, Arnaud et al. 2014. 'E3 Ligase Hei10: A Multifaceted Structure-Based Signaling Molecule with Roles within and beyond Meiosis'. Genes & Development 28(10): 1111–23.

de Massy, Bernard. 2013. 'Spp1 Links Sites of Meiotic DNA Double-Strand Breaks to Chromosome Axes'. Molecular Cell 49(1): 3–5.

Devine, Scott E, and Jef D Boeke. ‘Integration of the Yeast Retrotransposon Tyl Is Targeted to Regions Upstream of Genes Transcribed by RNA Polymerase III’.

Dion, Éric, Liangliang Li, Martine Jean, and François Belzile. 2007. ‘An Arabidopsis MLH1 Mutant Exhibits Reproductive Defects and Reveals a Dual Role for This Gene in Mitotic Recombination: Dual Role of AtMLH1 in Recombination’. The Plant Journal 51(3): 431–40.

Dong, Fenggao, and Jiming Jiang. ‘Non-Rabl Patterns of Centromere and Telomere Distribution in the Interphase Nuclei of Plant Cells’.

Dooner, Hugo, James English, Edward Ralston, and Edward Weck. 1986. ‘A Single Genetic Unit Specifies Two Transposition Functions in the Maize Element Activator’. Science 234(4773): 210–11.

Dover G. 1986. ‘Molecular Drive in Multigene Families: How Biological Novelties Arise, Spread and Are Assimilated’. TIG.

Dover, Gabriel. 1982. ‘Molecular Drive: A Cohesive Mode of Species Evolution’. Nature 299(5879): 111–17.

Drouaud, Jan et al. 2013. ‘Contrasted Patterns of Crossover and Non-Crossover at Arabidopsis Thaliana Meiotic Recombination Hotspots’ ed. Hong Ma. PLoS Genetics 9(11): e1003922.

Du, Yaqing, Christopher N. Topp, and R. Kelly Dawe. 2010. ‘DNA Binding of Centromere Protein C (CENPC) Is Stabilized by Single-Stranded RNA’ ed. Gregory P. Copenhaver. PLoS Genetics 6(2): e1000835.

Dubin, Manu J et al. 2015. ‘DNA Methylation in Arabidopsis Has a Genetic Basis and Shows Evidence of Local Adaptation’. eLife 4: e05255.

Dunleavy, Elaine M. et al. 2009. ‘HJURP Is a Cell-Cycle-Dependent Maintenance and Deposition Factor of CENP-A at Centromeres’. Cell 137(3): 485–97.

Dunleavy, Elaine M. et al. 2012. ‘The Cell Cycle Timing of Centromeric Chromatin Assembly in Drosophila Meiosis Is Distinct from Mitosis Yet Requires CAL1 and CENP-C’ ed. David M. Glover. PLoS Biology 10(12): e1001460.

Duret, Laurent, Gabriel Marais, and Christian Biémont. 2000. ‘Transposons but Not Retrotransposons Are Located Preferentially in Regions of High Recombination Rate in Caenorhabditis Elegans’. Genetics 156(4): 1661–69.

Dvorkina, Tatiana et al. 2021. ‘CentromereArchitect: Inference and Analysis of the Architecture of Centromeres’. Bioinformatics 37(Supplement_1): i196–204.

Eickbush, T. 2002. ‘Fruit Flies and Humans Respond Differently to Retrotransposons’. Current Opinion in Genetics & Development 12(6): 669–74.

Eickbush, Thomas H, and Danna G Eickbush. ‘Integration, Regulation, and Long-Term Stability of R2 Retrotransposons’.

Eickbush, Thomas H., and Varuni K. Jamburuthugoda. 2008. ‘The Diversity of Retrotransposons and the Properties of Their Reverse Transcriptases’. Virus Research 134(1–2): 221–34.

Evgen’ev, Michael B. et al. 1997. ‘Penelope , a New Family of Transposable Elements and Its Possible Role in Hybrid Dysgenesis in Drosophila Virilis’. Proceedings of the National Academy of Sciences 94(1): 196–201.

Fachinetti, Daniele et al. 2013. ‘A Two-Step Mechanism for Epigenetic Specification of Centromere Identity and Function’. Nature Cell Biology 15(9): 1056–66.

Falque, Matthieu et al. 2010. ‘Two Types of Meiotic Crossovers Coexist in Maize’. The Plant Cell 21(12): 3915–25.

Fang, Yufeng et al. 2020. ‘Long Transposon-Rich Centromeres in an Oomycete Reveal Divergence of Centromere Features in Stramenopila-Alveolata-Rhizaria Lineages’ ed. Rachel J. O’Neill. PLOS Genetics 16(3): e1008646.

Faulkner, Geoffrey J., and Piero Carninci. 2009. ‘Altruistic Functions for Selfish DNA’. Cell Cycle 8(18): 2895–2900.

Feng, Qinghua, John V Moran, Haig H Kazazian, and Jef D Boeke. 1996. ‘Human L1 Retrotransposon Encodes a Conserved Endonuclease Required for Retrotransposition’. Cell 87(5): 905–16.

Ferdous, Maheen et al. 2012. 'Inter-Homolog Crossing-Over and Synapsis in Arabidopsis Meiosis Are Dependent on the Chromosome Axis Protein AtASY3' ed. R. Scott Hawley. PLoS Genetics 8(2): e1002507.

Fernandes, Joiselle B, Piotr Wlodzimierz, and Ian R Henderson. 2019. 'Meiotic Recombination within Plant Centromeres'. Current Opinion in Plant Biology 48: 26–35.

Fernandes, Joiselle Blanche et al. 2018. 'FIGL1 and Its Novel Partner FLIP Form a Conserved Complex That Regulates Homologous Recombination' ed. Michael Lichten. PLOS Genetics 14(4): e1007317.

Ferree, Patrick M., and Satyaki Prasad. 2012. 'How Can Satellite DNA Divergence Cause Reproductive Isolation? Let Us Count the Chromosomal Ways'. Genetics Research International 2012: 1–11.

Ferri, Federica et al. 2009. 'Non-Coding Murine Centromeric Transcripts Associate with and Potentiate Aurora B Kinase'. Nucleic Acids Research 37(15): 5071–80.

Finn, R. D., J. Clements, and S. R. Eddy. 2011. 'HMMER Web Server: Interactive Sequence Similarity Searching'. Nucleic Acids Research 39(suppl): W29–37.

Finseth, Findley R., Yuzhu Dong, Arpiar Saunders, and Lila Fishman. 2015. 'Duplication and Adaptive Evolution of a Key Centromeric Protein in Mimulus , a Genus with Female Meiotic Drive'. Molecular Biology and Evolution 32(10): 2694–2706.

Finseth, Findley R., Thomas C. Nelson, and Lila Fishman. 2021. 'Selfish Chromosomal Drive Shapes Recent Centromeric Histone Evolution in Monkeyflowers' ed. Harmit S. Malik. PLOS Genetics 17(4): e1009418.

Fishman, Lila, and John H Willis. 2005. 'A Novel Meiotic Drive Locus Almost Completely Distorts Segregation in Mimulus (Monkeyflower) Hybrids'. Genetics 169(1): 347–53.

Fishman, Lila, and John H. Willis. 2008. 'Pollen Limitation and Natural Selection on Floral Characters in the Yellow Monkeyflower, Mimulus Guttatus'. New Phytologist 177(3): 802–10.

Flynn, Jullien M. et al. 2020. 'RepeatModeler2 for Automated Genomic Discovery of Transposable Element Families'. Proceedings of the National Academy of Sciences 117(17): 9451–57.

Foltz, Daniel R. et al. 2006. 'The Human CENP-A Centromeric Nucleosome-Associated Complex'. Nature Cell Biology 8(5): 458–69.

France, Martin G. et al. 2021. 'ZYP1 Is Required for Obligate Cross-over Formation and Cross-over Interference in Arabidopsis'. Proceedings of the National Academy of Sciences 118(14): e2021671118.

Francis, Kirk E. et al. 2007. 'Pollen Tetrad-Based Visual Assay for Meiotic Recombination in Arabidopsis'. Proceedings of the National Academy of Sciences 104(10): 3913–18.

French, Bradley T, and Aaron F Straight. 2019. 'CDK Phosphorylation of Xenopus Laevis M18 BP 1 Promotes Its Metaphase Centromere Localization'. The EMBO Journal 38(4). https://onlinelibrary.wiley.com/doi/10.15252/embj.2018100093 (December 30, 2022).

Fu, Shulan et al. 2013. 'De Novo Centromere Formation on a Chromosome Fragment in Maize'. Proceedings of the National Academy of Sciences 110(15): 6033–36.

Fuentes, Roven Rommel, Dick de Ridder, Aalt D J van Dijk, and Sander A Peters. 2022. 'Domestication Shapes Recombination Patterns in Tomato' ed. Michael Purugganan. Molecular Biology and Evolution 39(1): msab287.

Fujita, Yohta et al. 2007. 'Priming of Centromere for CENP-A Recruitment by Human HMis18α, HMis18β, and M18BP1'. Developmental Cell 12(1): 17–30.

Fukagawa, Tatsuo, and William C. Earnshaw. 2014. 'The Centromere: Chromatin Foundation for the Kinetochore Machinery'. Developmental Cell 30(5): 496–508.

Furuyama, Suzanne, and Sue Biggins. 2007. 'Centromere Identity Is Specified by a Single Centromeric Nucleosome in Budding Yeast'. Proceedings of the National Academy of Sciences 104(37): 14706–11.

Gamba, Riccardo, and Daniele Fachinetti. 2020. 'From Evolution to Function: Two Sides of the Same CENP-B Coin?' Experimental Cell Research 390(2): 111959.

Gao, Bo et al. 2020. 'Evolution of Pogo, a Separate Superfamily of IS630-Tc1-Mariner Transposons, Revealing Recurrent Domestication Events in Vertebrates'. Mobile DNA 11(1): 25.

Gao, Shenghan et al. 2022. HiCAT: A Tool for Automatic Annotation of Centromere Structure. Bioinformatics. preprint. http://biorxiv.org/lookup/doi/10.1101/2022.08.07.502881 (December 30, 2022).

Garrido-Ramos, Manuel. 2017. 'Satellite DNA: An Evolving Topic'. Genes 8(9): 230.

Geiss, Christian P. et al. 2014. 'CENP-A Arrays Are More Condensed than Canonical Arrays at Low Ionic Strength'. Biophysical Journal 106(4): 875–82.

Genet, Am J Hum. 'Chromosome-Specific Organization of Human Alpha Satellite DNA'.

Gent, Jonathan I et al. 2011. 'Distinct Influences of Tandem Repeats and Retrotransposons on CENH3 Nucleosome Positioning'. Epigenetics & Chromatin 4(1): 3.

Gent, Jonathan I., Na Wang, and R. Kelly Dawe. 2017. 'Stable Centromere Positioning in Diverse Sequence Contexts of Complex and Satellite Centromeres of Maize and Wild Relatives'. Genome Biology 18(1): 121.

Gershman, Ariel et al. 2022. 'Epigenetic Patterns in a Complete Human Genome'. Science 376(6588): eabj5089.

Gerton, Jennifer L., and R. Scott Hawley. 2005. 'Homologous Chromosome Interactions in Meiosis: Diversity amidst Conservation'. Nature Reviews Genetics 6(6): 477–87.

Girard, Chloe et al. 2015. 'AAA-ATPase FIDGETIN-LIKE 1 and Helicase FANCM Antagonize Meiotic Crossovers by Distinct Mechanisms' ed. Michael Lichten. PLOS Genetics 11(7): e1005369.

Glöckner, Gernot, and Andrew J. Heidel. 2009. 'Centromere Sequence and Dynamics in Dictyostelium Discoideum'. Nucleic Acids Research 37(6): 1809–16.

Gong, Zhiyun et al. 2012. 'Repeatless and Repeat-Based Centromeres in Potato: Implications for Centromere Evolution'. The Plant Cell 24(9): 3559–74.

Gorelick, Root, Jessica Carpinone, and Lindsay Jackson Derraugh. 2016. 'No Universal Differences between Female and Male Eukaryotes: Anisogamy and Asymmetrical Female Meiosis'. Biological Journal of the Linnean Society. https://academic.oup.com/biolinnean/article-lookup/doi/10.1111/bij.12874 (December 30, 2022).

Grabundzija, Ivana, Alison B. Hickman, and Fred Dyda. 2018. 'Helraiser Intermediates Provide Insight into the Mechanism of Eukaryotic Replicative Transposition'. Nature Communications 9(1): 1278.

Grandbois, Michel et al. 1999. 'How Strong Is a Covalent Bond?' Science 283(5408): 1727–30.

Greenfeder, S A, and C S Newlon. 1992. 'Replication Forks Pause at Yeast Centromeres'. MOL. CELL. BIOL. 12.

Guenatri, Mounia, Delphine Bailly, Christèle Maison, and Geneviève Almouzni. 2004. 'Mouse Centric and Pericentric Satellite Repeats Form Distinct Functional Heterochromatin'. Journal of Cell Biology 166(4): 493–505.

Guérin, Thomas M., and Stéphane Marcand. 2022. 'Breakage in Breakage–Fusion–Bridge Cycle: An 80-Year-Old Mystery'. Trends in Genetics 38(7): 641–45.

Gutbrod, Michael J., and Robert A. Martienssen. 2020. 'Conserved Chromosomal Functions of RNA Interference'. Nature Reviews Genetics 21(5): 311–31.

Hajra, Sujata, Santanu Kumar Ghosh, and Makkuni Jayaram. 2006. 'The Centromere-Specific Histone Variant Cse4p (CENP-A) Is Essential for Functional Chromatin Architecture at the Yeast 2-Mm Circle Partitioning Locus and Promotes Equal Plasmid Segregation'. Journal of Cell Biology 174(6): 779–90.

Hall, Sarah E., Gregory Kettler, and Daphne Preuss. 2003. 'Centromere Satellites From Arabidopsis Populations: Maintenance of Conserved and Variable Domains'. Genome Research 13(2): 195–205.

Harrington, John J. et al. 1997. 'Formation of de Novo Centromeres and Construction of First-Generation Human Artificial Microchromosomes'. Nature Genetics 15(4): 345–55.

Hartung, Frank et al. 2007. 'The Catalytically Active Tyrosine Residues of Both SPO11-1 and SPO11-2 Are Required for Meiotic Double-Strand Break Induction in Arabidopsis'. The Plant Cell 19(10): 3090–99.

He, Yan et al. 2017. 'Genomic Features Shaping the Landscape of Meiotic Double-Strand-Break Hotspots in Maize'. Proceedings of the National Academy of Sciences 114(46): 12231–36.

Heckmann, Stefan et al. 2014. 'Alternative Meiotic Chromatid Segregation in the Holocentric Plant Luzula Elegans'. Nature Communications 5(1): 4979.

Henikoff, Jorja G., Jitendra Thakur, Sivakanthan Kasinathan, and Steven Henikoff. 2015. 'A Unique Chromatin Complex Occupies Young α-Satellite Arrays of Human Centromeres'. Science Advances 1(1): e1400234.

Henikoff, Steven, Kami Ahmad, and Harmit S. Malik. 2001. 'The Centromere Paradox: Stable Inheritance with Rapidly Evolving DNA'. Science 293(5532): 1098–1102.

Herr, A. J., M. B. Jensen, T. Dalmay, and D. C. Baulcombe. 2005. 'RNA Polymerase IV Directs Silencing of Endogenous DNA'. Science 308(5718): 118–20.

Heyer, Wolf-Dietrich, Kirk T. Ehmsen, and Jie Liu. 2010. 'Regulation of Homologous Recombination in Eukaryotes'. Annual Review of Genetics 44(1): 113–39.

Hickman, Alison B., and Fred Dyda. 2016. 'DNA Transposition at Work'. Chemical Reviews 116(20): 12758–84.

Higgins, James D. et al. 2005. 'The Arabidopsis Synaptonemal Complex Protein ZYP1 Is Required for Chromosome Synapsis and Normal Fidelity of Crossing Over'. Genes & Development 19(20): 2488–2500.

Higgins, James D. et al. 2008. 'AtMSH5 Partners AtMSH4 in the Class I Meiotic Crossover Pathway in Arabidopsis Thaliana , but Is Not Required for Synapsis'. The Plant Journal 55(1): 28–39.

Higgins, James D., Susan J. Armstrong, F. Christopher H. Franklin, and Gareth H. Jones. 2004. 'The Arabidopsis MutS Homolog AtMSH4 Functions at an Early Step in Recombination: Evidence for Two Classes of Recombination in Arabidopsis'. Genes & Development 18(20): 2557–70.

Hildebrand, Erica M., and Sue Biggins. 2016. 'Regulation of Budding Yeast CENP-A Levels Prevents Misincorporation at Promoter Nucleosomes and Transcriptional Defects' ed. Beth A. Sullivan. PLOS Genetics 12(3): e1005930.

Hill, Hunter J., and Kent G. Golic. 2022. 'Chromosome Tug of War: Dicentric Chromosomes and the Centromere Strength Hypothesis'. Cells 11(22): 3550.

Hoffmann, Sebastian et al. 2016. 'CENP-A Is Dispensable for Mitotic Centromere Function after Initial Centromere/Kinetochore Assembly'. Cell Reports 17(9): 2394–2404.

Hoffmann, Sebastian et al. 2020. 'A Genetic Memory Initiates the Epigenetic Loop Necessary to Preserve Centromere Position'. The EMBO Journal 39(20). https://onlinelibrary.wiley.com/doi/10.15252/embj.2020105505 (December 30, 2022).

Hofstatter, Paulo G. et al. 2022. 'Repeat-Based Holocentromeres Influence Genome Architecture and Karyotype Evolution'. Cell 185(17): 3153-3168.e18.

Hori, Tetsuya et al. 2008. 'CCAN Makes Multiple Contacts with Centromeric DNA to Provide Distinct Pathways to the Outer Kinetochore'. Cell 135(6): 1039–52.

Hori, Tetsuya et al. 2017. 'Association of M18BP1/KNL2 with CENP-A Nucleosome Is Essential for Centromere Formation in Non-Mammalian Vertebrates'. Developmental Cell 42(2): 181-189.e3.

Hori, Tetsuya, Wei-Hao Shang, Kozo Takeuchi, and Tatsuo Fukagawa. 2013. 'The CCAN Recruits CENP-A to the Centromere and Forms the Structural Core for Kinetochore Assembly'. Journal of Cell Biology 200(1): 45–60.

Hosouchi, T. 2002. 'Physical Map-Based Sizes of the Centromeric Regions of Arabidopsis Thaliana Chromosomes 1, 2, and 3'. DNA Research 9(4): 117–21.

Hou, Xueren et al. 2022. 'A Near-Complete Assembly of an Arabidopsis Thaliana Genome'. Molecular Plant 15(8): 1247–50.

Hoyt, Savannah J. et al. 2022. 'From Telomere to Telomere: The Transcriptional and Epigenetic State of Human Repeat Elements'. Science 376(6588): eabk3112.

Hsieh, Chia-Ling, Jing Xia, and Haifan Lin. 2020. 'MIWI Prevents Aneuploidy during Meiosis by Cleaving Excess Satellite RNA'. The EMBO Journal 39(16). https://onlinelibrary.wiley.com/doi/10.15252/embj.2019103614 (December 30, 2022).

Hsu, Che-Wei, Cheng-Yu Lo and Cheng-Ruei Lee. 2019. 'On the postglacial spread of human commensal *Arabidopsis thaliana*: journey to the East'. New Phytologist 222(3): 1447-1457.

International Human Genome Sequencing Consortium et al. 2001. 'Initial Sequencing and Analysis of the Human Genome'. Nature 409(6822): 860–921.

Ishikura, Shuhei et al. 2020. 'ZFAT Binds to Centromeres to Control Noncoding RNA Transcription through the KAT2B–H4K8ac–BRD4 Axis'. Nucleic Acids Research 48(19): 10848–66.

Ito, Hidetaka. 2012. 'Small RNAs and Transposon Silencing in Plants: Small RNAs and Transposon'. Development, Growth & Differentiation 54(1): 100–107.

Iwata-Otsubo, Aiko et al. 2017. 'Expanded Satellite Repeats Amplify a Discrete CENP-A Nucleosome Assembly Site on Chromosomes That Drive in Female Meiosis'. Current Biology 27(15): 2365-2373.e8.

Jansen, Lars E.T., Ben E. Black, Daniel R. Foltz, and Don W. Cleveland. 2007. 'Propagation of Centromeric Chromatin Requires Exit from Mitosis'. Journal of Cell Biology 176(6): 795–805.

Janssen, Aniek, Serafin U. Colmenares, and Gary H. Karpen. 2018. 'Heterochromatin: Guardian of the Genome'. Annual Review of Cell and Developmental Biology 34(1): 265–88.

Jantsch, Verena et al. 2004. 'Targeted Gene Knockout Reveals a Role in Meiotic Recombination for ZHP-3, a Zip3-Related Protein in Caenorhabditis Elegans'. Molecular and Cellular Biology 24(18): 7998–8006.

Jiang, Caifu et al. 2014. 'Environmentally Responsive Genome-Wide Accumulation of de Novo Arabidopsis Thaliana Mutations and Epimutations'. Genome Research 24(11): 1821–29.

Jiang, Jiming, James A Birchler, Wayne A Parrott, and R Kelly Dawe. 2003. 'A Molecular View of Plant Centromeres'. Trends in Plant Science 8(12): 570–75.

Johannes, Frank et al. 2009. 'Assessing the Impact of Transgenerational Epigenetic Variation on Complex Traits' ed. Peter M. Visscher. PLoS Genetics 5(6): e1000530.

Johnson, Whitney L, and Aaron F Straight. 2017. 'RNA-Mediated Regulation of Heterochromatin'. Current Opinion in Cell Biology 46: 102–9.

Jones, Gareth H., and F. Chris H. Franklin. 2006. 'Meiotic Crossing-over: Obligation and Interference'. Cell 126(2): 246–48.

Jones, R. N. 1991. 'B-Chromosome Drive'. The American Naturalist 137(3): 430–42.

Jun, Suckjoon, and Bela Mulder. 2006. 'Entropy-Driven Spatial Organization of Highly Confined Polymers: Lessons for the Bacterial Chromosome'. Proceedings of the National Academy of Sciences 103(33): 12388–93.

Kabeche, Lilian, Hai Dang Nguyen, Rémi Buisson, and Lee Zou. 2018. 'A Mitosis-Specific and R Loop–Driven ATR Pathway Promotes Faithful Chromosome Segregation'. Science 359(6371): 108–14.

Kanellopoulou, Chryssa et al. 2005. 'Dicer-Deficient Mouse Embryonic Stem Cells Are Defective in Differentiation and Centromeric Silencing'. Genes & Development 19(4): 489–501.

Kapitonov, Vladimir V., and Jerzy Jurka. 2006. 'Self-Synthesizing DNA Transposons in Eukaryotes'. Proceedings of the National Academy of Sciences 103(12): 4540–45.

Kapitonov, Vladimir V, and Jerzy Jurka. 2008. 'A Universal Classification of Eukaryotic Transposable Elements Implemented in Repbase'. g e n e ti c s.

Keeney, S., and M.J. Neale. 2006. 'Initiation of Meiotic Recombination by Formation of DNA Double-Strand Breaks: Mechanism and Regulation'. Biochemical Society Transactions 34(4): 523–25.

Kerzendorfer, C. et al. 2006. 'The Arabidopsis Thaliana MND1 Homologue Plays a Key Role in Meiotic Homologous Pairing, Synapsis and Recombination'. Journal of Cell Science 119(12): 2486–96.

Kleckner, N. 1996. 'Meiosis: How Could It Work?' Proceedings of the National Academy of Sciences 93(16): 8167–74.

Kobayashi, Masaaki et al. 2014. 'Genome-Wide Analysis of Intraspecific DNA Polymorphism in "Micro-Tom", a Model Cultivar of Tomato (Solanum Lycopersicum)'. Plant and Cell Physiology 55(2): 445–54.

Kobbe, Daniela et al. 2008. 'AtRECQ2, a RecQ Helicase Homologue from Arabidopsis Thaliana , Is Able to Disrupt Various Recombinogenic DNA Structures in Vitro'. The Plant Journal 55(3): 397–405.

Kong, Augustine et al. 2014. 'Common and Low-Frequency Variants Associated with Genome-Wide Recombination Rate'. Nature Genetics 46(1): 11–16.

Koonin, Eugene V, and Mart Krupovic. 2017. 'Polintons, Virophages and Transpovirons: A Tangled Web Linking Viruses, Transposons and Immunity'. Current Opinion in Virology 25: 7–15.

Koumbaris, George et al. 2011. 'FoSTeS, MMBIR and NAHR at the Human Proximal Xp Region and the Mechanisms of Human Xq Isochromosome Formation'. Human Molecular Genetics 20(10): 1925–36.

Kramara, J., B. Osia, and A. Malkova. 2018. 'Break-Induced Replication: The Where, The Why, and The How'. Trends in Genetics 34(7): 518–31.

Krupovic, Mart, Kira S. Makarova, and Eugene V. Koonin. 2022. 'Cellular Homologs of the Double Jelly-Roll Major Capsid Proteins Clarify the Origins of an Ancient Virus Kingdom'. Proceedings of the National Academy of Sciences 119(5): e2120620119.

Kumar, Rajeev et al. 2019. 'Antagonism between BRCA2 and FIGL1 Regulates Homologous Recombination'. Nucleic Acids Research 47(10): 5170–80.

Kumekawa, Norikazu, Tsutomu Hosouchi, Hisano Tsuruoka, and Hirokazu Kotani. 'The Size and Sequence Organization of the Centromeric Region of Arabidopsis Thaliana Chromosome 5'. 7(6).

Kumon, Tomohiro et al. 2021. 'Parallel Pathways for Recruiting Effector Proteins Determine Centromere Drive and Suppression'. Cell 184(19): 4904-4918.e11.

Kumon, Tomohiro, and Michael A. Lampson. 2022. 'Evolution of Eukaryotic Centromeres by Drive and Suppression of Selfish Genetic Elements'. Seminars in Cell & Developmental Biology 128: 51–60.

Kunyavskaya, Olga et al. 2022. 'Automated Annotation of Human Centromeres with HORmon'. Genome Research 32(6): 1137–51.

Kuromori, Takashi et al. 2008. 'Homologous Chromosome Pairing Is Completed in Crossover Defective Atzip4 Mutant'. Biochemical and Biophysical Research Communications 370(1): 98–103.

Kursel, Lisa E, Hannah McConnell, Aida Flor A de la Cruz, and Harmit S Malik. 2021. 'Gametic Specialization of Centromeric Histone Paralogs in Drosophila Virilis'. Life Science Alliance 4(7): e202000992.

Kursel, Lisa E, Frances C Welsh, and Harmit S Malik. 2020. 'Ancient Coretention of Paralogs of Cid Centromeric Histones and Cal1 Chaperones in Mosquito Species' ed. Amanda Larracuente. Molecular Biology and Evolution 37(7): 1949–63.

Kurzbauer, Marie-Therese, Clemens Uanschou, Doris Chen, and Peter Schlögelhofer. 2012. 'The Recombinases DMC1 and RAD51 Are Functionally and Spatially Separated during Meiosis in Arabidopsis'. The Plant Cell 24(5): 2058–70.

Lambing, Christophe, Pallas C. Kuo, et al. 2020. 'ASY1 Acts as a Dosage-Dependent Antagonist of Telomere-Led Recombination and Mediates Crossover Interference in Arabidopsis'. Proceedings of the National Academy of Sciences 117(24): 13647–58.

Lambing, Christophe, Andrew J. Tock, et al. 2020. 'Interacting Genomic Landscapes of REC8-Cohesin, Chromatin, and Meiotic Recombination in Arabidopsis'. The Plant Cell 32(4): 1218–39.

Lambing, Christophe, F. Chris H. Franklin, and Chung-Ju Rachel Wang. 2017. 'Understanding and Manipulating Meiotic Recombination in Plants'. Plant Physiology 173(3): 1530–42.

Lamesch, Philippe et al. 2012. 'The Arabidopsis Information Resource (TAIR): Improved Gene Annotation and New Tools'. Nucleic Acids Research 40(D1): D1202–10.

Lando, David et al. 2012. 'Quantitative Single-Molecule Microscopy Reveals That CENP-A Cnp1 Deposition Occurs during G2 in Fission Yeast'. Open Biology 2(7): 120078.

Langdon, Tim et al. 2003. 'A High-Copy-Number CACTA Family Transposon in Temperate Grasses and Cereals'. Genetics 163(3): 1097–1108.

Langley, Sasha A, Karen H Miga, Gary H Karpen, and Charles H Langley. 2019. 'Haplotypes Spanning Centromeric Regions Reveal Persistence of Large Blocks of Archaic DNA'. eLife 8: e42989.

Lebrigand, Kevin, Virginie Magnone, Pascal Barbry, and Rainer Waldmann. 2020. 'High Throughput Error Corrected Nanopore Single Cell Transcriptome Sequencing'. Nature Communications 11(1): 4025.

Leclerc, Simon, and Katsumi Kitagawa. 2021. 'The Role of Human Centromeric RNA in Chromosome Stability'. Frontiers in Molecular Biosciences 8: 642732.

Lee, Hye-Ran et al. 2005. 'Chromatin Immunoprecipitation Cloning Reveals Rapid Evolutionary Patterns of Centromeric DNA in Oryza Species'. Proceedings of the National Academy of Sciences 102(33): 11793–98.

Lee, Sung-Il, and Nam-Soo Kim. 2014. 'Transposable Elements and Genome Size Variations in Plants'. Genomics & Informatics 12(3): 87.

Lenormand, Thomas et al. 2016. 'Evolutionary Mysteries in Meiosis'. Philosophical Transactions of the Royal Society B: Biological Sciences 371(1706): 20160001.

Lermontova, Inna et al. 2006. 'Loading of Arabidopsis Centromeric Histone CENH3 Occurs Mainly during G2 and Requires the Presence of the Histone Fold Domain'. The Plant Cell 18(10): 2443–51.

Lermontova, Inna et al. 2015. 'Centromeric Chromatin and Its Dynamics in Plants'. The Plant Journal 83(1): 4–17.

Li, Wenzhu, and Xiangwei He. 2020. 'Inverted Meiosis: An Alternative Way of Chromosome Segregation for Reproduction'. Acta Biochimica et Biophysica Sinica 52(7): 702–7.

Li, Xuexian, and R. Kelly Dawe. 2009. 'Fused Sister Kinetochores Initiate the Reductional Division in Meiosis I'. Nature Cell Biology 11(9): 1103–8.

Li, Yafei et al. 2018. 'HEIP1 Regulates Crossover Formation during Meiosis in Rice'. Proceedings of the National Academy of Sciences 115(42): 10810–15.

Li, Yubin, and Hugo K Dooner. 2009. 'Excision of Helitron Transposons in Maize'. Genetics 182(1): 399–402.

Lin, F L, K Sperle, and N Sternberg. 1984. 'Model for Homologous Recombination during Transfer of DNA into Mouse L Cells: Role for DNA Ends in the Recombination Process'. MOL. CELL. BIOL. 4.

Lin, Xiaoying et al. 1999. 'Sequence and Analysis of Chromosome 2 of the Plant Arabidopsis Thaliana'. 402.

Ling, Yick Hin, and Karen Wing Yee Yuen. 2019. 'Centromeric Non-Coding RNA as a Hidden Epigenetic Factor of the Point Centromere'. Current Genetics 65(5): 1165–71.

Linheiro, Raquel S., and Casey M. Bergman. 2012. 'Whole Genome Resequencing Reveals Natural Target Site Preferences of Transposable Elements in Drosophila Melanogaster' ed. Jason E. Stajich. PLoS ONE 7(2): e30008.

Liu, Hong et al. 2015. 'Mitotic Transcription Installs Sgo1 at Centromeres to Coordinate Chromosome Segregation'. Molecular Cell 59(3): 426–36.

Liu, Yao-Guang, and Yuanling Chen. 2007. 'High-Efficiency Thermal Asymmetric Interlaced PCR for Amplification of Unknown Flanking Sequences'. BioTechniques 43(5): 649–56.

Lopez, Virginia et al. 2015. 'Cytokinesis Breaks Dicentric Chromosomes Preferentially at Pericentromeric Regions and Telomere Fusions'. Genes & Development 29(3): 322–36.

Lv, Jian et al. 2020. 'Generation of Paternal Haploids in Wheat by Genome Editing of the Centromeric Histone CENH3'. Nature Biotechnology 38(12): 1397–1401.

Macaisne, Nicolas et al. 2008. 'SHOC1, an XPF Endonuclease-Related Protein, Is Essential for the Formation of Class I Meiotic Crossovers'. Current Biology 18(18): 1432–37.

Macaisne, Nicolas, Julien Vignard, and Raphaël Mercier. 2011. 'SHOC1 and PTD Form an XPF–ERCC1-like Complex That Is Required for Formation of Class I Crossovers'. Journal of Cell Science 124(16): 2687–91.

Maheshwari, Shamoni et al. 2015. 'Naturally Occurring Differences in CENH3 Affect Chromosome Segregation in Zygotic Mitosis of Hybrids' ed. Kirsten Bomblies. PLOS Genetics 11(1): e1004970.

Maheshwari, Shamoni et al. 2017. 'Centromere Location in Arabidopsis Is Unaltered by Extreme Divergence in CENH3 Protein Sequence'. Genome Research 27(3): 471–78.

Maison, C., J.- P. Quivy, A. V. Probst, and G. Almouzni. 2010. 'Heterochromatin at Mouse Pericentromeres: A Model for De Novo Heterochromatin Formation and Duplication during Replication'. Cold Spring Harbor Symposia on Quantitative Biology 75(0): 155–65.

Malik, Harmit S., and Steven Henikoff. 2009. 'Major Evolutionary Transitions in Centromere Complexity'. Cell 138(6): 1067–82.

Malik, Harmit S., Danielle Vermaak, and Steven Henikoff. 2002. 'Recurrent Evolution of DNA-Binding Motifs in the Drosophila Centromeric Histone'. Proceedings of the National Academy of Sciences 99(3): 1449–54.

Mandáková, Terezie, Petra Hloušková, Marcus A. Koch, and Martin A. Lysak. 2020. 'Genome Evolution in Arabideae Was Marked by Frequent Centromere Repositioning'. The Plant Cell 32(3): 650–65.

Mandrioli, Mauro, and Gian Carlo Manicardi. 2020. 'Holocentric Chromosomes'. PLOS Genetics 16(7): e1008918.

Manthei, Kelly A., and James L. Keck. 2013. 'The BLM Dissolvasome in DNA Replication and Repair'. Cellular and Molecular Life Sciences 70(21): 4067–84.

Marimuthu, Mohan P. A. et al. 2021. 'Epigenetically Mismatched Parental Centromeres Trigger Genome Elimination in Hybrids'. Science Advances 7(47): eabk1151.

Marques, André et al. 2015. 'Holocentromeres in Rhynchospora Are Associated with Genome-Wide Centromere-Specific Repeat Arrays Interspersed among Euchromatin'. Proceedings of the National Academy of Sciences 112(44): 13633–38.

Marshall, Owen J., Anderly C. Chueh, Lee H. Wong, and K.H. Andy Choo. 2008. 'Neocentromeres: New Insights into Centromere Structure, Disease Development, and Karyotype Evolution'. The American Journal of Human Genetics 82(2): 261–82.

Mateo, Lidia, and Josefa González. 2014. 'Pogo-like Transposases Have Been Repeatedly Domesticated into CENP-B-Related Proteins'. Genome Biology and Evolution 6(8): 2008–16.

May, Bruce P et al. 2005. 'Differential Regulation of Strand-Specific Transcripts from Arabidopsis Centromeric Satellite Repeats' ed. John Doebley. PLoS Genetics 1(6): e79.

McClintock, Barbara. 1938. 'THE PRODUCTION OF HOMOZYGOUS DEFICIENT TISSUES WITH MUTANT CHARACTERISTICS BY MEANS OF THE ABERRANT MITOTIC BEHAVIOR OF RING-SHAPED CHROMOSOMES'. Genetics 23(4): 315–76.

McClintock, Barbara. 1941a. 'THE ASSOCIATION OF MUTANTS WITH HOMOZYGOUS DEFICIENCIES IN ZEA MAYS'. Genetics 26(5): 542–71.

McEwen, Bruce F. et al. 2001. 'CENP-E Is Essential for Reliable Bioriented Spindle Attachment, but Chromosome Alignment Can Be Achieved via Redundant Mechanisms in Mammalian Cells' ed. Ted Salmon. Molecular Biology of the Cell 12(9): 2776–89.

McKinley, Kara L., and Iain M. Cheeseman. 2016a. 'The Molecular Basis for Centromere Identity and Function'. Nature Reviews Molecular Cell Biology 17(1): 16–29.

McKinley, Kara L., and Iain M. Cheeseman. 2016b. 'The Molecular Basis for Centromere Identity and Function'. Nature Reviews Molecular Cell Biology 17(1): 16–29.

McKinley, Kara L., and Iain M. Cheeseman. 2014. 'Polo-like Kinase 1 Licenses CENP-A Deposition at Centromeres'. Cell 158(2): 397–411.

McMahill, Melissa S, Caroline W Sham, and Douglas K Bishop. 2007. 'Synthesis-Dependent Strand Annealing in Meiosis' ed. Michael Lichten. PLoS Biology 5(11): e299.

McNulty, Shannon M., Lori L. Sullivan, and Beth A. Sullivan. 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.e6.

Mellone, Barbara G. et al. 2011. 'Assembly of Drosophila Centromeric Chromatin Proteins during Mitosis' ed. Sue Biggins. PLoS Genetics 7(5): e1002068.

Melters, Daniël P et al. 2013. 'Comparative Analysis of Tandem Repeats from Hundreds of Species Reveals Unique Insights into Centromere Evolution'. Genome Biology 14(1): R10.

Meluh, Pamela B et al. 1998. 'Cse4p Is a Component of the Core Centromere of Saccharomyces Cerevisiae'. Cell 94(5): 607–13.

Meraldi, Patrick, AndrewD McAinsh, Esther Rheinbay, and PeterK Sorger. 2006. 'Phylogenetic and Structural Analysis of Centromeric DNA and Kinetochore Proteins'. Genome Biology 7(3): R23.

Mercier, R., and M. Grelon. 2008. 'Meiosis in Plants: Ten Years of Gene Discovery'. Cytogenetic and Genome Research 120(3–4): 281–90.

Mercier, Raphaël et al. 2005. 'Two Meiotic Crossover Classes Cohabit in Arabidopsis'. Current Biology 15(8): 692–701.

Mercier, Raphaël et al. 2015. 'The Molecular Biology of Meiosis in Plants'. Annual Review of Plant Biology 66(1): 297–327.

Meštrović, Nevenka et al. 2015. 'Structural and Functional Liaisons between Transposable Elements and Satellite DNAs'. Chromosome Research 23(3): 583–96.

Mézard, Christine, Julien Vignard, Jan Drouaud, and Raphaël Mercier. 2007. 'The Road to Crossovers: Plants Have Their Say'. Trends in Genetics 23(2): 91–99.

Miga, Karen H. et al. 2014. 'Centromere Reference Models for Human Chromosomes X and Y Satellite Arrays'. Genome Research 24(4): 697–707.

Miga, Karen H. et al.. 2020. 'Telomere-to-Telomere Assembly of a Complete Human X Chromosome'. Nature 585(7823): 79–84.

Miga, Karen H, and Beth A Sullivan. 2021. 'Expanding Studies of Chromosome Structure and Function in the Era of T2T Genomics'. Human Molecular Genetics: ddab214.

Miller, Joseph T et al. 1998. 'Retrotransposon-Related DNA Sequences in the Centromeres of Grass Chromosomes'. Genetics 150(4): 1615–23.

Mishra, Prashant K., Mary Baum, and John Carbon. 2007. 'Centromere Size and Position in Candida Albicans Are Evolutionarily Conserved Independent of DNA Sequence Heterogeneity'. Molecular Genetics and Genomics 278(4): 455–65.

Mitchell, A. R., J. R. Gosden, and D. A. Miller. 1985. 'A Cloned Sequence, P82H, of the Alphoid Repeated DNA Family Found at the Centromeres of All Human Chromosomes'. Chromosoma 92(5): 369–77.

Molina, Wagner Franco, Pablo A. Martinez, Luiz Antônio C. Bertollo, and Claudio Juan Bidau. 2014. 'Evidence for Meiotic Drive as an Explanation for Karyotype Changes in Fishes'. Marine Genomics 15: 29–34.

Moran, E Sanchez et al. 'Chiasma Formation in Arabidopsis Thaliana Accession Wassileskija and in Two Meiotic Mutants'.

Morgan, Chris et al. 2021. 'Diffusion-Mediated HEI10 Coarsening Can Explain Meiotic Crossover Positioning in Arabidopsis'. Nature Communications 12(1): 4674.

Murchison, Elizabeth P. et al. 2005. 'Characterization of Dicer-Deficient Murine Embryonic Stem Cells'. Proceedings of the National Academy of Sciences 102(34): 12135–40.

Murray, Robert W., and Ramasubbu Jeyaraman. 1985. 'Dioxiranes: Synthesis and Reactions of Methyldioxiranes'. The Journal of Organic Chemistry 50(16): 2847–53.

Musacchio, Andrea, and Arshad Desai. 2017. 'A Molecular View of Kinetochore Assembly and Function'. Biology 6(4): 5.

Musilova, P., S. Kubickova, J. Vahala, and J. Rubes. 2013. 'Subchromosomal Karyotype Evolution in Equidae'. Chromosome Research 21(2): 175–87.

Nagaki, Kiyotaka, and Minoru Murata. 2005. 'Characterization of CENH3 and Centromere-Associated DNA Sequences in Sugarcane'. Chromosome Research 13(2): 195–203.

Naish, Matthew et al. 2021. 'The Genetic and Epigenetic Landscape of the Arabidopsis Centromeres'. Science 374(6569): eabi7489.

Nakano, Megumi et al. 2008. 'Inactivation of a Human Kinetochore by Specific Targeting of Chromatin Modifiers'. Developmental Cell 14(4): 507–22.

Nardi, Isaac K. et al. 2016. 'Licensing of Centromeric Chromatin Assembly through the Mis18α-Mis18β Heterotetramer'. Molecular Cell 61(5): 774–87.

Nasuda, S. et al. 2005. 'Stable Barley Chromosomes without Centromeric Repeats'. Proceedings of the National Academy of Sciences 102(28): 9842–47.

Nechemia-Arbely, Yael et al. 2017. 'Human Centromeric CENP-A Chromatin Is a Homotypic, Octameric Nucleosome at All Cell Cycle Points'. Journal of Cell Biology 216(3): 607–21.

Neumann, Pavel, Huihuang Yan, and Jiming Jiang. 2007. 'The Centromeric Retrotransposons of Rice Are Transcribed and Differentially Processed by RNA Interference'. Genetics 176(2): 749–61.

Novak, P. et al. 2013. 'RepeatExplorer: A Galaxy-Based Web Server for Genome-Wide Characterization of Eukaryotic Repetitive Elements from next-Generation Sequence Reads'. Bioinformatics 29(6): 792–93.

Novitski, E. 1952. 'THE GENETIC CONSEQUENCES OF ANAPHASE BRIDGE FORMATION IN DROSOPHILA'. Genetics 37(3): 270–87.

Nurk, Sergey et al. 2022. 'The Complete Sequence of a Human Genome'.

Ogura, Yutaka, Fukashi Shibata, Hiroshi Sato, and Minoru Murata. 2004. 'Characterization of a CENP-C Homolog in Arabidopsis Thaliana'. Genes & Genetic Systems 79(3): 139–44.

Ohkuni, Kentaro, and Katsumi Kitagawa. 2011. 'Endogenous Transcription at the Centromere Facilitates Centromere Activity in Budding Yeast'. Current Biology 21(20): 1695–1703.

Ohzeki, Jun-ichirou et al. 2016. 'KAT7/HBO1/MYST2 Regulates CENP-A Chromatin Assembly by Antagonizing Suv39h1-Mediated Centromere Inactivation'. Developmental Cell 37(5): 413–27.

Ohzeki, Jun-ichirou, Vladimir Larionov, William C. Earnshaw, and Hiroshi Masumoto. 2015. 'Genetic and Epigenetic Regulation of Centromeres: A Look at HAC Formation'. Chromosome Research 23(1): 87–103.

Ohzeki, Jun-ichirou, Koichiro Otake, and Hiroshi Masumoto. 2020. 'Human Artificial Chromosome: Chromatin Assembly Mechanisms and CENP-B'. Experimental Cell Research 389(2): 111900.

Okada, Masahiro et al. 2006. 'The CENP-H–I Complex Is Required for the Efficient Incorporation of Newly Synthesized CENP-A into Centromeres'. Nature Cell Biology 8(5): 446–57.

Okada, Teruaki et al. 2007. 'CENP-B Controls Centromere Formation Depending on the Chromatin Context'. Cell 131(7): 1287–1300.

Ólafsson, Guðjón, and Peter H. Thorpe. 2020. 'Polo Kinase Recruitment via the Constitutive Centromere-Associated Network at the Kinetochore Elevates Centromeric RNA' ed. Beth A. Sullivan. PLOS Genetics 16(8): e1008990.

Osman, Christof, Dennis R. Voelker, and Thomas Langer. 2011. 'Making Heads or Tails of Phospholipids in Mitochondria'. Journal of Cell Biology 192(1): 7–16.

Osman, Kim et al. 2011. 'Pathways to Meiotic Recombination in Arabidopsis Thaliana'. New Phytologist 190(3): 523–44.

Ossowski, Stephan et al. 2010. 'The Rate and Molecular Spectrum of Spontaneous Mutations in Arabidopsis Thaliana'. Science 327(5961): 92–94.

Otsu, Nobuyuki. 'A Tlreshold Selection Method from Gray-Level Histograms'.

Paar, Vladimir, Ivan Basar, Marija Rosandic, and Matko Gluncic. 2007. 'Consensus Higher Order Repeats and Frequency of String Distributions in Human Genome'. Current Genomics 8(2): 93–111.

Page, Scott L., and R. Scott Hawley. 2004. 'THE GENETICS AND MOLECULAR BIOLOGY OF THE SYNAPTONEMAL COMPLEX'. Annual Review of Cell and Developmental Biology 20(1): 525–58.

Pan, Dongqing et al. 2017. 'CDK-Regulated Dimerization of M18BP1 on a Mis18 Hexamer Is Necessary for CENP-A Loading'. eLife 6: e23352.

———. 2019. 'Mechanism of Centromere Recruitment of the CENP-A Chaperone HJURP and Its Implications for Centromere Licensing'. Nature Communications 10(1): 4046.

Pardo-Manuel de Villena, Fernando, and Carmen Sapienza. 2001. 'Nonrandom Segregation during Meiosis: The Unfairness of Females'. Mammalian Genome 12(5): 331–39.

Plohl, Miroslav, Nevenka Meštrović, and Brankica Mravinac. 2014. 'Centromere Identity from the DNA Point of View'. Chromosoma 123(4): 313–25.

Pontier, Dominique et al. 2012. 'NERD, a Plant-Specific GW Protein, Defines an Additional RNAi-Dependent Chromatin-Based Pathway in Arabidopsis'. Molecular Cell 48(1): 121–32.

Poulter, Russell T M, and Margi I Butler. 'Tyrosine Recombinase Retrotransposons and Transposons'.

Pradillo, Mónica, Javier Varas, Cecilia Oliver, and Juan L. Santos. 2014. 'On the Role of AtDMC1, AtRAD51 and Its Paralogs during Arabidopsis Meiosis'. Frontiers in Plant Science 5. http://journal.frontiersin.org/article/10.3389/fpls.2014.00023/abstract (December 30, 2022).

Pritham, Ellen J., Tasneem Putliwala, and Cédric Feschotte. 2007. 'Mavericks, a Novel Class of Giant Transposable Elements Widespread in Eukaryotes and Related to DNA Viruses'. Gene 390(1–2): 3–17.

Purgato, Stefania et al. 2015. 'Centromere Sliding on a Mammalian Chromosome'. Chromosoma 124(2): 277–87.

Qiao, Huanyu et al. 2014. 'Antagonistic Roles of Ubiquitin Ligase HEI10 and SUMO Ligase RNF212 Regulate Meiotic Recombination'. Nature Genetics 46(2): 194–99.

Quadrana, Leandro et al. 2019. 'Transposition Favors the Generation of Large Effect Mutations That May Facilitate Rapid Adaption'. Nature Communications 10(1): 3421.

Quénet, Delphine, and Yamini Dalal. 2014. 'A Long Non-Coding RNA Is Required for Targeting Centromeric Protein A to the Human Centromere'. eLife 3: e26016.

Quénet, Delphine, David Sturgill, Marin Olson, and Yamini Dalal. 2017. CENP-A Associated LncRNAs Influence Chromosome Segregation in Human Cells. Cell Biology. preprint. http://biorxiv.org/lookup/doi/10.1101/097956 (December 30, 2022).

Rabanal, Fernando A et al. 2022. 'Pushing the limits of HiFi assemblies reveals centromere diversity between two *Arabidopsis thaliana* genomes'. Nucleic Acids Research, 50(21): 12309–12327

Raynard, Steven, Wendy Bussen, and Patrick Sung. 2006. 'A Double Holliday Junction Dissolvasome Comprising BLM, Topoisomerase IIIα, and BLAP75'. Journal of Biological Chemistry 281(20): 13861–64.

Reinders, Jon, and Jerzy Paszkowski. 2009. 'Unlocking the Arabidopsis Epigenome'. Epigenetics 4(8): 557–63.

Reynolds, April et al. 2013. 'RNF212 Is a Dosage-Sensitive Regulator of Crossing-over during Mammalian Meiosis'. Nature Genetics 45(3): 269–78.

Rhind, Nicholas et al. 2011. 'Comparative Functional Genomics of the Fission Yeasts'. Science 332(6032): 930–36.

Ribeiro, Tiago et al. 2017. 'Centromeric and Non-Centromeric Satellite DNA Organisation Differs in Holocentric Rhynchospora Species'. Chromosoma 126(2): 325–35.

Rice, William. 2020. Why Do Centromeres Evolve So Fast: BIR Replication, Hypermutation, Transposition, and Molecular-Drive. LIFE SCIENCES. preprint. https://www.preprints.org/manuscript/202012.0669/v1 (December 30, 2022).

Rigal, Mélanie et al. 2016. 'Epigenome Confrontation Triggers Immediate Reprogramming of DNA Methylation and Transposon Silencing in Arabidopsis Thaliana F1 Epihybrids'. Proceedings of the National Academy of Sciences 113(14). https://pnas.org/doi/full/10.1073/pnas.1600672113 (December 30, 2022).

Rizvi, Syed Meraj Azhar, Hemant Kumar Prajapati, and Santanu Kumar Ghosh. 2018. 'The 2 Micron Plasmid: A Selfish Genetic Element with an Optimized Survival Strategy within Saccharomyces Cerevisiae'. Current Genetics 64(1): 25–42.

Rocchi, M et al. 2012. 'Centromere Repositioning in Mammals'. Heredity 108(1): 59–67.

Romanova, L.Y. et al. 1996. 'Evidence for Selection in Evolution of Alpha Satellite DNA: The Central Role of CENP-B/PJα Binding Region'. Journal of Molecular Biology 261(3): 334–40.

Ross, Justyne E., Kaitlin Stimpson Woodlief, and Beth A. Sullivan. 2016. 'Inheritance of the CENP-A Chromatin Domain Is Spatially and Temporally Constrained at Human Centromeres'. Epigenetics & Chromatin 9(1): 20.

Rowan, Beth A, Vipul Patel, Detlef Weigel, and Korbinian Schneeberger. 2015. 'Rapid and Inexpensive Whole-Genome Genotyping-by-Sequencing for Crossover Localization and Fine-Scale Genetic Mapping'. G3 Genes|Genomes|Genetics 5(3): 385–98.

Salomé, P A et al. 2012. 'The Recombination Landscape in Arabidopsis Thaliana F2 Populations'. Heredity 108(4): 447–55.

Sandmann, Michael et al. 2017. 'Targeting of Arabidopsis KNL2 to Centromeres Depends on the Conserved CENPC-k Motif in Its C Terminus'. The Plant Cell 29(1): 144–55.

Sanei, Maryam et al. 2011. 'Loss of Centromeric Histone H3 (CENH3) from Centromeres Precedes Uniparental Chromosome Elimination in Interspecific Barley Hybrids'. Proceedings of the National Academy of Sciences 108(33). https://pnas.org/doi/full/10.1073/pnas.1103190108 (December 30, 2022).

Sanyal, Kaustuv, Mary Baum, and John Carbon. 2004. 'Centromeric DNA Sequences in the Pathogenic Yeast Candida Albicans Are All Different and Unique'. Proceedings of the National Academy of Sciences 101(31): 11374–79.

Sart, Desirée du et al. 1997. 'A Functional Neo-Centromere Formed through Activation of a Latent Human Centromere and Consisting of Non-Alpha-Satellite DNA'. Nature Genetics 16(2): 144–53.

Sasaki, Eriko, Taiji Kawakatsu, Joseph R. Ecker, and Magnus Nordborg. 2019. 'Common Alleles of CMT2 and NRPE1 Are Major Determinants of CHH Methylation Variation in Arabidopsis Thaliana' ed. Claudia Köhler. PLOS Genetics 15(12): e1008492.

Schaper, Elke et al. 2015. 'TRAL: Tandem Repeat Annotation Library'. Bioinformatics 31(18): 3051–53.

Schleiffer, Alexander et al. 2012. 'CENP-T Proteins Are Conserved Centromere Receptors of the Ndc80 Complex'. Nature Cell Biology 14(6): 604–13.

Schmitz, Robert J. et al. 2011. 'Transgenerational Epigenetic Instability Is a Source of Novel Methylation Variants'. Science 334(6054): 369–73.

Schreiber, Mona et al. 2022. 'Recombination Landscape Divergence Between Populations Is Marked by Larger Low-Recombining Regions in Domesticated Rye' ed. Michael Purugganan. Molecular Biology and Evolution 39(6): msac131.

Schueler, Mary G. et al. 2010. 'Adaptive Evolution of Foundation Kinetochore Proteins in Primates'. Molecular Biology and Evolution 27(7): 1585–97.

Schwacha, Anthony, and Nancy Kleckner. 1997. 'Interhomolog Bias during Meiotic Recombination: Meiotic Functions Promote a Highly Differentiated Interhomolog-Only Pathway'. Cell 90(6): 1123–35.

Séguéla-Arnaud, Mathilde et al. 2015. 'Multiple Mechanisms Limit Meiotic Crossovers: TOP3α and Two BLM Homologs Antagonize Crossovers in Parallel to FANCM'. Proceedings of the National Academy of Sciences 112(15): 4713–18.

Séguéla-Arnaud, Mathilde et al. 2016. 'RMI1 and TOP3α Limit Meiotic CO Formation through Their C-Terminal Domains'. Nucleic Acids Research: gkw1210.

Serra, Heïdi et al. 2018. 'Massive Crossover Elevation via Combination of HEI10 and Recq4a Recq4b during Arabidopsis Meiosis'. Proceedings of the National Academy of Sciences 115(10): 2437–42.

Sevim, Volkan, Ali Bashir, Chen-Shan Chin, and Karen H. Miga. 2016. 'Alpha-CENTAURI: Assessing Novel Centromeric Repeat Sequence Variation with Long Read Sequencing'. Bioinformatics 32(13): 1921–24.

Shang, Wei-Hao et al. 2010. 'Chickens Possess Centromeres with Both Extended Tandem Repeats and Short Non-Tandem-Repetitive Sequences'. Genome Research 20(9): 1219–28.

Shelby, Richard D., Karine Monier, and Kevin F. Sullivan. 2000. 'Chromatin Assembly at Kinetochores Is Uncoupled from DNA Replication'. Journal of Cell Biology 151(5): 1113–18.

Shen, Yi et al. 2012. 'The Role of ZIP4 in Homologous Chromosome Synapsis and Crossover Formation in Rice Meiosis'. Journal of Cell Science: jcs.090993.

Shi, Jinghua et al. 2010. 'Widespread Gene Conversion in Centromere Cores' ed. Harmit S. Malik. PLoS Biology 8(3): e1000327.

Shilo, Shay et al. 2015. 'DNA Crossover Motifs Associated with Epigenetic Modifications Delineate Open Chromatin Regions in Arabidopsis'. The Plant Cell 27(9): 2427–36.

Shin, Jinwoo et al. 2018. 'MUN (MERISTEM UNSTRUCTURED), Encoding a SPC24 Homolog of NDC80 Kinetochore Complex, Affects Development through Cell Division in Arabidopsis Thaliana'. The Plant Journal 93(6): 977–91.

Shinohara, Akira, Hideyuki Ogawa, and Tomoko Ogawa. 1992. 'Rad51 Protein Involved in Repair and Recombination in S. Cerevisiae Is a RecA-like Protein'. Cell 69(3): 457–70.

Shrestha, Roshan L. et al. 2017. 'Mislocalization of Centromeric Histone H3 Variant CENP-A Contributes to Chromosomal Instability (CIN) in Human Cells'. Oncotarget 8(29): 46781–800.

Sidhu, Gaganpreet K. et al. 2015. 'Recombination Patterns in Maize Reveal Limits to Crossover Homeostasis'. Proceedings of the National Academy of Sciences 112(52): 15982–87.

Slotte, Tanja et al. 2013. 'The Capsella Rubella Genome and the Genomic Consequences of Rapid Mating System Evolution'. Nature Genetics 45(7): 831–35.

Smit, A F, and A D Riggs. 1996. 'Tiggers and DNA Transposon Fossils in the Human Genome.' Proceedings of the National Academy of Sciences 93(4): 1443–48.

Stankovic, Ana et al. 2017. 'A Dual Inhibitory Mechanism Sufficient to Maintain Cell-Cycle-Restricted CENP-A Assembly'. Molecular Cell 65(2): 231–46.

Su, Handong et al. 2019. 'Centromere Satellite Repeats Have Undergone Rapid Changes in Polyploid Wheat Subgenomes'. The Plant Cell 31(9): 2035–51.

Sujiwattanarat, Penporn et al. 2015. 'Higher-Order Repeat Structure in Alpha Satellite DNA Occurs in New World Monkeys and Is Not Confined to Hominoids'. Scientific Reports 5(1): 10315.

Sullivan, Beth A, and Gary H Karpen. 2004. 'Centromeric Chromatin Exhibits a Histone Modification Pattern That Is Distinct from Both Euchromatin and Heterochromatin'. Nature Structural & Molecular Biology 11(11): 1076–83.

Sullivan, Beth A., and Huntington F. Willard. 1998. 'Stable Dicentric X Chromosomes with Two Functional Centromeres'. Nature Genetics 20(3): 227–28.

Sullivan, K F, M Hechenberger, and K Masri. 1994. 'Human CENP-A Contains a Histone H3 Related Histone Fold Domain That Is Required for Targeting to the Centromere.' Journal of Cell Biology 127(3): 581–92.

Sullivan, Lori L. et al. 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): 457–70.

Sultana, Tania, Alessia Zamborlini, Gael Cristofari, and Pascale Lesage. 2017. 'Integration Site Selection by Retroviruses and Transposable Elements in Eukaryotes'. Nature Reviews Genetics 18(5): 292–308.

Sun, Xiaoji et al. 2015. 'Transcription Dynamically Patterns the Meiotic Chromosome-Axis Interface'. eLife 4: e07424.

Sundararajan, Kousik, and Aaron F. Straight. 2022. 'Centromere Identity and the Regulation of Chromosome Segregation'. Frontiers in Cell and Developmental Biology 10: 914249.

Suzuki, Aussie et al. 2014. 'The Architecture of CCAN Proteins Creates a Structural Integrity to Resist Spindle Forces and Achieve Proper Intrakinetochore Stretch'. Developmental Cell 30(6): 717–30.

Talbert, Paul, and Steven Henikoff. 2022. 'Centromere Drive: Chromatin Conflict in Meiosis'. Current Opinion in Genetics & Development 77: 102005.

Tanaka, Koichi, Hui Li Chang, Ayano Kagami, and Yoshinori Watanabe. 2009. 'CENP-C Functions as a Scaffold for Effectors with Essential Kinetochore Functions in Mitosis and Meiosis'. Developmental Cell 17(3): 334–43.

Tanaka, Tomoyuki U., Michael J. R. Stark, and Kozo Tanaka. 2005. 'Kinetochore Capture and Bi-Orientation on the Mitotic Spindle'. Nature Reviews Molecular Cell Biology 6(12): 929–42.

Tang, Yu et al. 2017. 'MTOPVIB Interacts with AtPRD1 and Plays Important Roles in Formation of Meiotic DNA Double-Strand Breaks in Arabidopsis'. Scientific Reports 7(1): 10007.

Teixeira, José R. et al. 2018. 'Concurrent Duplication of Drosophila Cid and Cenp-C Genes Resulted in Accelerated Evolution and Male Germline-Biased Expression of the New Copies'. Journal of Molecular Evolution 86(6): 353–64.

Thakur, Jitendra, and Steven Henikoff. 2018. 'Unexpected Conformational Variations of the Human Centromeric Chromatin Complex'. Genes & Development 32(1): 20–25.

Thieme, Michael et al. 2017. 'Inhibition of RNA Polymerase II Allows Controlled Mobilisation of Retrotransposons for Plant Breeding'. Genome Biology 18(1): 134.

Tian, Tian et al. 2022. 'Structural Insights into Human CCAN Complex Assembled onto DNA'. Cell Discovery 8(1): 90.

Toby, Garabet G., Wahiba Gherraby, Thomas R. Coleman, and Erica A. Golemis. 2003. 'A Novel RING Finger Protein, Human Enhancer of Invasion 10, Alters Mitotic Progression through Regulation of Cyclin B Levels'. Molecular and Cellular Biology 23(6): 2109–22.

Tsukahara, Sayuri et al. 2009. 'Bursts of Retrotransposition Reproduced in Arabidopsis'. Nature 461(7262): 423–26.

Tsukahara, Sayuri et al. 2012. 'Centromere-Targeted de Novo Integrations of an LTR Retrotransposon of Arabidopsis Lyrata'. Genes & Development 26(7): 705–13.

Underwood, Charles J. et al. 2018. 'Epigenetic Activation of Meiotic Recombination near Arabidopsis Thaliana Centromeres via Loss of H3K9me2 and Non-CG DNA Methylation'. Genome Research 28(4): 519–31.

Underwood, Charles J., and Kyuha Choi. 2019. 'Heterogeneous Transposable Elements as Silencers, Enhancers and Targets of Meiotic Recombination'. Chromosoma 128(3): 279–96.

Voullaire, Lucille et al. 1999. 'Trisomy 20p Resulting from Inverted Duplication and Neocentromere Formation'. American Journal of Medical Genetics 85(4): 403–8.

Vrielynck, Nathalie et al. 2016. 'A DNA Topoisomerase VI–like Complex Initiates Meiotic Recombination'. Science 351(6276): 939–43.

Wade, C. M. et al. 2009. 'Genome Sequence, Comparative Analysis, and Population Genetics of the Domestic Horse'. Science 326(5954): 865–67.

Wahl, Herbert A. 1940. 'CHROMOSOME NUMBERS AND MEIOSIS IN THE GENUS CAREX'. American Journal of Botany 27(7): 458–70.

Wang, Bo et al. 2022. 'High-Quality Arabidopsis Thaliana Genome Assembly with Nanopore and HiFi Long Reads'. Genomics, Proteomics & Bioinformatics 20(1): 4–13.

Wang, Kejian et al. 2012. 'The Role of Rice HEI10 in the Formation of Meiotic Crossovers' ed. F. Chris H. Franklin. PLoS Genetics 8(7): e1002809.

Wang, Na, Jonathan I. Gent, and R. Kelly Dawe. 2021. 'Haploid Induction by a Maize Cenh3 Null Mutant'. Science Advances 7(4): eabe2299.

Wang, Ying et al. 2006. 'Euchromatin and Pericentromeric Heterochromatin: Comparative Composition in the Tomato Genome'. Genetics 172(4): 2529–40.

Waye, John S., and Huntington F. Willard. 1985. 'Chromosome-Specific Alpha Satellite DNA: Nucleotide Sequence Analysis of the 2.0 Kilobasepair Repeat from the Human X Chromosome'. Nucleic Acids Research 13(8): 2731–43.

Weide, R. et al. 1998. 'Paracentromeric Sequences on Tomato Chromosome 6 Show Homology to Human Satellite III and to the Mammalian CENP-B Binding Box'. Molecular and General Genetics MGG 259(2): 190–97.

Wells, Jonathan N., and Cédric Feschotte. 2020. 'A Field Guide to Eukaryotic Transposable Elements'. Annual Review of Genetics 54(1): 539–61.

White, Thomas A., Magnus Bordewich, and Jeremy B. Searle. 2010. 'A Network Approach to Study Karyotypic Evolution: The Chromosomal Races of the Common Shrew (Sorex Araneus) and House Mouse (Mus Musculus) as Model Systems'. Systematic Biology 59(3): 262–76.

Wibowo, Anjar Tri et al. 2022. 'Predictable and Stable Epimutations Induced during Clonal Plant Propagation with Embryonic Transcription Factor' ed. Nathan M. Springer. PLOS Genetics 18(11): e1010479.

Wicker, Thomas et al. 2007. 'A Unified Classification System for Eukaryotic Transposable Elements'. Nature Reviews Genetics 8(12): 973–82.

Wijnker, E, and H Dejong. 2008. 'Managing Meiotic Recombination in Plant Breeding'. Trends in Plant Science 13(12): 640–46.

Wijnker, Erik et al. 2013. 'The Genomic Landscape of Meiotic Crossovers and Gene Conversions in Arabidopsis Thaliana'. eLife 2: e01426.

Wilhelm, M. 2001. 'Reverse Transcription of Retroviruses and LTR Retrotransposons'. 58.

Williams, Byron C., Terence D. Murphy, Michael L. Goldberg, and Gary H. Karpen. 1998. 'Neocentromere Activity of Structurally Acentric Mini-Chromosomes in Drosophila'. Nature Genetics 18(1): 30–38.

Wright, Stephen I., Newton Agrawal, and Thomas E. Bureau. 2003. 'Effects of Recombination Rate and Gene Density on Transposable Element Distributions in Arabidopsis Thaliana'. Genome Research 13(8): 1897–1903.

Yadav, Vikas et al. 2018. 'RNAi Is a Critical Determinant of Centromere Evolution in Closely Related Fungi'. Proceedings of the National Academy of Sciences 115(12): 3108–13.

Yang, Sihai et al. 2012. 'Great Majority of Recombination Events in Arabidopsis Are Gene Conversion Events'. Proceedings of the National Academy of Sciences 109(51): 20992–97.

Yao, Jianhui et al. 2013. 'Plasticity and Epigenetic Inheritance of Centromere-Specific Histone H3 (CENP-A)-Containing Nucleosome Positioning in the Fission Yeast'. Journal of Biological Chemistry 288(26): 19184–96.

Ye, Anna A., Stuart Cane, and Thomas J. Maresca. 2016. 'Chromosome Biorientation Produces Hundreds of Piconewtons at a Metazoan Kinetochore'. Nature Communications 7(1): 13221.

Yelina, Nataliya E. et al. 2012. 'Epigenetic Remodeling of Meiotic Crossover Frequency in Arabidopsis Thaliana DNA Methyltransferase Mutants' ed. Gregory S. Barsh. PLoS Genetics 8(8): e1002844.

Yelina, Nataliya E. et al. 2015. 'DNA Methylation Epigenetically Silences Crossover Hot Spots and Controls Chromosomal Domains of Meiotic Recombination in Arabidopsis'. Genes & Development 29(20): 2183–2202.

Yin, Yan et al. 2011. 'The E3 Ubiquitin Ligase Cullin 4A Regulates Meiotic Progression in Mouse Spermatogenesis'. Developmental Biology 356(1): 51–62.

Yoda, Kinya et al. 2000. 'Human Centromere Protein A (CENP-A) Can Replace Histone H3 in Nucleosome Reconstitution in Vitro'. Proceedings of the National Academy of Sciences 97(13): 7266–71.

Yu, Zhouliang et al. 2015. 'Dynamic Phosphorylation of CENP-A at Ser68 Orchestrates Its Cell-Cycle-Dependent Deposition at Centromeres'. Developmental Cell 32(1): 68–81.

Yunis, Jorge J., and Walid G. Yasmineh. 1971. 'Heterochromatin, Satellite DNA, and Cell Function: Structural DNA of Eucaryotes May Support and Protect Genes and Aid in Speciation.' Science 174(4015): 1200–1209.

Zhang, Cheng et al. 2012. 'The Arabidopsis Thaliana DSB Formation (AtDFO) Gene Is Required for Meiotic Double-Strand Break Formation: AtDFO Is Required for DSB Formation'. The Plant Journal 72(2): 271–81.

Zhang, Liangran et al. 2014. 'Topoisomerase II Mediates Meiotic Crossover Interference'. Nature 511(7511): 551–56.

Zhang, Z., and W. I. Wood. 2003. 'A Profile Hidden Markov Model for Signal Peptides Generated by HMMER'. Bioinformatics 19(2): 307–8.

Ziolkowski, Piotr A. et al. 2017. 'Natural Variation and Dosage of the HEI10 Meiotic E3 Ligase Control Arabidopsis Crossover Recombination'. Genes & Development 31(3): 306–17.