Josip Juraj Strossmayer University of Osijek

University of Dubrovnik

Ruđer Bošković Institute, Zagreb

University Postgraduate Interdisciplinary Doctoral Study

Molecular Biosciences

Martina Pavlek

Mechanisms of satellite DNA sequence dynamics in complex genomes

doctoral thesis

Zagreb, 2015.

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište Josipa Jurja Strossmayera u Osijeku **Doktorska disertacija**
Sveučilište u Dubrovniku
Institut Ruđer Bošković
Sveučilišni poslijediplomski interdisciplinarni
doktorski studij Molekularne bioznanosti

Znanstveno područje: Prirodne znanosti
Znanstveno polje: Biologija

Mehanizmi uključeni u dinamiku satelitnih DNA u kompleksnim genomima

Martina Pavlek

Disertacija je izrađena u: Laboratoriju za strukturu i funkciju heterokromatina, Institut Ruđer Bošković, Zagreb

Mentor/i: dr. sc. Nevenka Meštrović Radan

Kratki sažetak doktorske disertacije:
Satelitne DNA su uzastopno ponovljene nekodirajuće sekvence smještene uglavnom u (peri)centromernom području. Mehanizmi njihovog nastanka, evolucije i širenja su još uvijek nedovoljno poznati i to posebice u necentromernim područjima. Analize sekvenci biblioteke satelitnih DNA u srodnom vrstama oblića roda *Meloidogyne* otkrile su jednostavnu i kompleksnu organizaciju te moguću ulogu kratkih očuvanih motiva kao promotora genomskih rearanžmana. Potraga za novim satelitnim DNA u sekvenciranom genomu kornjaša *T. castaneum* otkrila je 9 novih necentromernih satelitnih DNA, jednoliko zastupljenih u hetero i eukromatinu. Intenzivna izmjena među homolognim i nehomolognim kromosomima ukazuje na učinkovit mehanizam širenja ovih sekvenci u necentormernim regijama.

Broj stranica: 104
Broj slika: 38
Broj tablica: 11
Broj literaturnih navoda: 110
Jezik izvornika: engleski
Ključne riječi: satelitna DNA, očuvane regije, organizacija višeg reda, rearanžmani sekvenci, cjelokupna analiza genoma, *Meloidogyne* sp., *Tribolium castaneum*

Datum obrane: 10.4.2015.

Stručno povjerenstvo za obranu:
1. dr. sc. Ivica Rubelj, viši znanstveni suradnik Instituta Ruđer Bošković u Zagrebu, predsjednik
2. dr. sc. Nevenka Meštrović Radan, znanstvena suradnica Instituta Ruđer Bošković u Zagrebu, mentorica i član
3. prof. dr. sc. Vera Cesar, redovita profesorica Sveučilišta Josipa Jurja Strossmayera u Osijeku Odjela za biologiju, član
4. dr. sc. Branka Bruvo Mađarić, znanstvena suradnica Instituta Ruđer Bošković u Zagrebu, zamjena člana

Disertacija je pohranjena u: Nacionalnoj i sveučilišnoj knjižnici Zagreb, Ul. Hrvatske bratske zajednice 4, Zagreb;Gradskoj i sveučilišnoj knjižnici Osijek, Europska avenija 24, Osijek; Sveučilištu Josipa Jurja Strossmayera u Osijeku, Trg sv. Trojstva 3, Osijek

BASIC DOCUMENTATION CARD

Josip Juraj Strossmayer University of Osijek **PhD thesis**
University of Dubrovnik
Ruđer Bošković Institute
University Postgraduate Interdisciplinary Doctoral Study of
Molecular biosciences

Scientific Area: Natural sciences
Scientific Field: Biology

Mechanisms of satellite DNA sequence dynamics in complex genomes

Martina Pavlek

Thesis performed at: Laboratory for structure and function of heterochromatin, Ruđer Bošković Institute, Zagreb

Supervisor/s: dr. sc. Nevenka Meštrović Radan

Short abstract:
Satellite DNAs are tandemly repeated non coding sequences preferentially located in (peri)centromeric part of the genome. Mechanisms involved in their genesis, evolution and spread in complex genomes are still not well understood, especially in non-centromeric regions. Sequence analyses of satDNAs in the library of sister species from nematode genus *Meloidogyne* revealed simple and complex type (HOR) of organization and proposed short conserved motifs as possible promotors of genomic rearrangements. Genome-wide search for new satDNAs in sequenced genome of coleopteran *T. castaneum* led to identification of 9 non-centromeric satDNA families, evenly distributed within the putative heterochromatin and euchromatin. Extensive exchange between homologous and non-homologous chromosomes, suggests efficient propagation mechanism of tandem repeats in non-centromeric regions.

Number of pages: 104
Number of figures: 38
Number of tables: 11
Number of references: 110
Original in: english
Key words: satellite DNAs, conserved regions, higher-order repeats, sequence rearrangements, genome-wide analysis, *Meloidogyne* sp., *Tribolium castaneum*

Date of the thesis defense: 10.4.2015.

Reviewers:
1. dr. sc. Ivica Rubelj, PhD, Senior Associated Researcher, Ruđer Bošković Institute, Zagreb
2. dr. sc. Nevenka Meštrović Radan, PhD, Associated Researcher, Ruđer Bošković Institute, Zagreb
3. prof. dr. sc. Vera Cesar, PhD, Josip Juraj Strossmayer University of Osijek
4. dr. sc. Branka Bruvo Mađarić, PhD, Associated Researcher, Ruđer Bošković Institute, Zagreb

Thesis deposited in: National and University Library in Zagreb, Ul. Hrvatske bratske zajednice 4, Zagreb; City and University Library of Osijek, Europska avenija 24, Osijek; Josip Juraj Strossmayer University of Osijek, Trg sv. Trojstva 3, Osijek

Research for this PhD thesis was performed in the Laboratory for Structure and Function of Heterochromatin, Division of Molecular Biology, Ruđer Bošković Institut, Zagreb, under the supervision of dr.sc. Nevenka Meštrović Radan. Graduate studies were completed under the program of University Postgraduate Interdisciplinary Doctoral Study of Molecular biosciences.

Acknowledgments

I thank …

… my supervisor Nevenka for great faith in me, for exceptional guidance and for all the support and liberty in my research

… Marta Žižek and Ana Car for contributing to the lab work and Yevgeniy Gelfand for reprograming TRDB database for my research

… Eva and Tanja for making lab work so much fun, for all technical help and for inspiring and constructive discussions

… Brankica and Miro for useful suggestions and advices during my research

… informatics geniuses **Marko**, Filip and Hrvoje for solving all unsolvable problems on my computer and thus making large part of my research easier

… all colleagues and friends from the Institute who made my PhD years less a job and more a fun ride

… my parents for their understanding and support during this long path, to my brothers and all the family who supported and believed in me all the way

Table of contents

1. INTRODUCTION

1.1. Satellite DNA sequence features

Satellite DNAs (satDNAs) are tandemly repeated non coding DNA sequences that make a large part of almost all eukaryotic genomes, e.g. 50% of human genome is made of repetitive sequences (Lander et al. 2001) and 30% of pericentromeric satellite DNA; 20% of fruit fly (Kapitonov and Jurka 2003) and 40% of rice (Goff et al. 2002). Genomic regions extremely enriched in satDNAs are (peri)centromeric regions of chromosomes. A comprehensive bioinformatics analysis of centromeric satDNAs in a number of animal and plant species confirmed the rapid evolution of DNA sequences in these areas (Melters et al. 2013) although there are some satellite DNAs that are conserved in nucleotide sequence for long periods of time. Recent progress in genome sequencing has revealed that satDNAs also represent a substantial fraction of noncentromeric regions (Warburton et al. 2008). Absence of coding potential, extreme diversity of satDNAs and lack of direct evidence for any possible function(s) of satDNAs in a genome addresses an important question: why and how do satellite sequences accumulate in a genome?

SatDNAs are organized as long arrays of head-to-tail linked basic repeat units, monomers. Monomers are often A+T rich and can vary in size from few nucleotides only to more than 1 kb, building arrays up to 100 Mb in length. Nevertheless, monomer length between 150-180 and 300-360 bp was observed in many satellite DNAs and can be considered as evolutionarily favored. The current hypothesis links preferred monomer length and the length of DNA wrapped around 1 or 2 nucleosomes as a requirement that may facilitate regular phasing of nucleosomes in the centromere region (Zhang et al. 2013, Heslop-Harrison and Schwarzacher 2013).

Although satellite sequence can be extremely divergent, a common feature of many satDNAs is irregular distribution of sequence variability along the monomer sequence and formation of conserved sequence segments, probably because of evolution under selective constraints (Plohl et al. 2008). The most prominent examples are found in rice (Lee et al. 2006), nematodes (Meštrović et al. 2006), *Arabidopsis* and human (Hall et al. 2003). For example, short conserved motifs detected between centromeric satellite DNAs of rice and maize may represent functional elements originating from the ancestral sequence, arising about 50– 70 Myr ago (Cheng et al. 2002).

However, among all detected conserved regions, the only function is assigned to the CENP-B box of alpha satDNA in human and other primates. It is a 17 bp long motif that binds

to the centromere CENP-B protein. It has been proposed that CENP-B participates in human centromere assembly (Masumoto et al. 1993). CENP-B is highly conserved protein in all mammalian centromeres, from human to marsupials (Earnshaw and Tomkiel 1992, Bulazel et al. 2006). Motifs highly similar in sequence to CENP-B box have also been found in divers mammalian (Kipling et al. 1995, Alkan et al. 2011) (Fig. 1.1.) and non-mammalian species (López and Edström 1998, Fantaccione et al. 2005) as a part of completely unrelated satellite DNAs but their functionality was not evaluated. In addition to its putative centromeric role of CENP-B protein, extensive sequence similarity of the CENP B and *pogo*-like transposases opens a speculative possibility that DNA-CENP B protein complex also promote recombination processes involved in maintenance of the satellite DNA arrays (Casola et al. 2008, Jaco et al. 2008).

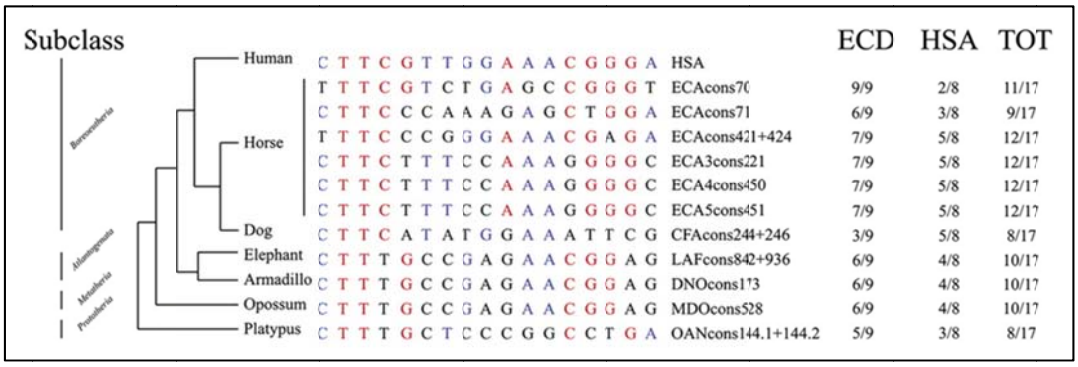


Figure 1.1. CENPB box-like motifs extracted from consensus sequences. Conserved bases in the evolutionarily conserved domain (ECD) have been reported in red and conserved bases compared with human (HSA), other than the ECD domain, are reported in blue. The number of total conserved bases is reported in *last* column. On the left side is a phylogenetic tree according to Prasad et al. 2008. (figure taken from Alkan et al. 2011)

Another feature of satellite DNAs determined by their nucleotide sequence is secondary and tertiary structure, namely dyad structures and sequence- induced bent helix axis which can be involved in heterochromatin formation (Jonstrup et al. 2008) and specific recognition of DNA- binding protein components of the heterochromatin (Radic et al. 1992), respectively. Short inverted sequence segments, often detected in satellite monomers, form dyad structures and may be recognized by mechanisms related to transposition. Bent helix axis of the satellite monomer and a resulting structure of the DNA molecule composed of tandemly repeated monomers are induced by periodic distribution of nucleotides,

particularly by distribution of short tracts of As and/or Ts phased with a turn of double helix (Martínez-Balbás et al. 1990).

1.2. Evolution of satellite DNA

1.2.1. Dynamic of satellite DNA evolution

As a part of non-coding genome high mutation rate is tolerated and indeed satellite DNAs are one of the most rapidly evolving DNA sequences in eukaryotic genome that show differences even among closely related species (Henikoff et al. 2001). In contrast, sequence divergence between monomers of the same family is often low, usually up to 15% (see for example King and Cummings 1997). However, the divergence can be much higher in some cases, such as in monomeric human alpha-satellite with repeating units divergence up to 30% (Rudd and Willard 2004) or lower as detected in homogenous satDNA of bees with only 1.4 % monomer divergence (Tares et al. 1993). High monomer homogeneity in satDNA family is achieved by non-independent evolution of monomers through the two-level process of molecular drive. Mutations are homogenized throughout members of a satDNA family by mechanisms of non-reciprocal sequence transfer and concomitantly fixed within a group of reproductively linked organisms as a result of random assortment of genetic material (Dover 1982, 1986). The consequence is concerted evolution of monomers with higher sequence similarity of a satellite family among the same then between different species and finally formation of species specific satellite DNAs (Fig. 1.2.1.a). Level of sequence variability in a satellite DNA is therefore equilibrium between the process of accumulation of mutations and the rate of their spread (or elimination) among satellite monomers.

Homogenization mechanisms also affect repeat organization of some satDNAs, resulting in higher-order repeats (HORs) (see section 2.3.1. Complex organization of alpha satDNAs). An important outcome of mechanisms involved in homogenization is a higher degree of sequence similarity observed among adjacent repeats than among those retrieved at random. Monomers can often be clustered into satDNA subfamilies which are usually chromosome-specific (Dover 1986, Willard and Waye 1987). Sequence divergences accumulate because of higher homogenization efficiency among adjacent monomers than among those positioned in different arrays on the same chromosome, and progressively, on homologous and heterologous chromosomes (Dover 1986).

Besides gradual sequence evolution of satDNAs in separate lineages, it has been also suggested that changes in the number of copies could produce species-specific satDNA as a result of a differential amplification of satDNAs which coexist in closely related species building satellite DNA library (Fig. 1.2.1.b). This model has originally been suggested by Fry and Salser (1977) analyzing a satellite DNA from the kangaroo rat, but the first experimental evidence of this concept was provided by studying satellite DNAs in insects of the genus *Palorus* (Meštrović et al. 1998). Study of the 4 unrelated species-specific dominant centromeric satellite DNAs revealed presence of low-copy counterparts of each of them in every examined species. Comparisons of high-copy and low-copy monomer variants of these satellites showed complete lack of any species-diagnostic mutations. Copy number changes may be accompanied by rapid change of satellite DNA profile and can explain species-specificity of satellite profiles even when satellite sequences remain "frozen" during long evolutionary periods (Mravinac et al. 2002). Not only distinct satellite DNAs, but also monomer variants of the single family can be distributed in genomes in the form of a library. For example, broad distribution of BIV160 satDNA indicates that library of similar variants existed in bivalve species for about 540 million years (Plohl et al. 2010). Until now, satellite DNA libraries were detected in various plant and animal taxa, probably representing the most common mode of satellite DNA evolution for example (Lin and Li 2006, del Bosque et al. 2011, Koukalova et al. 2010).

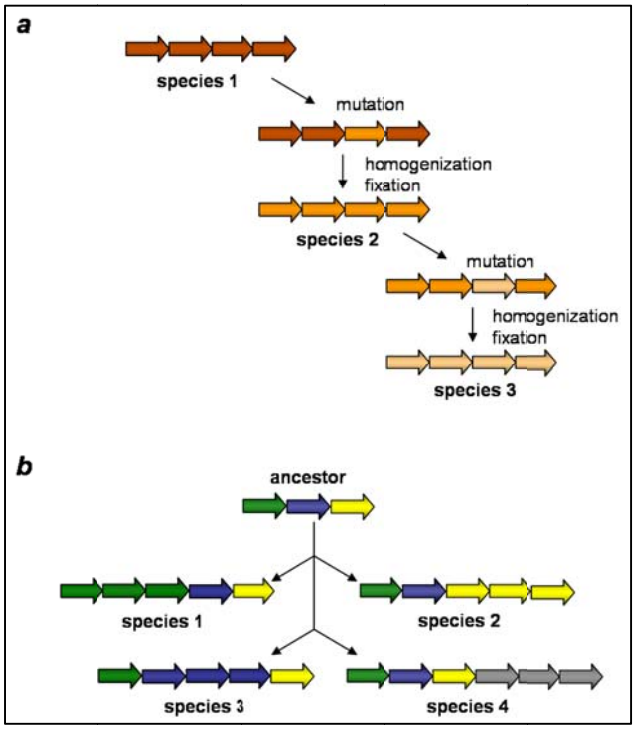


Figure 1.2.1. Schematic representation of satellite DNA evolutionary concepts. **a)** Concerted evolution. Satellite DNA is changed due to gradual accumulation of sequence divergence. **b)** Satellite DNA library concept. Variation in satellite profiles is obtained by changes in copy number. (figure taken from Plohl et al. 2012)

1.2.2. Mechanisms of satellite DNA evolution

Mechanisms by which satellite DNAs are homogenized are unequal crossover, gene conversion, rolling circle replication and transposition (Fig. 1.2.2. and Fig. 1.2.3.) (Dover 1986, Smith 1976, Stephan 1986). Unequal crossover and gene conversion have been identified as the most widespread mechanism involved in satellite DNA dynamics through recombination during meiosis when homologue chromosomes become physically linked (Mahtani and Willard 1998, Talbert and Henikoff 2010). Unequal exchange can lead to expansion of new repeat variants and/or formation of higher-order repeats, as well as eliminating variation in monomers (Fig. 1.2.3.). Subsequent unequal crossovers between pairs of tandem array blocks either tandemly duplicate or delete an integral number of blocks.

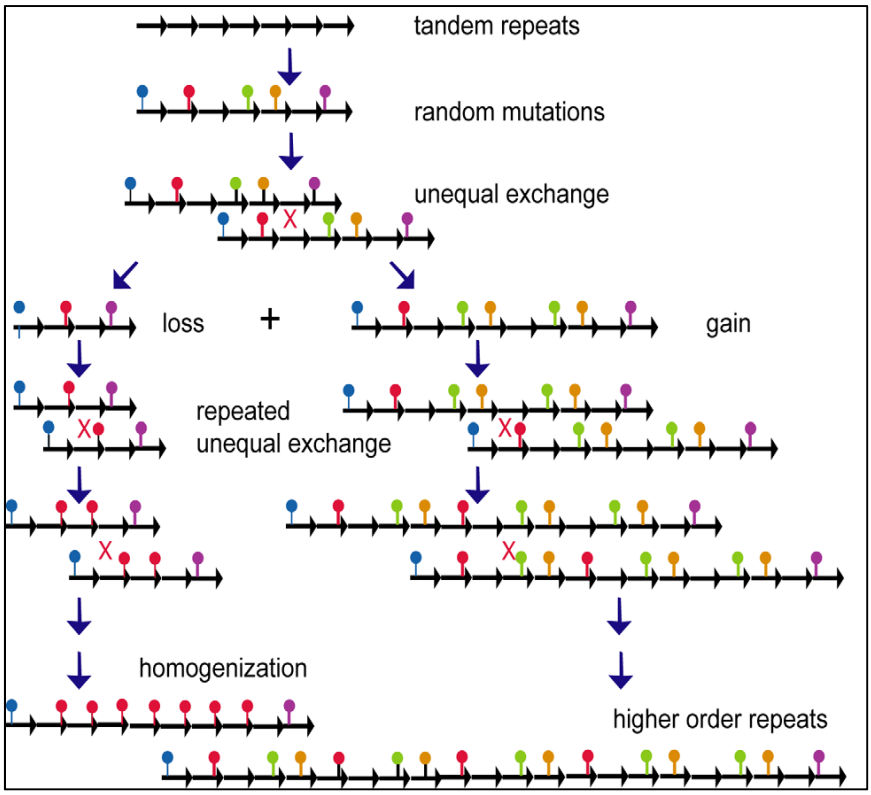


Figure 1.2.2. Unequal exchange in satellite arrays. Identical tandem satellite repeats become diversified over time by mutation. Unequal exchange results in gain or loss of tandem repeats. Repeated exchange can lead to homogenization of satellite repeats (left). If the unit of exchange consists of multiple diverged monomers, higher-order repeats are generated (right). (figure taken from Talbert and Henikoff 2010)

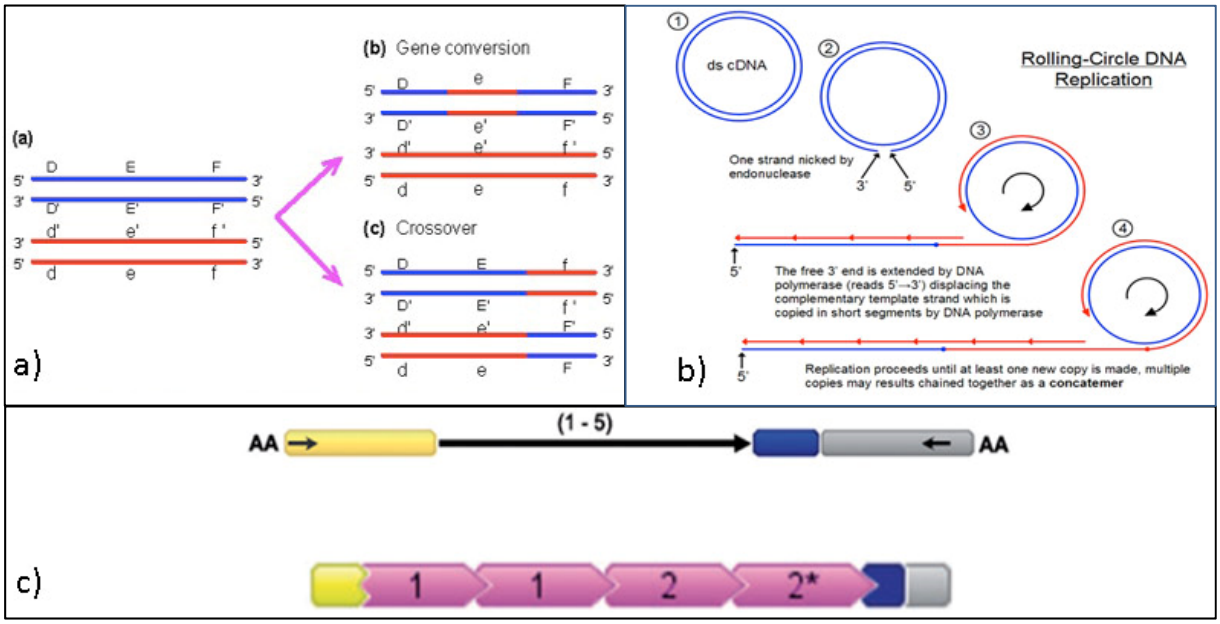


Figure 1.2.3. Schematic representation of **a)** gene conversion, **b)** rolling circle replication and **c)** transposition - modular structure of one transposable element and the same element with 4 tandem repeats as core elements.

Gene conversion is the nonreciprocal exchange of genetic material that is initiated by DNA double-strand breaks and repaired by copying a short (usually 2 kb or less) stretch of the homologous chromosome. Dispersion of satellite DNAs on heterologous chromosomes can be explained by other two mechanisms, rolling circle replication and transposition. The discovery of human extrachromosomal circular DNA originating from satDNA supports the idea that excision, rolling circle replication and reinsertion of eccDNA plays a significant role in the evolution of satellite repeats. Growing number of reports indicate a link between transposable elements and satDNAs in the genome suggesting that efficient dispersion of satellite sequences throughout the genome can be facilitated by mechanisms related to transposition. For example, 2-6 tandem repeats that are related in sequence and monomer length to broadly distributed BIV160 satellite family among mollusks, are found as a part of sequences resembling MITEs in clam *Crassostrea virginica* (Plohl et al. 2010, Gaffney et al. 2003). Also, high copy number MITE, named DTC84, is characterized in the clam *Donax trunculus* and one of its organization features is presence of core repeats (Šatović and Plohl 2013). More examples are found in *Drosophila* sp. (Miller et al. 2000) and in the cycas *Zamia paucijuga* (Cafasso et al. 2003) but the true nature of mechanism(s) that expand fragments of mobile or mobile- like elements into long arrays of satellite DNAs is not known. While evolution of sequence segments from mobile elements to satellite DNAs seems to be a

logical scenario, the possibility that satellite DNA fragments were simply captured by mobile elements is also open (Kejnovsky et al. 2006). At this point it may be speculated that both pathways are possible and that a sequence can be repeatedly reverted from one organizational form to the other. Recent study of repetitive sequences in rice centromere also indicate that segmental duplication of large arrays of satellite repeats is primarily responsible for the amplification of satellite repeats (Ma and Jackson 2006).

1.3. Satellite DNAs and centromeres

The centromere is a chromosomal locus responsible for the faithful segregation of genetic material during cell division. The majority of eukaryotes studied in terms of centromeric DNA have monocentric chromosomes with large regional centromeres (Fig. 1.3.1.a). The centromere includes the core or functional centromere domain, a specialized locus at which microtubules attach to the complex multiprotein structure of the kinetochore in order to segregate chromosomes in mitosis and meiosis. Two classes of highly abundant repetitive sequences, satDNAs and transposable elements (TEs), represent major DNA components of many centromeric regions of monocentric chromosomes. Centromere paradox is that despite of the extreme diversity of satDNA sequences in this region, centromere function is very highly preserved throughout all eukaryotes. It is proposed that fast evolving satellite DNAs push adaptive evolution of centromeric histones (CENH3) which in turn serve as a link between fast evolving satellite DNAs and preserved kinetochores (Henikoff et al. 2001). In the absence of a universal DNA sequence, species-specific histone H3 variant CENH3 is the most prominent protein identifier of centromere function.

Centromeric regions harbor as much as megabases of satellite DNA even though much less is enough for centromere function. For example, functional part of *Drosophila* sp. centromere is 15-40 kb long satDNA array or 30-70 kb in artificial human chromosome (Okamoto et al. 2007). Generally, huge differences in copy number of satellites in (peri)centromeric regions among homologous and heterologous chromosomes are very common (Plohl et al. 2012) even between individuals (Altemose et al. 2014).

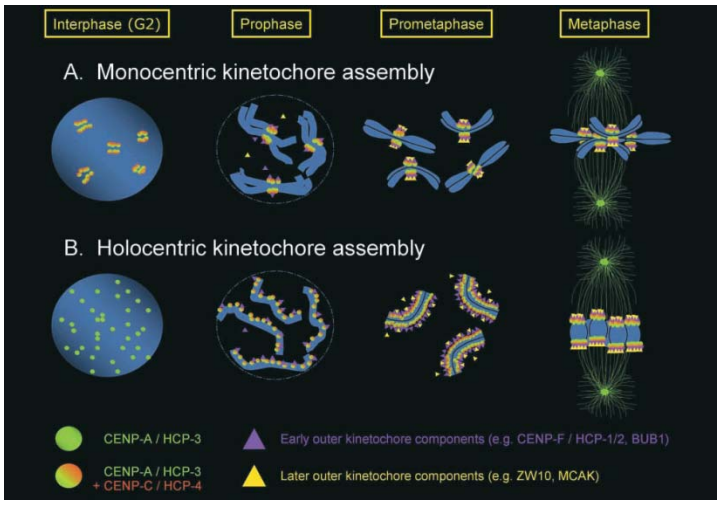


Figure 1.3.1. Assembly of kinetochores on monocentric (A) and holocentric (B) chromosomes (figure taken from Dernburg 2001)

Holocentric centromere organization, with kinetochore forming in a plate shape along the whole length of the chromosomes (Dernburg 2001) (Fig. 1.3.1.b), has arisen independently at least 13 times during species evolution (Mola and Papeschi 2006) but DNA sequences underlying these centromeres are very poorly known. However, comprehensive study of satDNA in centromeric regions of different plant and animal species show that genomic content of tandem repeats in holocentric species differs greatly (Melters et al. 2013). Centromeric function in holocentric species, based on immunodetection of CENH3 homologs, has been intensively analyzed only in the nematode *Caenorhabditis elegans* and in few other species. For example, *C. elegans* genome contains only a few tandem repeats (Hillier et al. 2007) while ChIP analysis shows that ~50 % of this genome is associated with CENH3, but CENH3 loci are not correlated with repeat density (Gassmann et al. 2013). In contrast, comprehensive characterization of holocentric *Luzula elegans* shows that 61 % of its genome is built of highly repetitive DNAs, including over 30 highly divergent satellite families, while 33 % of the genome comprises Ty1/copia LTR retrotransposons of the Angela clade (Heckmann et al. 2013).

1.3.1. Complex organization of alpha satDNAs

One of the most extensively studied repetitive DNA families is the primate specific alpha satellite DNA. Monomer of human alpha satellite is 171- bp long monomer and satellite arrays are categorized into two basic types according to their genomic organization and sequence properties: monomeric and higher order repeats (HOR). HOR fraction represent centromere core which is associated with centromere function, while monomeric arrays constitute the flanking pericentromeric heterochromatin and are not involved in centromeric activity (Fig. 1.3.2.).

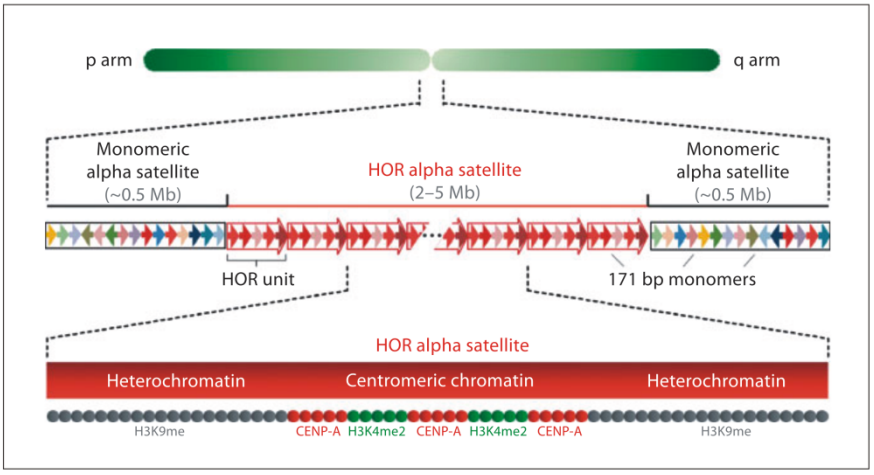


Figure 1.3.2. Structural organization of human (peri)centromeric regions. A typical human chromosome is schematically delineated, emphasizing (peri)centromeric regions. Small arrows in different colors represent single monomers of alpha satellite DNA, while HOR units are indicated by large red arrows. A fraction of HORalpha satellite forms centromeric chromatin, built from subdomains of nucleosomes containing centromeric histone CENP- A (red circles) interspersed with histone H3 dimethylated at lysine 4 ($H_3K_4me_2$) (green circles). The remainder of alpha satellite DNA is assembled into heterochromatin enriched for nucleosomes containing histone H3 methylated at lysine 9 (H_3K_9me) (grey circles). (figure taken from Plohl et al. 2012 and based on Schueler and Sullivan 2006)

HORs are based on multimeric repeats with 2 to over 20 diverged monomers. Monomers within a HOR show an average pairwise sequence similarity of ~70%, until mutual HOR similarity is 97–100%. HOR arrays spread up to several megabases and in most cases appear to be chromosome specific. HOR alpha satellite associates with CENH3 and some monomers within HOR units contain the CENP- B box. Monomeric arrays are heterogeneous without any ordered periodicity and individual repeats have identity of 50-100%. These heterogeneous arrays are frequently interrupted by other satellites and interspersed repetitive elements such as LINEs, SINEs, and LTRs (Plohl et al. 2012).

Alpha satellite DNA is a widespread repetitive family within the primate lineage and it shows concerted manner of evolution (Willard and Waye 1987). All characterized primate species have a common monomer length of ~170 bp but monomer sequence and structural characteristics are very diverse among species (Fig. 1.3.3.). During evolution, at the time of the very first amplification steps, several different variants emerged (Alexandrov et al. 2001) and by their combining more complex repeating units ascended, among which the simplest are dimeric HOR units found in the genomes of Old and New World monkeys. They are also characterized by no chromosome specificity which appeared within the last 25– 35 Myr of primate evolution.

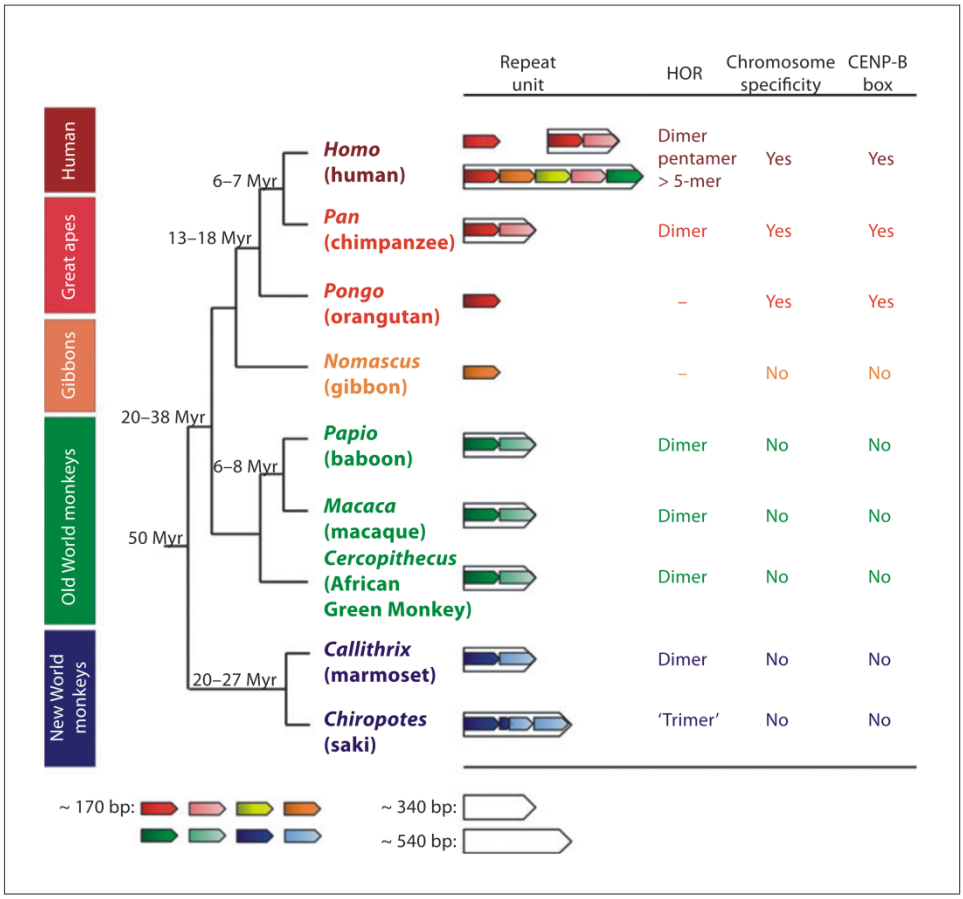


Figure 1.3.3. Structural properties of alpha satellite DNA in primates. Schematic illustration of repeating units. The form of HOR units, chromosome specificity of satellite suprachromosomal families as well as the presence of CENP- B box within the sequence is indicated. Phylogenetic relationships and approximate divergence dates are derived from the tree of living primates (Perelman et al. 2011). (figure taken from (Plohl et al. 2012)

1.4. Genome-wide analyses of satellite DNAs

Satellite DNAs are very unevenly distributed among genomes. In contrast to centromeric satDNAs which have been characterized in many plant and animal species studies of tandem repeats out of centromere regions were mainly carried out for a micro- and minisatellites. Since they are mostly part of repetitive heterochromatin, which generally lacks genes and other unique sequences, they are almost absent in most sequenced genomes which makes them hard to study. The repetitive nature of satDNAs makes them hard to sequence and even harder to assemble. For example, in coleopteran *Tribolium castaneum* 30% of assembled genome are repetitive sequences (Wang et al. 2008) but high copy centromeric satellite TCAST, whose abundance is experimentally estimated to as much as 35% (Feliciello et al. 2011), is highly underrepresented in assembled genome (Richards et al. 2008). However, recent development of appropriate bioinformatics algorithms and

programs together with raising abundance of sequenced genomes opens a possibility for extensive and throughout analyses of whole genomes in terms of satDNA profile.

Warburton and others (2008) analyzed repetitive profile in euchromatic portion of human genome. They discovered it harbors many satellite DNAs, even 10 kb long fragments of tandemly repeated sequences (it was assumed that satellite DNAs in euchromatin are dispersed and that they don't form long arrays) which contained huge variety of monomer lenghts, from several to more than 1000 bp. In human, Satellites 2 and 3 show extreme variability in length of satellite DNA arrays, from 7 to 98 Mb, on Y chromosome among members of same population (Altemose et al. 2014). Comprehensive bioinformatics analysis of two mouse whole-genome shotgun assemblies characterized 62 newly TR families distributed on all chromosomes which make up a kind of unique chromosome bar code (Komissarov et al. 2011).

Recent studies of satDNAs in euchromatic genome region also suggest a role in modulation of gene regulation (Stam et al. 2002) in disease-associated gene mutation and acumulation the differences between closely related species which may have a phenotypic effects (Paar et al. 2011).

1. 5. Model organisms

1.5.1. Satellite DNAs in *Meloidogyne* spp

Species from genus *Meloidogyne* are root-knot plant-parasitic nematodes that cause vast damage in agriculture. The recent completion of two root-knot nematode genomes *M. incognita* (Abad et al. 2008) and *M. hapla* (Opperman et al. 2008) emphasized them as model organisms of metazoan plant parasitic species (Bird et al. 2009). In recent years extensive studies on satDNAs in many *Meloidogyne* species recovered that satellite sequence in these species evolve according to library model of evolution (Meštrović et al. 2006). Analysis of the distribution of sequence variability among three related satellite DNAs from the library revealed highly structured monomers, composed of alternating lowly variable, moderately variable and highly variable domains. Interestingly, comparison of satellite DNA sequences cloned from each species revealed that the entire monomer sequence is uniformly conserved, even in domains characterized as highly variable, although species are separated for about 45 Myr. An exceptionally complex pattern of sequence

variability was found in a family of satellite DNAs of root- knot nematode species from the genus *Meloidogyne* indicated two phases in evolution of satellites in the library (Fig. 1.5.1.).

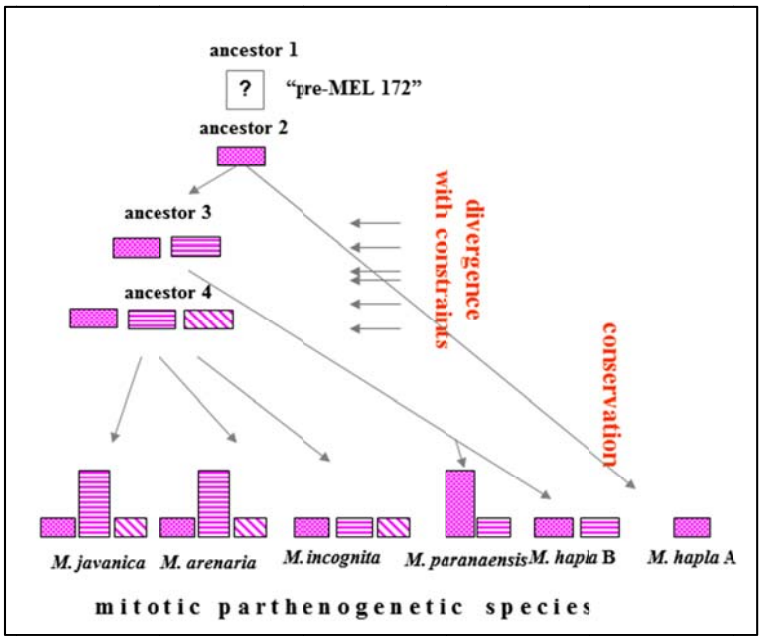


Figure 1.5.1. The two phases in evolution of MEL172 satellites from Meloidogyne species (Meštrović et al. 2006). The first phase is formation of satDNAs in the library which are shaped and spread under selective pressure due to functional interactions. The second phase includes sequence conservation and persistence of satDNAs in the library for long time-periods.

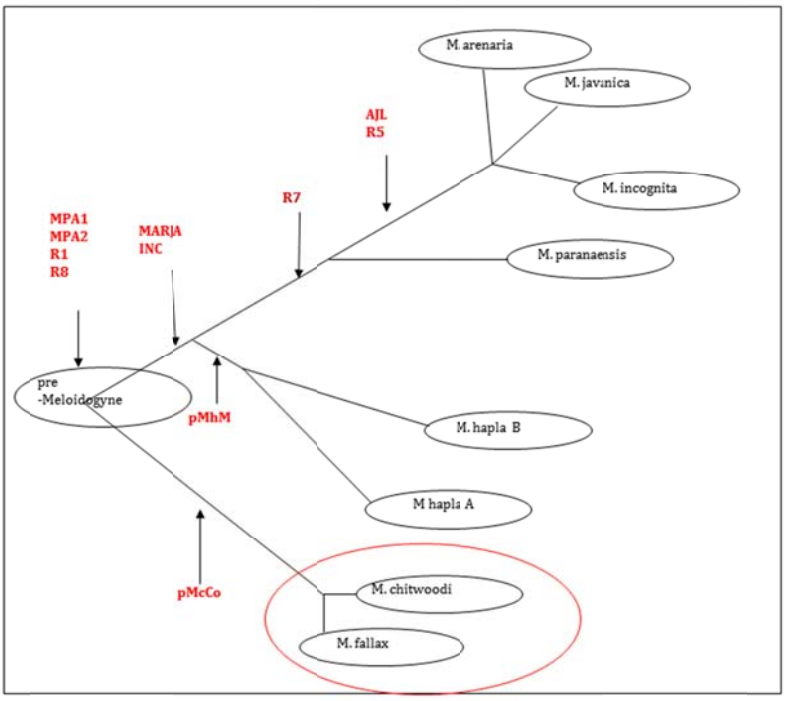


Figure 1.5.2. Root-knot nematode most parsimonious hypothetic evolutionary scenario based on the distribution of satellite DNAs. Arrows indicate the origin and distribution of satellite DNAs in root-knot nematodes. (taken from Meštrović et al. 2009, modified)

In addition, overall analysis of satDNAs in *Meloidogyne* species revealed their distribution as an informative character that can explain some aspects of evolutionary relationships, while interspecies sequence divergence did not bear any relevant information (Meštrović et al. 2009) (Fig. 1.5.1).

Previous work showed that recently separated and reproductively isolated sister species *M. fallax* and *M. chitwoodi* share one, unique, satellite DNA from pMcCo family (Castagnone-Sereno et al. 1998) (Fig. 1.5.2.). Also, *M. chitwoodi* besides that one has another 5 satellite DNAs from the same familiy which altogether form two subfamilies, subfamily 1 (1a, 1b, 1c and 1d satDNAs) and subfamily 2 (2a and 2b satDNAs) (Castagnone-Sereno et al. 1998). Possible existence of a satellite library in these recently separated species makes them an ideal system for exploring the mechanisms involved in satellite DNA formation and spread and possible requirements on their sequences.

1.5.2. Satellite DNAs in *Tribolium* spp.

The cosmopolitan insect genus *Tribolium* comprises 33 species, including the major global pests of stored grain and cereal commodities for human consumption. Among them *Tribolium castaneum*, also known as the red flour beetle, represents a powerful model organism for studies of insect development, population genetics as well as comparative genomics. An extensive research has been done on satDNAs in many *Tribolium* species (Juan et al. 1993, Mravinac et al. 2004, Ugarković et al. 1996, Mravinac et al. 2005, Žinić et al. 2000, Mravinac and Plohl 2007, Mravinac and Plohl 2010). Single species-specific satellite DNAs distributed on (peri)centromeres of all chromosomes of the complement dominate in the majority of *Tribolium* species making up to 40% of the whole genomes. The genome of *T. castaneum* has recently been sequenced at sevenfold redundancy and annotated (Richards et al. 2008). The third version of the assembly (Baylor College of Medicine; Tcas_3.0) containing more than 90% of the sequenced genome has been assembled into 10 chromosomes (Kim et al. 2010). It is the first coleopteran genome that has been sequenced thus representing the largest and the most species diverse eukaryotic order. *T. castaneum* has a large blocks of (peri)centromeric heterochromatin uniformly distributed on all chromosomes, characterized by abundant species specific satellite DNA (TCAST satDNA) which comprise 35% of the genome (Feliciello et al. 2011). However, due to the presence and abundance of nearly identical satellite monomers *T. castaneum* centromeric satDNA has

been poorly represented by only 0.3% in genome sequence assembly (by RepeatScout) with the majority of sequenced pericentromeric satDNA included in unassembled reads (Wang et al. 2008). Given that the estimated genome size of 204 Mb is 44 Mb larger than the assembled genome sequence, it is likely that the centromeric portion of satellite DNA is omitted from assembly. Three complementary approaches used for *de novo* identification of repetitive DNA content and distribution recovered more than 30% of repetitive DNA in the assembled *T. castaneum* genome (Wang et al. 2008). Analysis using Tandem Repeat Finder with parameters ≥ 2 copies, <500 pb of length, alignment score to report 30 and alignment parameters 2,7,7 recovered 4.9 % of total tandem repeats (2.5 % satellite DNA). Second approach using RepeatScout (≥50 bp) identified even 26% of repetitive DNA. Further, 4475 repeat families obtained by Repeat Scout were divided based on the percent of the genome that they occupy into classes: HighA, HighB, Mid and Low. Repetitive DNAs are not uniformly distributed among chromosomes: CH7 contains the least, while CH2, CH3, CH8, CH9 and CH10 contain the most (Wang et al. 2008). Distribution and density of repetitive DNA on each chromosome is shown in Figure 1.5.3. Position of putative heterochromatin (grey square under the plot for each chromosome in Fig. 1.5.3.) is determined by accumulation of HighA class repetitive families (Wang et al. 2008). Interestingly, in comparison with *Drosophila* genome microsatellites and minisatellites (1-6 bp and 7-100 bp per repeat unit, respectively) are less abundant in *Tribolium.* On the contrary, satellites over 100 nucleotides, which are quite rare in *Drosophila*, are prevalent in *Tribolium* genome.

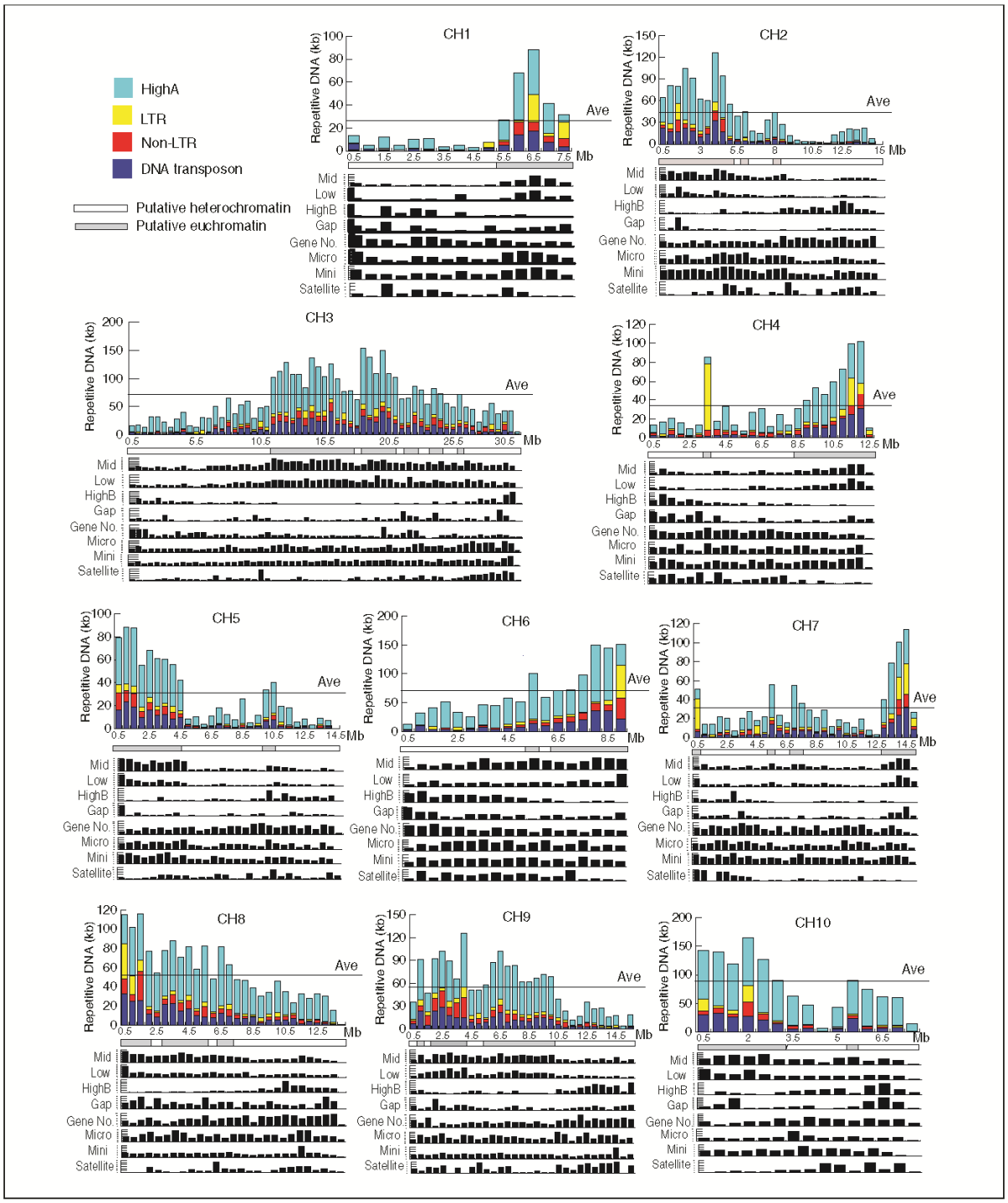


Figure 1.5.3. Density and distribution of repetitive DNA on each chromosome of *T. castaneum*. The total length (kb) of repetitive DNA in each 500 kb interval along the chromosome is plotted. The 300 kb long uncaptured gaps were not included in the chromosomes. The HighA class includes the 360 bp satellite. Gene number, gap length and distribution of other repetitive classes within the 500 kb intervals are shown below the main graph for each chromosome. The combined average of HighA repeats and TE per 500 kb along the chromosome is depicted as a black line. (Figure taken from Wang et al. 2008)

2. AIMS

Key question about satellite DNA evolution concerns the nature of mechanisms that drive formation and spread of novel tandem repeats in genomes. Satellite library in recently separated *Meloidogyne* species makes them an ideal system for exploring the mechanisms involved in satellite DNA formation and possible requirements on satellite sequences. Analysis of organization, sequence features and phylogenetic relationships of monomers in five divergent satDNAs of the library shared by *M. fallax* and *M. chitwoodi* will be performed. In addition, the genome of *T. castaneum* has recently been sequenced and identification of repetitive DNA content recovered more than 30% of repetitive DNA. Genome-wide profiling of tandem repeats on assembled *T. castaneum* genome using Tandem Repeat Finder will be done. Further, localization of the most prominent satDNAs in relation to the centromeric satDNA and investigation of relationships among tandem repeats of particular satDNA family from different loci and structure of local genomic features will be performed. The expected results should expand the knowledge about the mechanisms involved in genesis and propagation of satellite sequences in complex genomes.

3. MATERIALS AND METHODS

3.1. Materials

Laboratory chemicals and material used in this study are listed in Tables 3.1. to 3.5. Commonly used buffers and solutions are listed in Table 3.1., commercial kits in Table 3.2. and enzymes in Table 3.3. Material is listed in Table 3.4. and other chemicals Table 3.5.

Molecular marker O'GeneRuler DNA Ladder Mix (Fermentas) was used for sample size estimation and Lambda DNA 50ng/µl was used to determine sample concentration.

Table 3.1. Buffers and solutions

Buffers and solutions	**Contents**
G buffer	0.1 M NaCl, 0.01 M Tris-HCl (pH 8.0), 25 mM EDTA (pH 8.0), 0.5% SDS
TE buffer	10 mM Tris-HCl (pH 8.0), 1 mM EDTA (pH 8.0)
TAE	40 mM Tris, 20 mM acetic acid, 1 mM EDTA (pH 8.0)
SSCx20	3 M NaCl, 0.3 M Na-citrat (pH 7.0)
PBS	137 mM NaCl, 2.7 mM KCl, 10.1 mM Na_2HPO_4, 1.8 mM KH_2PO_4 (pH 7.4)
Southern hybridization solution	0.25 M Na_2HPO_4 pH 7.2, 1 mM EDTA, 20% SDS, 0.5% blocking reagent
Southern buffer 1	0.1 M maleic acid, 3 M NaCl, 0.3% Tween 20 (pH 8.0)
Southern buffer 2	1% blocking reagent in buffer 1 (Southern)
Southern buffer 3	0.1 M Tris-HCl (pH 9.5), 0.1 M NaCl
Washing buffer (Southern)	20 mM Na_2HPO_4 pH 7.2, 1 mM EDTA, 1% SDS)
Fixative solution FISH	acetic acid : aps. ethanol (1:3, v/v)
FISH buffer 1	5% (v/v) 1M $MgCl_2$ in PBS buffer
FISH buffer 2	2.7% (v/v) 37% formaldehyde, 97.3% (v/v) FISH buffer 1
FISH denaturation solution	70% (v/v) fomamide in 2XSSC buffer
DeSO4 buffer	20% (m/v) $DeSO_4$, 50 mM $NaPO_4$ (pH 7.0) in 4xSSC buffer
FISH hybridization solution	60% (v/v) formamide, 40% (v/v) $DeSO_4$ buffer
FISH washing buffer	50% (v/v) formamide, 50% (v/v) 2xSSC buffer
4M buffer	5% (m/v) blocking reagent in 4xSSC buffer
4T buffer	0.05% (v/v) Tween 20 in 4xSSC buffer
DAPI solution	50ng/ml DAPI in 2xSSC buffer
Antifade reagent	2.33% (m/v) DABCO, 8% (v/v) reH_2O, 2% (v/v) 1M Tris-HCl buffer (pH 8.0), 90% (v/v) glycerol
NaOH 0.4 M	
KCl 0.075 M	
Na-acetate 3 M (pH 7.4)	

Table 3.2. Commercial kits

Kit	Manufacturer
QIAquick Gel Extraction Kit	Qiagen
QIAquick PCR Purification Kit	Qiagen
pGEM-T Easy Vector SystemI	Promega
GoTaq Flexi DNA Polymerase	Promega
High Pure Plasmid Isolation Kit	Roche
Nick Translation Mix	Roche
Cy3 PCR Labeling Master	Jena Bioscience
DNeasy Tissue Kit	Qiagen

Table 3.3. Enzymes

Enzyme	Manufacturer
T4 DNA ligase	Promega
RNase A	Roche
Restriction enzymes: HindII, MboI, HpaII, AluI, HaeIII, HinfI, RsaI, DraI, EcoRI	Fermentas, Roche, BioLabs
Proteinase K	Roche

Table 3.4. Material

Matrials	Manufacturer
Nylon mambranes, positively charged	Roche
Röntgen film	Amersham

Table 3.5. Other chemicals

Other chemicals	Manufacturer
Cy3 Reactive Dye	Amersham
Cy3-dUTP-PCR	Jena Bioscience
Biotinylated dNTP Mixture	BioLabs
Sodium dodecyl sulfate (SDS)	Sigma
Agarose LE	Roche
X-gal	Sigma
IPTG	Invitrogene
Blocking reagent	Roche
CPD-Star	Roche
Deoxynucletide Soultion Mix	BioLabs
Fluorescent Avidin D	Vector Laboratories
Biotinylated Anti-Avidin D	Vector Laboratories
Streptavidin-AP-conjugate	Roche
Colcemide	Roche
Pepsine	Sigma
Tween 20	Promega
Fomamide	Sigma
Formaldehyde	Sigma
Acetic acid	Kemika
Phenol	Sigma
Phenol:chloroform	Sigma
Kopleksal III (EDTA-Na_2x2H_2O)	Kemika
Trizma base (Tris)	Sigma
Ethanol aps.	Kemika
Chloride acid	Kemika
DAPI	Serva Feinbiochemica
DABCO	Sigma
Glycerol	Kemika
Bacto Tryptone, Bacto Yeast extract	Becton, Dickinson & Co.
Ampicilin	Sigma
Luria Agar	Sigma

Cloning vectors pUC18 (Fermentas) and pGEM T-Easy vector (Promega) were used to transform *Escherichia coli* Subcloning Efficiency DH5α (Invitrogene) and XL10-Gold (Agilent Technologies) chemo competent and ultracompetent cells and ElectroMAX Stbl4 electrocompetent cells (Invitrogene). Bacteria were grown on liquid (for 1 liter: 10g tryptone, 5g yeast extract, 5g NaCl and ampicillin with final concentration of 100 μg/ml) and solid medium (15g of agar added per one liter of liquid medium).

Primers used for *Meloidogyne* species are listed in Table 3.6. and their position (except 2bL and R) is marked in Supplementary figures 4.1.2. and 4.1.3. Primers used for *T. castanuem* are listed in Table 3.7. Positions of Tcastan1 and Tcastan2 are marked in Figure 4.2.3., for R66_F and R66_R in Figure 4.2.16. and for all others in Supplementay figure 4.2.3.

Table 3.6. *Meloidogyne* spp.primers

Sequence name	**Primer name**	**Sequence 5'-3'**
1a satDNA	1aL	CCAAATTCAGCAAATTTCCAACGAT
	1aR	AATCCATCGACTAGTTTTTGAG
1a satDNA (HOR specific)	1aL	CCAAATTCAGCAAATTTCCAACGAT
	1a'R	GGGGAAGGAATATTTTTGAACTTTT
1b satDNA	1bL	CATATCTCTCAAAGCCTTCT
	1bR	TCGGAAGCATATTCGCTGT
1c satDNA	1cL	TCGATTCACCTCTTCATCCTC
	1cR	GGGGGGAGAATGGATACTTTG
2a satDNA	2aL	CCTCTTTCGAATGATATATGAATC
	2aR	TTCAGTAAGTTATGAGACTTGTTCC
2b satDNA	2bL	GGACTTATGAAATTGTAGGTCAGT
	2bR	GCTCTTTCGAATGATATATGAATC
U1	U1L	GGTGTAGAAGAGACAAGCCTC
	U1R	AGGGTGTTCCTTTACTCCTTC
U2	U2L	CTTGTTAGATATTTACAATTTTGG
	U2R	ATTCCCATTCTATATAGATGATG
M. fallax SCAR	Ff2	CCATTTCTGCTAAATGCCAAACTA
	Rf	GGACACAGTAATTCATGAGCTAG
M. chitwoodi SCAR	Fc2	GGCATTGACGTGCTCCGAGAGT
	Rc	GGTCTGAGTGAGGACAAGAGTA

Table 3.7. *T. castaneum* primers

Sequence name	Primer name	Sequence 5'-3'
Cluster 1 (Cl1)	kl1_F	AAGTCGGCTACGACTAACCGTTC
	kl1_R	TTGCAAATTTGGATTCCGCCCGG
Cluster 2 (Cl2)	kl2_F	TATACGCAAAATGAGCCGC
	kl2_R	AAAGTCGTAGAGCAATGCGG
Cluster 3 (Cl3)	kl3_F	CACCAAATTTGGTCGAAAATGAC
	kl3_R	CGTGTTTAAATCCTCAGAACTTGC
Cluster 4 (Cl4)	kl4_F	GTTTGTTCCAGTGAATTCTGCGG
	kl4_R	CCGTTTTGCTCTACGACTTTTAG
Cluster 5 (Cl5)	kl5_F	GGTGTTGAAAAGTCATAARTTGAGTG
	kl5_R	GAGCCGGTGTACACAACATT
Cluster 7 (Cl7)	kl7_F	CGACGCATGGGTCAATCTAAGACA
	kl7_R	ATTCGAAACTTTTCAAAAAAATTGG
Cluster 8 (Cl8)	kl8_F	GAATCGTCCGAAATAAGCCG
	kl8_R	CTGAAAACGCCTTATTCTGGC
Cluster 9 (Cl9)	kl9_F	TCATGTTCGGACAAACACC
	kl9_R	TTTTTTACAGTCGAAGGCCC
Cluster 10 (Cl10)	kl10_F	GACAGATTTGGAATCCTTAGAC
	kl10_R	CTACGATTCGTAGTTTTGGAG
TCAST	Tcastan1	TGTAGGACTAACCATAAGCG
	Tcastan2	CAATGTTTGAGACGAAGACG
plasmid primers	M13 N	GTAAAACGACGGCCAGT
	M13 R	CAGGAAACAGCTATGAC
R66-like region	R66_F	TTCATATGGCTTCTCCGTTGG
	R66_R	TATTTACTGAGGTATTGAATTTGAT

3.1.1. Animal material

The *Meloidogyne* spp. isolates used in this study were chosen from the living collection maintained at INRA, Sophia Antipolis, France. The geographic origin of both studied species was The Netherlands; Spijkenisse for *M. chitwoodi* and Baexem for *M. fallax*. Nematodes were maintained on tomatoes (*Lycopersicon esculentum* cv. Saint Pierre) grown at 20°C in a greenhouse. *Tribolium castaneum* culture (laboratory strain) used in this study was obtained from Central Science Laboratory (Sand Hutton, York, UK). Insects were maintained on flour and kept in glass jars at room temperature, in a laboratory at Ruđer Bošković Institute.

3.2. Methods

3.2.1. Sampling and DNA Isolation

Tribolium castanem genomic DNA was isolated from adults. Approximately 70 specimens were frozen in liquid air (~ -196°C), grinded, suspended in G buffer (100 mg tissue → 1 ml G buffer) containing 0.2 mg/ml of Proteinase K and left for overnight incubation at 50°C. After two rounds of phenol and phenol:chlorophorm (1:1) extraction, upper faze (containing DNA) was transferred to 2 volumes of 100% ice-cold ethanol and 0.1 volume of sodium acetate and precipitated by centrifugation for 15 minutes. The pallet was re-suspended in TE buffer. Total nematode genomic DNA was purified from 50 - 100 µl eggs using the DNeasy Tissue Kit according to the manufacturer's instructions. RNA was removed by incubation with RNaseA, 20µg/ml, for 1 hour at 37°C. Quality and quantity of DNA was checked by electrophoresis on 1% agarose gel by comparison with DNA markers of known length and concentration. Concentration was also checked on BioSpec-nano Spectrophotometer in Laboratory for molecular ecotoxicology at Ruđer Bošković Institute.

3.2.2. PCR Analyses, Cloning and Sequencing

Primers listed in Tables 3.6. and 3.7. were constructed based on consensus sequence of each of the satellite families during this work and were used to amplify specific DNA fragments in polymerase chain reaction (PCR). The reaction mixture consisted of reaction buffer, 1.5 mM $MgCl_2$, 0.2 mM dNTPs (Biolabs), 0.5 U GoTaq DNA polymerase, 0.4 mM of each primer and 20 ng of genomic DNA. The PCR cycling parameters used were as follows: 2 min initial denaturation at 94°C, followed by 30 cycles of: 94°C for 30 sec, 58°C for 30 sec, and 72°C for 1 min. Final extension was at 72°C for 10 min. Annealing temperatures for *T. castaneum* satDNAs were adjusted: 48°C for Cl10, 48.8°C for Cl9, 55°C for Cl7, 58°C for Cl1 and Cl5 and 60°C for clusters 2, 3, 4 and 8. PCR products were purified using QIAquick PCR Purification Kit and separated by electrophoresis on agarose gel. Fragments of interest were cut from the gel, purified by QIAquick Gel Extraction Kit and ligated in pGEM T-Easy vector with T4 DNA ligase. 3µl of ligation mixture was used to transform *Escherichia coli* Subcloning Efficiency DH5α Competent, XL10-Gold Ultracompetent and ElectroMAX Stbl4 electrocompetent cells. All transformations were done according to manufacturer instructions. Bacteria were grown on solid medium overnight on 37°C (chemocompetent) and 30°C (electrocompetent) and recombinant clones were selected based on blue-white

selection (40 µl X-gal and IPTG was added to each plate). Specifically, clones containing plasmid vector were selected based on antibiotic resistance and clones with recombinant plasmid were recognized by their white color opposite to blue clones which had plasmids without DNA sequence of interest. White colonies were transferred to liquid medium, grown over night to amplify and to isolate plasmids of interest using High Pure Plasmid Isolation Kit. Sequence of cloned DNA segments was determinated in Macrogene sequence centers in South Korea and Netherlands.

3.2.3. Southern and Dot Blot Analyses

Based on sequences of cloned fragments, recombinant clone for each satDNA was selected to be used as mold for oligonucleotide probe construction by polymerase chain reaction (selected clones are marked in Supplementary Figure 4.2.3.). The probes were labeled with biotin-16-dUTP (nucleotide mixture was: 0.5mM dATP, 0.5mM dCTP, 0.5mM dGTP, 0.4mM dTTP, 0.1 mM biotin-16-dUTP) and primers used were plasmid primers M13 (5'GTAAAACGACGGCCAGT3' and 5'CAGGAAACAGCTATGAC3') with which 234 bp of plasmid sequence is also amplified (a part on each side of the inserted fragment). The probe for TCAST satellite was labeled with Cy3-dUTP (Cy3 Reactive Dye and Cy3-dUTP-PCR) by nick translation using Nick Translation Mix and by polymerase chain reaction using Cy3 PCR Labeling Master, according to the protocol provided by manufacturers. The mold for TCAST probe synthesis was dimer obtained after digestion of *T. castaneum* genomic DNA with AluI and cloned into the plasmid vector pUC18. Primers used in PCR reacition were specific primers Tcastan1 and Tcastan2.

For Southern blot hybridization restriction analyses were done on all investigated genomic DNAs with different restriction endonucleases (Table 3.3.), preferentially with ones that cut once in a monomer. Fragments obtained by digestion of 5 to 8 µg of *T. casteneum* and ~3 µg of nematode genomic DNA were separated by electrophoresis on agarose gel and transferred to nylon membrane by alkaline transfer in 0.4M NaOH. After overnight transfer the membrane was washed 2x5min in 2xSSC buffer, dried, fixed by baking at 120°C for 20 minutes and then incubated for 2-4 hours in Southern hybridization solution at 60-68°C (prehybridization). For hybridization, the probe was denaturated by cooking for 10 minutes and then added to fresh Southern hybridization solution for overnight hybridization at the same temperature as prehybridization. Final probe concentration was 10-20 ng/ml. Next day

the detection procedure was: 3x20 minutes in washing buffer at 2°C below hybridization temperature, 5 minutes in Southern buffer 1 at room temperature, incubation for one hour in Southern buffer 2, incubation for 30 minutes in Southern buffer 2 with added streptavidin-AP-conjugate (1:10000), 5x10 minutes washing in Southern buffer 1 and incubation 2x5 minutes in Southern buffer 3. Finally, chemiluminescent detection of the signal was carried out by adding CDP-Star reagent to the membrane and leaving it for 15 minutes to several hours or overnight to expose the Röntgen film.

The abundance of satellite DNA sequences was estimated by quantitative dot blot analysis using a series of genomic DNA dilutions. Known concentrations of satellite monomers, excised from a plasmid, were dot-blotted and used as a calibration curve.

3.2.4. Fluorescence in situ hybridization (FISH)

Two colored fluorescence in situ hybridization (FISH) was carried out on *T. castaneum* male gonads. To determine positions of new satDNAs related to (peri)centromeric regions a mixture of Cy3 TCAST probe and biotin labeled probe for each of new satDNAs was made and in order to investigate localization of Cl5 repeats and their flanking regions we used mixture of monomer Cy3 probe amplified by kl5_F and kl5_R primers and flanking region biotin probe amplified by R66_F and R66_R primers. Slides were prepared by "squash„ technique. Male gonads were isolated from adults, transferred to colcemide (10 μg/ml) for 1 hour, then to 75 mM KCl for 15 minutes (or less) to produce hypotonic shock after which they were fixed in fixative solution for 10 minutes, transferred to slide with 100 μl of 45% acetic acid, covered with covering glass, firmly squashed, frozen in liquid air and after removing cover glass left to dry on room temperature. After drying slides can be stored on -20°C or used for hybridization right away. Pretreatment for FISH was: 5 minutes washing in preheated (37°C) 2xSSC buffer, 1 hour incubation with RNase A (20mg/ml in 2xSSC) on 37°C, 3x5 minutes washing in 2xSSC, 10 minutes incubation with pepsin (100 μg/ml in 10 mM HCl) on 37°C, 2x5 minutes washing in PBS buffer, 5 minutes washing in FISH buffer 1, 10 minutes incubation in FISH buffer 2, 5 minutes washing in PBS buffer, dehydration through series of ice cold ethanol, 70%, 90% and 100%, 3 minutes in each and then drying at room temperature. Denaturation of sample is carried out in FISH denaturation solution at 70°C for 2 minutes followed by another series of ethanol dehydration drying at room temperature. 100 ng of each probe (TCAST and one of new satellite DNAs) was mixed together and dried,

resuspended in 10 µl of FISH hybridization solution, denaturated at 75°C for 5 minutes and then applied to the sample which was left for overnight incubation at 37°C. Detection procedure was as follows: 4x5 minutes washing in FISH washing buffer at 37°C, 3x5 minutes washing in 2xSSC buffer at 37°C, 30 minutes incubation at 37°C in 4M buffer, 30 minutes incubation at 37°C in 4M buffer with fluorescent Avidin D (1:500), 3x5 minutes washing in 4T buffer, 20 minutes incubation at 37°C in 4M buffer with biotinylated Anti-Avidin D (1:100), 3x5 minutes washing in 4T buffer, 20 minutes incubation at 37°C in 4M buffer with fluorescent Avidin D (1:2000), 3x5 minutes washing in 4T buffer, 5 minutes washing in PBS buffer, dehydration through series of ice cold ethanol, 70%, 90% and 100%, 5 minutes in each and then drying at room temperature. After dyeing in DAPI solution, a drop of antifade reagent was applied on the sample which was then covered with cover glass. All signals were viewed through appropriate filters for blue (DAPI), red (CY3) and green (FITC) fluorescence, using Opton Leitz microscope equipped with Pixera Pro150ES digital camera at the Division of Molecular Biology, Department of Biology, Faculty of Science in Zagreb.

3.2.5. Bioinformatics methods

Sequenced genome of *Tribolium castanem* was downloaded in fasta format from the web page ftp://ftp.bioinformatics.ksu.edu/pub/BeetleBase/3.0/ in the form of 10 chromosomes and 2153 unassembled reads. Sequence of each chromosome was uploaded to Tandem repeats database (TRDB) (Gelfand et al. 2006), https://tandem.bu.edu/cgi-bin/trdb/trdb.exe, where it was analyzed with tandem repeats finding algorithm (TRF) (Benson 1999) for tandemly repeated sequences. Conditions that can be adjusted are alignment parameters (match, mismatch, indels) and minimum alignment score to report. The program allows the result of TRF (a series of arrays) to be filtered out for copy number (number of repeats in the array), pattern size (size of the monomer unit) and for redundant sequences. After all chromosomes have been processed with TRF and filtered out, remaining arrays have been merged in one file which was then analyzed with clustering tool. Conditions that can be adjusted are cutoff value (from 95 to 60% of similarity) and appliance of heuristical, DUST and PAM algorithm. The result of these analyses is formation of several clusters - groups of arrays that in this case represent putative satellite DNAs. Sequences of all arrays from selected clusters were downloaded in fasta format. Left and right flanking region of all clustered arrays was also downloaded, 4000 bp in length. For each of the arrays

flanking regions had to be checked for monomer residues (sometimes several full length copies) and in some cases manually adjusted. This happens because in the case of 10 to 15 bp (or more) long deletions TRF algorithm doesn't recognize the sequence after the deletion as the same array. TRF analyses were done on unassembled reads also, just to understand the rough distribution of putative satellites in them.

All *T. castaneum* downloaded and *Meloidogyne* sp. cloned sequences were blasted against NCBI GenBank Database and Repbase Update (a collection of repetitive DNAs) (Jurka et al. 2005) to check similarity with published sequences. The sequences were further analyzed in programs BioEdit 7.0.9.0. (Hall 1999) and Geneious 5.5.6. Alignments of all monomers of each of the putative satellite DNAs were done using ClustalW algorithm (Thompson et al. 1997) which also constructed consensus sequences that were used to produce specific primers for each of the putative satellite DNAs.

The same alignments (without truncated monomers from the beginning and the end of the array) were used for phylogenetic analyses. First step was to select best-fit model of nucleotide substitution for each of the alignments which was done by jModelTest 2.1.3. (Darriba et al. 2012). Obtained model and other parameters were then applied in phyML 3.0. (Guindon and Gascuel 2003) and PAUP (Swofford 2002) programs to build a phylogenetic tree for each of the alignments. The trees were adjusted and displayed in MEGA 3.1 (Kumar et al. 2004), FigTree 1.3.1. and CorelX3 programs.

Sequenced genomes of two species from the same genus as studied nematode model organisms, *Meloidogyne incognita* and *M. hapla*, were searched for specific sequence motifs.

4. RESULTS

4.1. Sequence analysis of satellite DNAs in *Meloidogyne chitwoodi* and *M. fallax*

4.1.1. Analysis of complex satDNA Arrays

Possibility of sample cross-contamination with other nematode DNA was excluded through PCR check of genomic DNAs with SCAR (sequence characterized amplified region) primers (listed in Table 3.6.) specific for *M. chitwoodi* and *M. fallax* species (Zijlstra 2000) (Fig. 4.1.1.).

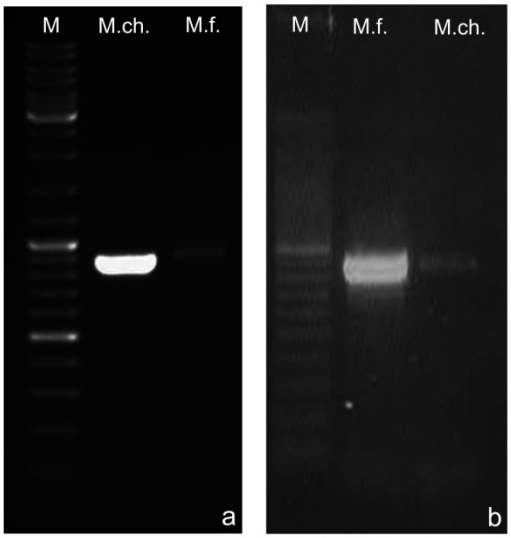


Figure 4.1.1. Electrophoretic separation of PCR products obtained by amplification of *M. chitwoodi* and *M. fallax* genomic DNAs using SCAR primers for a) *M. chitwoodi* and b) *M. fallax.*

PCR search for orthologue counterparts of satellite DNAs from *M. chitwoodi* in *M. fallax* confirmed the presence of 1a, 1b, 1c, 1d and 2a families while 2b has not been isolated (Fig. 4.1.2.). Amplification with primers specific for 1a, 2a and 2b satellite DNAs produced ladder of bands based on the monomer size. Amplification with primers specific for 1b produced bands of monomeric and dimeric size together with a fragment of about 1.5 kb in length, while amplification with 1c and 1d primers revealed complex but similar profiles (shown only for 1c).

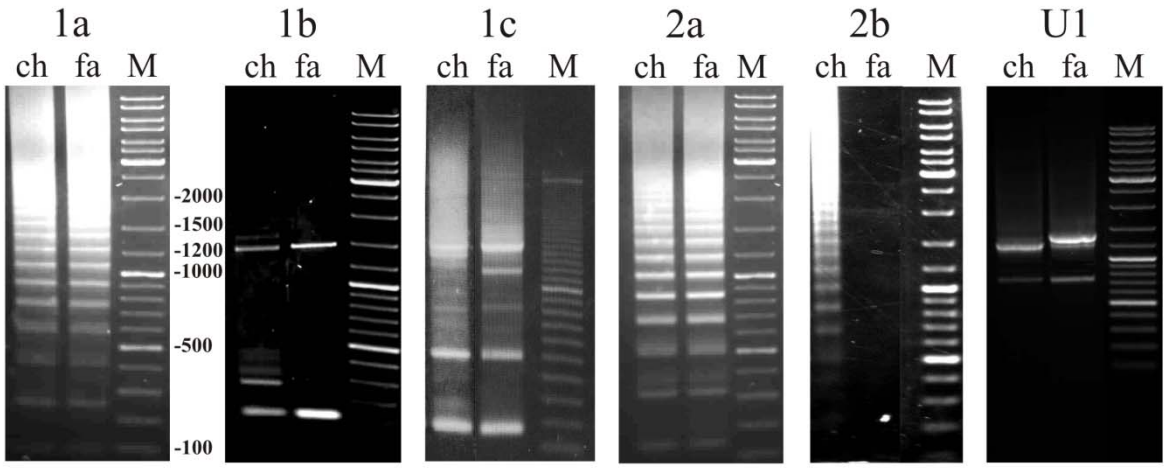


Figure 4.1.2. Electrophoretic separation of PCR products obtained by amplification of *M. chitwoodi* and *M. fallax* genomic DNAs using primers specific for 1a, 1b, 1c, 2a and 2b satellite DNAs. M is the DNA ladder marker.

For more detailed analyses of 1c satellite DNA, bands corresponding to multimeric size (i.e. ≥500 bp) obtained by amplification from both genomes were cloned and sequenced. All sequenced clones (59 altogether) are listed in Table 4.1.1.

Eight cloned fragments are composed of alternating 1c and 1d satellite DNA monomers which together define the dimeric unit, 338 bp long (169 bp x 2), organized in homogenous arrays in both genomes. Alignment of all 8 clones ($M_{1c}fa_n$ and $M_{1c}ch_n$; Table 4.1.1.) is given in Supplementary Fig. 4.1.1. Absence of a 170 bp based ladder in PCR amplification with 1c or 1d specific primers suggests that dimeric form composed of 1c and 1d monomers is the basic repeating unit of those two satellite families. Multiple sequence alignment of another 12 fragments ($H_{1c}fa_n$ and $H_{1c}ch_n$; Table 4.1.1.) revealed complex arrays composed of satellite DNA monomers 1a, 1b, a new 1b' variant, 1c, 1d and 2a together with so far uncharacterized sequence segment named U1. No relevant sequence homology of U1 with the studied satellite DNAs or any other sequence deposited in data bases was revealed with BLAST search. The alignment of complex HOR element, with each satellite family colored in different color, is given in Supplementary Fig. 4.1.2.

Table 4.1.1. Description of cloned satellite DNA arrays. In cloned satellite fragments, letters H, M and h indicate higher-order repeats, monomeric arrays, complex fragment, respectively. Then follows primer name (first subscript), species acronym and clone number (second subscript).

primers	species	Satellite fragments	Length (bp)
1c satDNA primers (1cL and 1cR)	*M. fallax*	$H_{1c}fa_2$	1353
		$H_{1c}fa_8$	1588
		$H_{1c}fa_{17}$	1530
		$H_{1c}fa_{18}$	1445
		$M_{1c}fa_8$	505
		$M_{1c}fa_{11}$	505
		$M_{1c}fa_6$	505
		$M_{1c}fa_7$	505
	M. chitwoodi	$H_{1c}ch_2$	1480
		$H_{1c}ch_3$	1480
		$H_{1c}ch_4$	1480
		$H_{1c}ch_6$	1480
		$H_{1c}ch_8$	1480
		$H_{1c}ch_9$	1523
		$H_{1c}ch_{11}$	1200
		$H_{1c}ch_{12}$	1533
		$M_{1c}ch_{13}$	505
		$M_{1c}ch_{16}$	505
		$M_{1c}ch_{10}$	505
		$M_{1c}ch_5$	505
U1 primers (U1L and U1R)	*M. fallax*	H_ufa_4	1422
		H_ufa_1	1419
		H_ufa_7	1419
		H_ufa_8	1252
		H_ufa_9	1253
		H_ufa_{10}	1419
		h_ufa_1	750
		h_ufa_2	750
		h_ufa_3	870
	M. chitwoodi	H_uch_{11}	1264
		H_uch_{21}	1269
		H_uch_{22}	1268
		H_uch_{23}	1266
		h_uch_1	750
		h_uch_2	750
		h_uch_3	870

primers	species	Satellite fragments	Length (bp)
1a satDNA primers (1aL and 1aR)	*M. fallax*	$M_{1a}fa_1$	910
		$M_{1a}fa_2$	903
		$M_{1a}fa_3$	903
		$M_{1a}fa_4$	901
		$M_{1a}fa_6$	902
		$M_{1a}fa_7$	906
	M. chitwoodi	$M_{1a}ch_3$	941
		$M_{1a}ch_4$	586
		$M_{1a}ch_5$	587
		$M_{1a}ch_6$	575
		$M_{1a}ch_8$	914
		$M_{1a}ch_9$	922
		$M_{1a}ch_{10}$	911
		$M_{1a}ch_{13}$	911
2a sat DNA primers (2aL and 2aR)	*M. fallax*	$M_{2a}fa_3$	801
		$M_{2a}fa_4$	487
		$M_{2a}fa_5$	487
		$M_{2a}fa_7$	487
	M. chitwoodi	$M_{2a}ch_1$	487
		$M_{2a}ch_2$	487
		$M_{2a}ch_3$	801
2b satDNA primers (2bL and 2bR)	*M. chitwoodi*	$M_{2b}ch_1$	500
		$M_{2b}ch_2$	500

In order to extend the segments of complex arrays, U1 specific PCR primers were constructed and used for amplification in both genomes. Obtained PCR products revealed fragments of expected lengths (~1200 and ~1400 bp) but also a shorter fragment of about 700 bp. Sequencing of longer fragments (H_uch_n and H_ufa_n; Table 4.1.1.) confirmed the same HOR organization and their alignment is shown together with sequences obtained with 1c primers in Supplementary Fig. 4.1.2. Schematic representation of HOR element, composed of U1, 1a, 1b, 1b', 1d, 1c and 2a units is given in Fig. 4.1.3.

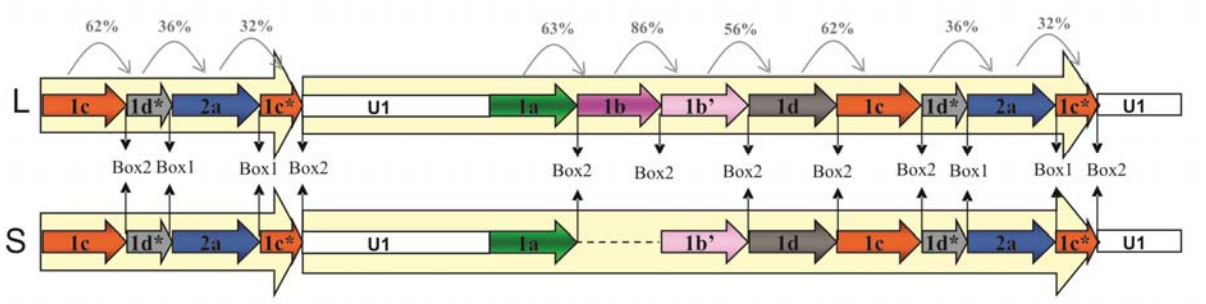


Figure 4.1.3. The long-L and short-S HOR sequence. The percent identity between monomers is written on arrows above the scheme. Box 1 and Box 2 in junction regions between different monomeres are indicated. 1d* and 1c* represent truncated 1d and 1c monomers.

Interestingly, 1d and 1c units at the end of HOR are truncated with complete monomer 2a located between them (Fig. 4.1.3.). Detailed sequence analyses showed high homogeneity of HOR units - 84 to 99% of mutual sequence identity. In contrast, neighboring monomers in HORs show a wide range of relationships: from relatively high sequence identity of 86% between 1b and 1b' variants, through moderate similarity of about 60% to apparently unrelated sequences sharing only 32% identity, such as detected between 2a and 1c monomers. The similarities between satDNAs are summarized in Table 4.1.2. and illustrated in Fig. 4.1.3.

Table 4.1.2. Mean of percent sequence identity between main groups of satDNAs monomers.

monomer length	Monomers name	Group name	n	1aHch$_n$, 1aHfa$_n$	1aMch$_n$, 1aMfa$_n$	1bHch$_n$, 1bHfa$_n$	1b'Hch$_n$ 1b'Hfa$_n$	1dHch$_n$, 1dHfa$_n$, 1dMch$_n$ 1dMfa$_n$	1cHch$_n$, 1cHfa$_n$, 1cMch$_n$, 1cMfa$_n$	2aHch$_n$, 2aHfa$_n$ 2aMch$_n$, 2aMfa$_n$	2bMch$_n$
170	1aHch$_n$, 1aHfa$_n$	1aH	26	94 (2.8)[a]							
169 (+5)	1aMch$_n$, 1aMfa$_n$	1aM	50	81 (1.6)	94 (2.1)						
170	1bHch$_n$, 1bHfa$_n$	1bH	10	63 (1.9)	64 (0.8)	99 (0.3)					
	1b'Hch$_n$ 1b'Hfa$_n$	1b'H	21	64 (1.2)	66 (0.9)	86(2.9)	93 (4.7)				
169	1dHch$_n$, 1dHfa$_n$, 1dMch$_n$, 1dMfa$_n$	1dMH	31	64 (1.9)	63 (2.0)	57 (0.5)	56 (0.7)	98 (1.7)			
169	1cHch$_n$, 1cHfa$_n$, 1cMch$_n$, 1cMfa$_n$	1cMH	31	56 (0.6)	53 (1.0)	52 (0.2)	51 (1.2)	62 (1.0)	99 (0.5)		
180	2aHch$_n$, 2aHfa$_n$ 2aMch$_n$, 2aMfa$_n$	2aMH	37	39 (0.9)	40 (1.0)	46 (0.6)	42 (1.0)	36 (0.7)	32 (0.5)	97 (1.6)	
179	2bMch$_n$	2bM	10	40 (1.1)	40 (0.8)	51 (0.5)	46 (1.6)	37 (0.6)	35 (0.6)	60 (0.9)	96 (1.3)

n-number of analyzed monomers

[a]average percent identity scores for each pairwise comparison are indicated in bold, while standard deviation (SD) is indicated in bracket

In addition, HOR segments revealed two variants which differ in the presence of 1b-type monomers. Long HOR variants have two consecutive monomers, 1b and 1b', that share sequence identity of 86%, while short HOR variants lack 1b monomer. Genomic DNA cut with the REs specific for 1c monomer sequence and probed with the labeled 1c monomer fragment supports the proposed HOR tandem organization (marked with asterisks on 1c part of Fig. 4.1.4.) Southern hybridization of genomic DNA with 1c indicates that long HOR variants prevail in *M. fallax* genome, while short variants seem to be more abundant in *M. chitwoodi*, as can be seen in Fig. 4.1.4.

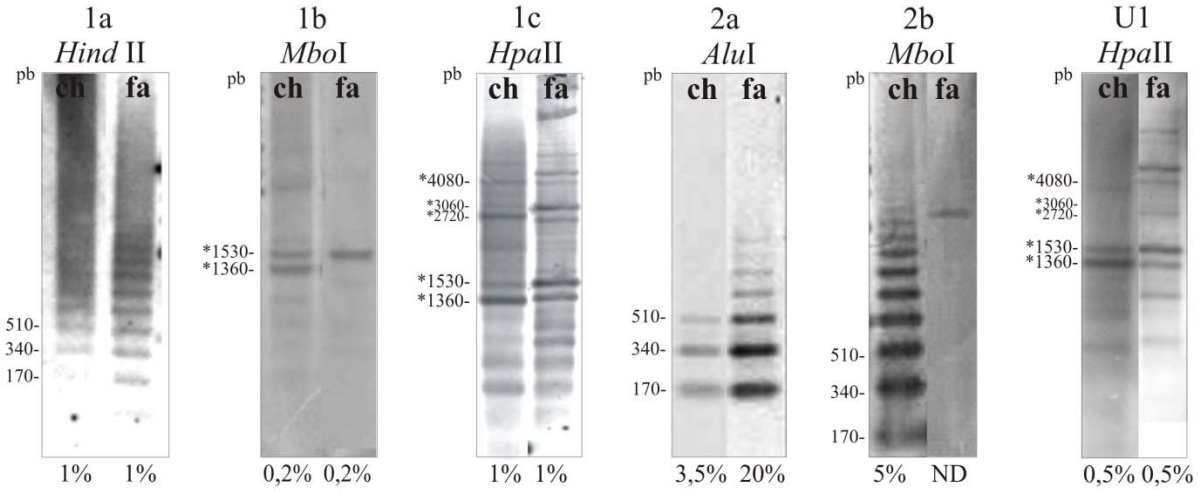


Figure 4.1.4. Southern hybridizations of *M. chitwoodi* and *M. fallax* genomic DNAs partially digested with RE-s and probed with 1a, 1b, 1c, 2a and 2b satDNA monomers and with U1 sequence. Approximate contribution of particular sequence in the genome, estimated by dot blot, is shown as a percentage indicated below Southern blots. HORs are indicated with asterisk. M is the DNA ladder marker. ND-not detectable.

Analyses of 6 cloned sequences obtained from 700 bp-long band amplified with U1 primers (h_ufa_n and h_uch_n) (Table 4.1.1.) revealed one additional complex fragment common for *M. fallax* and *M. chitwoodi*. Schematic representation of this fragment is shown in Figure 4.1.5. while alignment of those sequences is shown in Supplementary Fig. 4.1.3.

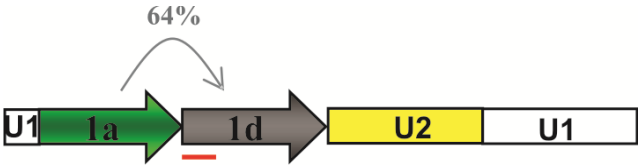


Figure 4.1.5. Schematic representation of complex fragment. The red line represents the overlapping segment of 1a and 1d monomers

These fragments are composed of complete 1a and 1d monomers linked to a novel 170 bp long fragment named U2 and flanked by U1 sequences. It has to be noted that a 62 bp-long perfectly conserved fragment of U1 is also found as a part of U2 sequence (marked with green box in Supplementary Fig. 4.1.3.) Tandem organization of the 700 bp complex fragment could not be proven by additional PCR analyses using U2 specific primers (constructed for this purpose) meaning that it is probably present in both genomes as an interspersed repeat.

4.1.2. Homogenous Monomeric Arrays

Ladder-like profile produced with 1a-specific primers gave fragments corresponding to multimers of 170 bp in both genomes. Cloning and sequencing of multimeric fragments ($M_{1a}fa_n$ and ,$M_{1a}ch_n$; Table 4.1.2.) revealed homogenous tandem arrays with 94% of mutual identity (Table 4.1.1.) composed of a variant of 1a satellite DNA sequence, named 1aM. Alignment of 14 1aM sequences, cut to monomers, is given in Supplementary Fig. 4.1.4. This 1a variant, 1aM, is different from the HOR variant therefore named 1aH. Average sequence identity between 1aH and 1aM variants is 81% (Table 4.1.1.) and their alignment is shown in Supplementary Fig. 4.1.5.

Southern blot with probe for 1aM-type satellite DNA confirmed tandem organization of 1aM variants (Fig. 4.1.4.). In addition, 1aH-specific primers were constructed to check if 1aH builds independent tandem arrays. PCR reaction did not reveal any ladder-like profile indicating that these variants are exclusively present as subunits of HORs.

PCR amplification with 1b primers revealed fragments whose length (~1400 bp) corresponds to HOR organization. According to primer position, fragments of monomeric and dimeric forms that appeared in the PCR reaction also originate from HORs. In support, Southern blot analysis of genomic DNAs with 1b showed hybridization signals only in bands corresponding to HOR arrays (Fig. 4.1.4.) emphasizing unique organization of 1b monomers exclusively in HORs in both genomes. PCR reaction with primers for 2a satellite confirmed its tandem organization as homogenous monomeric arrays in both genomes, as published previously (Philippe Castagnone-Sereno et al. 1998). This research revealed a new type of organization of this satellite - its presence in above analyzed HOR element in both genomes and difference in abundance with 3.5% of 2a in *M. chitwoodi* and 20% in *M. fallax* (Fig. 4.1.4.).

Examination of 2b satellite by PCR amplification and Southern blot recovered its exclusive presence in the *M. chitwoodi* genome in the form of high copy homogenous monomeric arrays (Fig. 4.1.2.). The only observed hybridization signal in *M. fallax* is the faint band which could represent a sporadic 2b sequence embedded in a longer DNA segment (Fig. 4.1.4.)

The abundance of all satDNA was estimated by quantitative dot blot analysis using a series of genomic DNA dilutions ranging from 50 to 200 ng. Satellite monomers, excised from a plasmid, were dot-blotted in the range between 0.05 and 1 ng, and used as a calibration curve. Fig. 4.1.6. shows dot blot result for 1c satDNA in *M. fallax* genomic DNA. Estimated abundance for all satDNAs is shown in Fig. 4.1.4. under Southern blot figures.

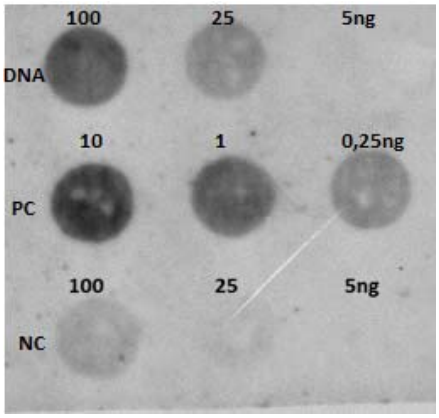


Figure 4.1.6. Dot blot for 1c satDNA in *M. fallax* genomic DNA.

4.1.3. Phylogenetic Analyzes of Monomers

All together 212 monomeric units from *M. chitwoodi* and *M. fallax* were used to examine phylogenetic relationships of all monomers, regardless to their organizational pattern and species of origin, in order to assess sequence dynamics of repetitive units in the closely related genomes. Based on multiple sequence alignment, presented in Supplementary Fig. 4.1.6., neighbor-joining phylogenetic analysis was performed which showed eight different clusters (1aH, 1aM, 1bH, 1b'H, 1cDH, 1dDH, 2aMH and 2bM; letters H, D, M indicate HOR, dimeric or monomeric organizational form, respectively) distributed in two main branches, satellite families of group 1 and group 2 (Fig 4.1.7.). Monomers within clusters could not be distinguished according to the species of origin nor was it possible to differentiate 1c, 1d and 2a monomers according to their array affiliation. As already observed, 1a satellite family splits in 1aM and 1aH according to their organizational origin while 1aH further clusters in two subgroups, based on short and long HOR forms. 1b monomers form two distinct groups, 1bH and 1b'H, related to their position in HORs.

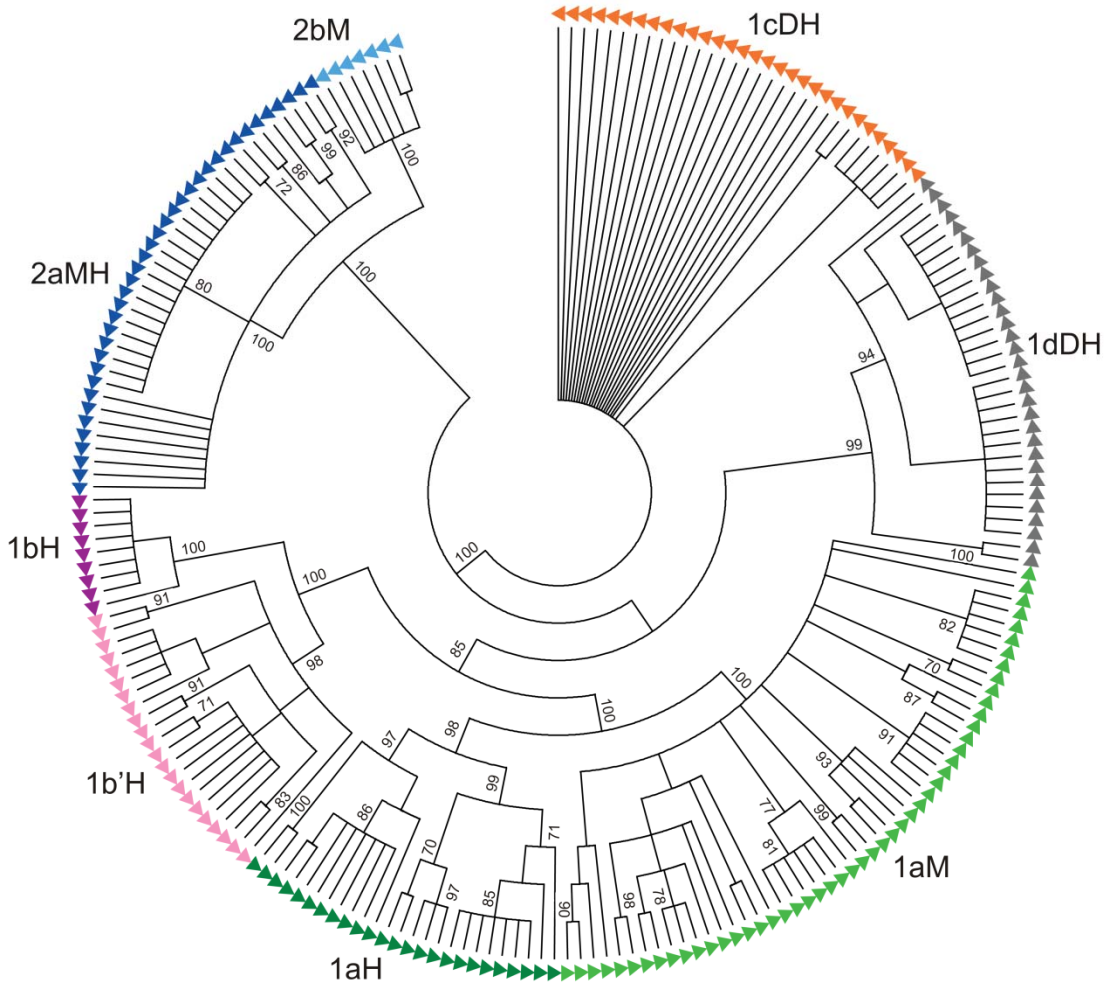


Figure 4.1.7. The phylogenetic tree of 1a, 1b, 1b' 1c, 1d, 2a and 2b monomers. Monomers from the HORs (H), dimeric (D) and monomeric arrays (M). Phylogenetic analysis of 212 monomers was performed by neighbor-joining method with bootstrap value of 100. Numbers at nodes indicate bootstrap values (100 replicates; only values greater than 70% are shown.

Sequence comparisons between monomer groups display three different levels of similarity (Table 4.1.2.). Similarity is high within 1bH group (86%) and between 1aM and 1aH (81%) monomer variants. Similarities within other satDNAs of group 1 and within satDNAs of group 2 are moderate, ranging from 51 to 66%. Sequence comparison and phylogenetic analyses between monomers of group 1 and 2 gives negligible similarities, 32–46% (Table 4.1.2.), and it can be supposed that these two groups might represent sequences of unrelated origin.

4.1.4. Conserved Motifs and Junctions Between Monomers

In contrast to the very low overall sequence similarity between some of the monomer groups, pairwise sequence alignment of consensus sequences of all 6 satellite families showed in Figure 4.1.8.A, and sliding window analysis (Fig 4.1.7.B) of all monomer sequences identified common domains of low variability. The grey shaded domain in Figure 4.1.8.A indicates the region of low variability shared among all satellite DNAs. Part of this region, 17 bp long segment, is a conserved block named Box 1 interesting because it remains conserved among highly divergent satellite DNAs like 1c and 2a that share only 32% identity while in the same time only one change characterizes the Box 1. Interestingly, transitions from truncated 1d to 2a monomer and from 2a to truncated 1c are located exactly at the Box 1. Significant degree of similarity is noted when conserved Box 1 sequences of all 6 satellite DNAs families are compared with the human CENP-B box. The comparison is shown in Figure 4.1.8.C with 6 satellite DNAs presented as a reverse complement. Six of them have 10–12 out of 17 nucleotides conserved and if bases essential for CENP-B binding in human are considered, 4–5 out of 9 remain conserved. The lowest identity is in exclusively HOR-included elements, 1b'H and 1bH, in which sequences may represent degenerate variants of the motif. This analysis was extended with the search for related motifs in two species from the same genus whose genomes have been sequenced, *M. incognita* and *M. hapla*. Preliminary results recovered no similarity in *M. hapla* genome but found different repetitive sequences with the Box 1 in unassembled part of *M. incognita* genome (Supplementary Fig. 4.1.7.). However, none of these repeats indicated any sequence similarity with satellite DNA sequences from *M. chitwoodi* and *M. fallax*.

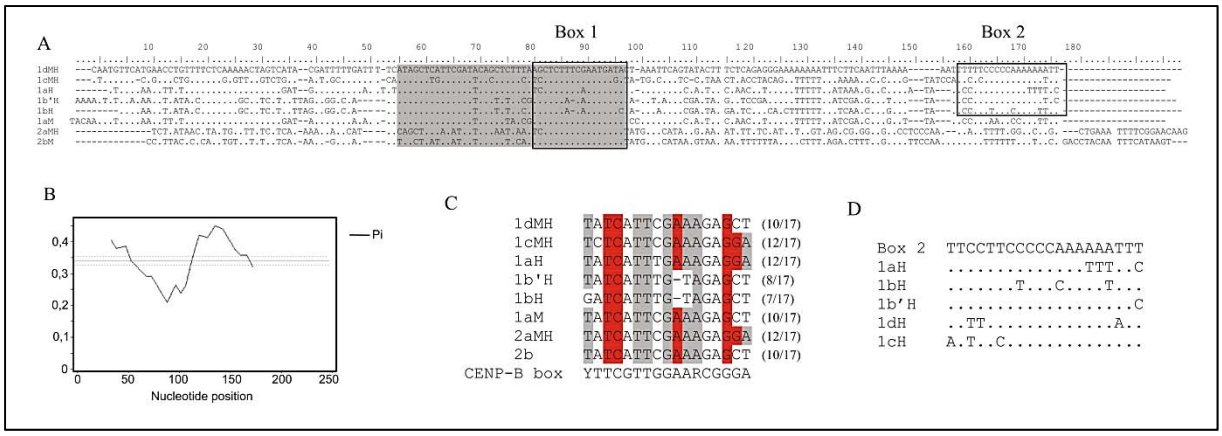


Figure 4.1.8. a) Consensus sequences of 1dMH, 1cMH, 1aH, 1bH, 1b'H, 1aM, 2aMH and 2bM satDNAs, determined according to the 50% majority rule. Conserved Box 1 and Box 2 are indicated within the boxed area, and shaded part represents a region of low variability. **b)** Identification of low variable domains by sliding window analysis by DnaSP. The average nucleotide variability P is shown by a solid line, and dashed lines represent 2-fold value of standard deviation. **c)** Comparison of two variants of Box 1 with the consensus of human CENP-B box. The reverse complementary sequence of Box 1 is presented. Identities between sequences are highlighted in grey, and bases considered essential to bind the CENP-B protein in human (Csink and Henikoff 1998) are highlighted in red. The number of total conserved bases is reported in brackets. **d)** Aligment of Box 2 sequences from HOR related monomers; positions identical to the overall consensus are shown with dots.

In HOR-related monomers of group 1 satellite DNAs (1aH, 1bH, 1b'H, 1cH and 1dH) there is another common region named Box 2, conserved between the group members. Its position is indicated by black box in Figure 4.1.8.A and in Supplementary Fig. 4.1.2. Alignment of consensus sequences of Box2, part of all satellites from group 1, is given in Figure 4.1.8.D. This region is 20 bp-long composed of T, C and A tracts and shows significant degree of mutual sequence identity with only few nucleotide changes. It is interesting that the Box 2 region is always found in HORs as a transition region between monomers from group 1. In addition, detailed analysis of the so-called complex fragment (Figure 4.1.5.) revealed that 1a monomer extends into 1d monomer in the 50 bp long overlapping region shared by both monomers. This whole segment is highly conserved, with only 6 nucleotide substitutions.

4.2. New satellite DNAs in the genome of coleopteran *Tribolium castaneum*

4.2.1. Identification of new satDNAs in genome of *T. castaneum*

First step in finding new satellite DNAs in sequenced genome of beetle *Tribolium castaneum* (Richards et al. 2008) was to analyze the genome with tandem repeats finder (TRF) algorithm (Benson 1999) that is implemented in Tandem Repeats Database (TRDB) (Gelfand et al. 2006). The genome is composed of 10 assembled chromosomes and 2153 unassembled reads (The third version of the assembly: Tcas_3.0). Their sizes are from 10 Mb of the smallest chromosome 1 (Ch1) to the 38 Mb big chromosome 3 (Ch3) (Table 4.2.1.). Each chromosome was analyzed separately while chromosome 3 had to be divided in two parts because of its large size. Alignment parameters were 2,7,7 (match, mismatch, indels) and minimum alignment score to report was 50. The result of this search was a list of arrays that were subsequently filtered out by monomer length between 100 and 500bp. This monomer size was chosen because that size is generally the most widespread and most of so far known satellites fall into that range. After that all redundant arrays (arrays with overlapping positions on chromosome) had to be removed from further analyses. They appear because program, when it is possible, offers more than one array type for the same sequence. For example, the same array can be recognized as 10 copies of 180 bp long monomers or as 5 copies of 360 bp long monomers. In those cases arrays with shorter monomer size were selected for further anylyses. 2960 arrays of tandem repeats with a total length of 3.25 Mb were obtained. They constitute 2.1% of the 156 Mb long *T. castaneum* assembled genome.

To explore trends of monomer length in those arrays and a possible correlation with copy number of monomers in arrays, they were divided into three classes: arrays with 2 monomers (634 arrays), 3-4 monomers (1563 arrays) and ≥5 monomers (763 arrays). Each class was analyzed separately (Figure 4.2.1.). Arrays with only 2 repeat units are predominantly built of monomers with length between 100 and 180 bp, while number of arrays drops with increased monomer size. Arrays with 3-4 monomers are most abundant when monomer size is in a narrow range between 160 and 180 bp. This analysis also indicates an increased number of arrays in the interval between 320 and 340 bp. Further increase of number of monomers in arrays (≥5 monomers) shows even more prominent domination of arrays with 160-180 bp long monomers and additional enrichment in the 320-

340 bp range. Furthermore, dramatic decrease in the number of long arrays is evident when monomer length increases above 340 bp.

Table 4.2.1. Size of each of the 10 chromosomes and unassembled reads (total length, length with captured and without captured and uncaptured gaps) and number of tandem arrays on each chromosome detected by TRF before and after filtering.

chromosome name	length with captured and uncaptrued gaps in bp	length with captured gaps in bp	length without captured and uncaptured gaps in bp	number of arrays before filtering	number of arrays after filtering (>5 copies, 100-500 bp)
Ch1	10877635	7277635	7017036	728	23
Ch2	20218415	14518415	14025453	1574	18
Ch3	38791480	28591480	27070658	5150	166
Ch4	13894384	12094384	11543342	1535	47
Ch5	19135781	14335781	13841583	1684	30
Ch6	13176827	8976827	8259034	2489	126
Ch7	20532854	15432854	14850616	1697	37
Ch8	18021898	13521898	12793837	2576	117
Ch9	21459655	15459655	14607456	2287	69
Ch10	11386040	7486040	7061652	2254	130
unassembled	41251169	22771169	20543936		
Σ	~228 Mb	~160 Mb	~151,6 Mb	21974	763

For all 10 chromosomes comparison of positions of short (<5 monomers/array) and long arrays (>5 monomers/array) in respect to regions of putative eu and heterochromatin, determined by Wang et al. (2008) according to the abundance of HighA repetitive class and TEs, was done and is shown in Figure 4.2.2. Chromosomes with highest proportion of satellite DNAs are chromosomes 3, 6, 8, 9 and 10 which is in accordance with previous study where the same chromosomes show accumulation of HighA class repetitive families obtained by ScoutRepeat approach (Wang et al. 2008). In our research distribution of short (<5 monomers) satDNA arrays (Figure 4.2.2.) showed almost uniform distribution along the whole chromosomes, including the HighA domain (putative heterochromatin). Interestingly, long arrays (≥5 monomers) showed higher tendency to reside in euchromatic regions, being less represented in HighA domains. Although observed trend of long array distribution could be due to the gaps in assembly of repetitive sequences, marked uncaptured gaps do not indicate any increased frequency in HighA domains than in chromosomal segments defined as euchromatic.

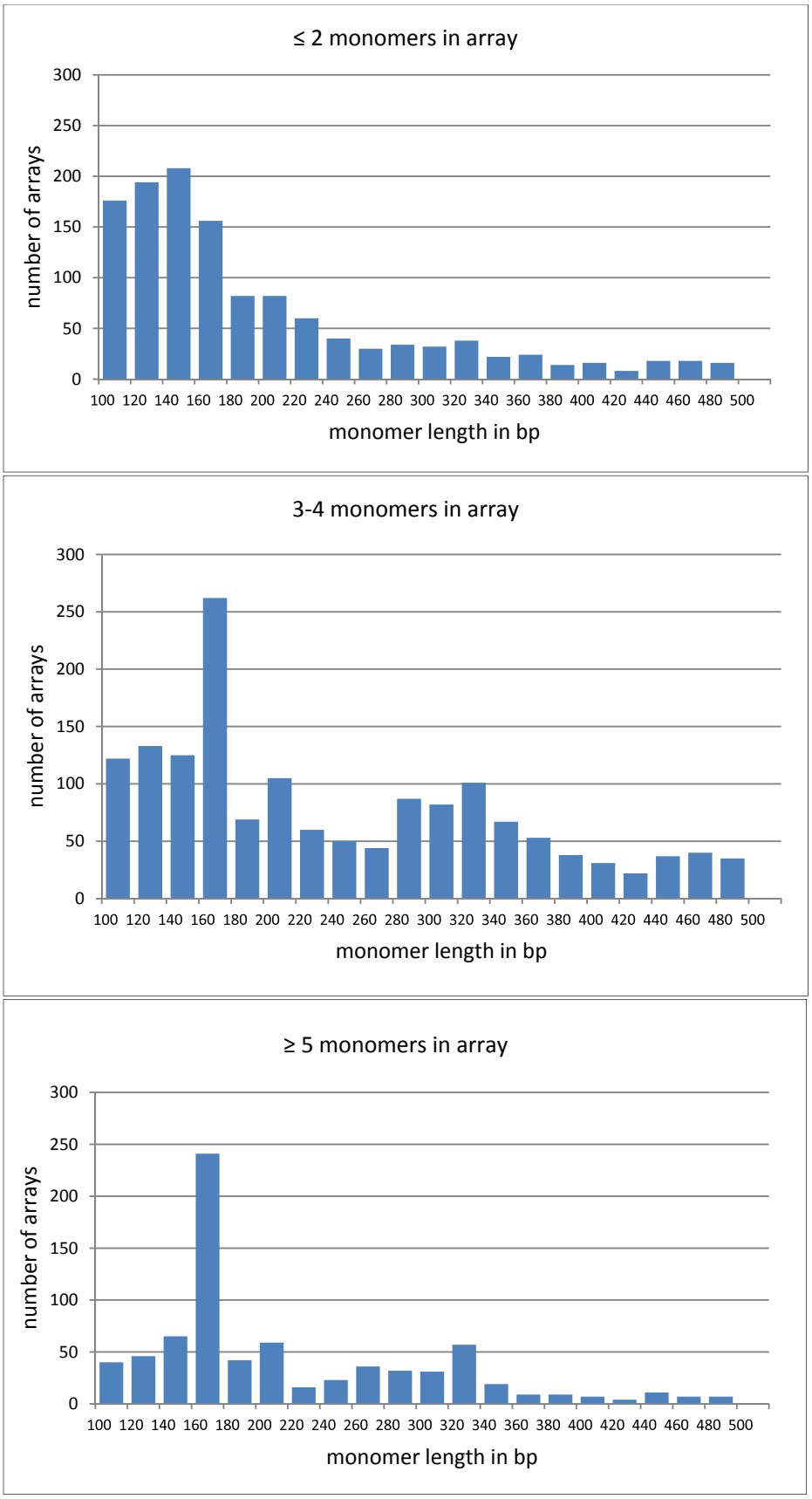


Figure 4.2.1. Correlation of monomer number in extracted TRF arrays and monomer length. Number of arrays is plotted as a function of monomer length for arrays with 2 monomers, 3-4 monomers and ≥5 monomers.

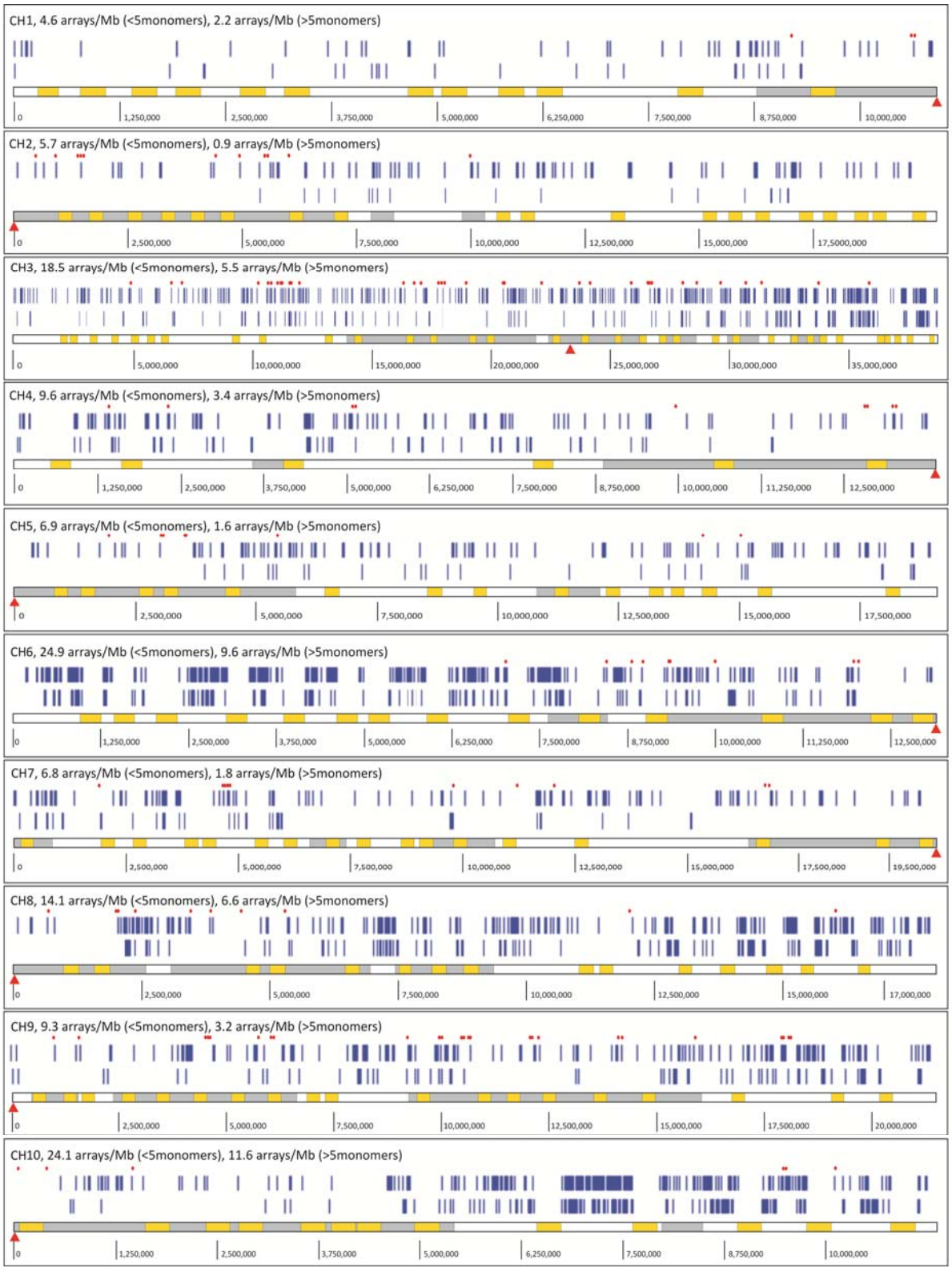


Figure 4.2.2. Genomic distribution of arrays with studied tandem repeats are superimposed on *T. castaneum* assembled chromosomes (CH1 to CH10) drawn according to Wang et al. (2008). Blue vertical bars represent short (<5 monomers/array, upper line) and long arrays (>5 monomers/array, lower line). The actual number of arrays per chromosome is indicated above each chromosme. Red dots correspond to centromeric TCAST satDNA found in the assembled genome. Red triangles indicate assumed position of the centromere and large blocks of centromeric heterochromatin. Horizontal bar represents putative

euchromatin (white) and heterochromatin (HighA domain, grey) regions as identified in Wang et al. (2008). Locations of the 300 kb placeholders were included to define uncaptured gaps (yellow bars).

4.2.2. The largest tandemly repeated DNA families

In order to explore the most abundant satDNAs in the assembled *T. castaneum* genome analyses were focused on arrays with ≥5 monomers obtained in the TRF output. This cut off level was selected for two reasons, to avoid a noise in phylogenetic analysis that would be caused by a large number of shorter arrays, and because analysis of preferential monomer length revealed positive correlation between number of repeats in arrays and monomer size typical for satDNAs (Figure 4.2.1.). Total length of all 763 arrays with ≥5 monomers is 1.63 Mb that constitute 1.04% of 156 Mb (160 Mb with captured gaps) large assembled *T. castaneum* genome (Kim et al. 2010) which is less than 2.5 % obtained by Wang et al. but they used more relaxed TRF parametrs, as mentioned in section 2.5.1. Arrays were further clustered based on profile similarity using tool called *Clustering* which is also integrated in TRDB. Conditions were as follows: P-value excluded (set to 0), cutoff value set at 70%, heuristical and DUST algorithm excluded, PAM algorithm included with default values (0.7 and 0.3). The result was 56 clusters with altogether 371 clustered arrays. The biggest cluster had 49 while last 40 clusters had only 2 arrays. Only three arrays represented (peri)centromeric satDNA and we extracted TCAST monomers from them. Using BLAST search TCAST monomers from unassembled reads were also extracted. Alignment of all obtained monomers recovered five subfamilies of TCAST pericentromeric satDNA (two were previously described (Ugarković et al. 1996; Feliciello et al. 2011), while others are newly determined) which have mutual sequence similarity of about 70% and are characterized by monomer length variation (Fig. 4.2.3.). In contrast to conservation of monomer length within satDNA family common for the most satDNAs, these monomers show large deletions (20-50 bp) and organization in form of interspersed monomers from different families (Feliciello et al. 2011). Taking into account that TRF analysis does not tolerate repeat variants with large deletions in array it was to be expected that pericentromeric satDNA was not considerably included in our output.

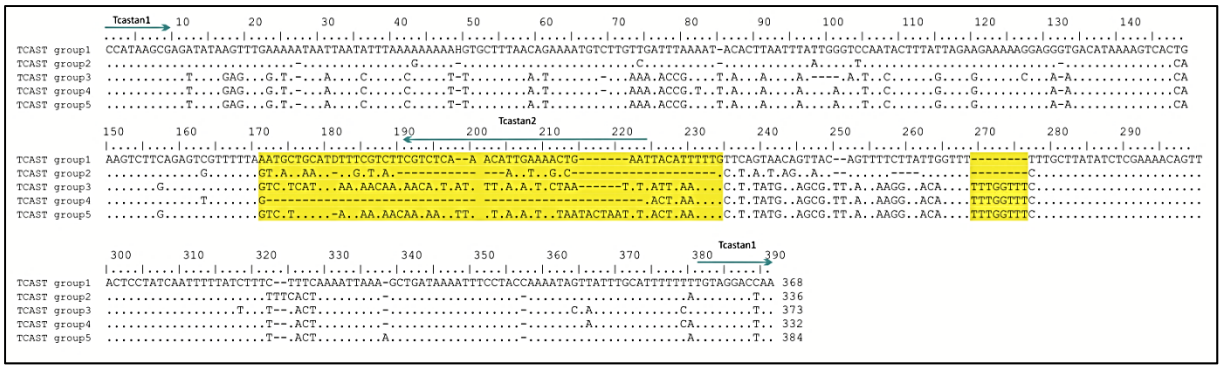


Figure 4.2.3. Alignemnt of 5 TCAST subfamilies. Positions identical to the first sequence are shown with dot and deletions are indicated with dash. Positions of primers Tcastan1 and Tcastan2 are marked with arrows above alignment.

In order to define distribution of this satDNA in the assembled genome, BLAST search of extracted arrays using consensus sequences of all subfamilies as queries was performed (red spots in Figure 4.2.2.). Only 130 short arrays mainly with 1 to 2 copies, distributed randomly along chromosomes and without any detectable preference towards the HighA domain, were determined.

Based on number (>=10) and distribution of arrays (at least two arrays on one chromosome and presence of cluster on at least two chromosomes) first 10 clusters were chosen for further analyses. Criteria are defined to enable comparative studies of monomers from different arrays on particular chromosome as well as comparisons of arrays among chromosomes. Alignments of all clusters were checked and cluster 6 was excluded from further analyses because it contains total divergent sequences. . After manual checking of all arrays from one cluster some have been removed from further analyses. The reasons were too long array length, for example arrays with 368, 240, 182, 170, 115 and 102 monomers that weren't confirmed in following FISH analyses, and arrays with both flanking regions made of unspecified (N) nucleotides (two arrays from cluster 5). These arrays are probably artificial and are a result of incorrect assembly process.

Structural characteristics of remaining 9 clusters are summarized in Table 4.2.2. They make up a little less than 1/3 of tandem repetitive sequences of assembled genome obtained by TRF. 7 out of 9 families have a monomer which can be grouped in the size-range of about 170bp and 300bp. Nucleotide sequences of all satellite families show high AT content (≥60%) and nucleotide diversity of monomers within family are in range from 10 to 28 %. Their abundance in assembled genome is in range from 0,006% up to 0,075% (9-

117kb). We also found that periodicity of AT tracts is prominent feature of all analyzed satDNAs. Number of monomers in obtained arrays was up to 54 copies which speaks in favor of the fact that noncentromerne regions of chromosomes are not resistant to the accumulation of long satellite DNA arrays. Similarity BLAST search against available data bases - NCBI GenBank Database and Repbase Update (Jurka et al. 2005) resulted in no significant similarity with any of so far known sequences suggesting that extracted tandem repeats are new, *T castaneum*-specific sequences. Local BLAST search of newly satDNAs with High, Mid and Low repetitive classes obtained by RepeatScout (Wang et al. 2008) recovered significant homology with Cl4 and Cl5 (Supplementary Figure 4.2.1.). Detailed analyses show high homology of 7 repetitive elements with monomers of Cl4 satDNA. These repetitive elements represent multimers with different monomer variants of the heterogenous Cl4 family (21% divergence between monomers; Table 4.2.2.) as well as part of monomer with different flanking regions. Three RepeatScout defined repetitive elements show homology with Cl5. These repeats are composed of Cl5 monomer and flanking regions. Monomers from other clusters have only homology with short AT rich tracts.

Table 4.2.2. Structural characteristics of 9 clusters obtained by TRF.

		Number of arrays per chromosome																
Cluster name	Number of arrays	Ch 1	Ch 2	Ch 3	Ch 4	Ch 5	Ch 6	Ch 7	Ch 8	Ch 9	Ch10	Max. number of monom-ers per array	Nucleotide diversity (Pi) of monomers in cluster±standard deviation	The average length of monomers (bp)	AT content	Number of monomers	Total repeat family lenght (kb)	Proportion of the genome (%) assembled reads	% of genome estimated by dot blot
Cl 1	46	0	1	13	1	1	8	3	4	10	5	31	0,20143±0,00320	166-173	66,2	489	83	0,053	>1
Cl 2	42	9	1	4	5	4	4	1	10	1	3	39	0,12138±0,00303	166-172	72,3	512	87	0,056	0,5
Cl 3	35	2	1	3	6	3	7	2	5	2	4	17	0,11734±0,006	205-219	74,8	230	48	0,031	0,2
Cl 4	33	0	1	5	0	2	3	8	10	4	0	50	0,21765±0,00296	168-176	69,7	426	73	0,047	0,5
Cl 5	30	0	1	6	3	3	5	1	7	1	3	28	0,09620±0,00182	270-338	73,1	384	117	0,075	>1
Cl 7	7	0	0	5	0	1	0	0	1	0	0	54	0,16874±0,00385	179-181	66	157	28	0,018	0,5
Cl 8	10	0	0	4	1	2	1	0	2	0	0	12	0,28310±0,00129	109-114	68,9	83	9	0,006	0,2
Cl 9	10	0	0	2	0	0	1	0	1	4	2	24	0,16321±0,00733	161-167	67,3	120	20	0,013	0,2
Cl 10	10	0	0	2	0	1	3	0	1	0	3	10	0,25299±0,00481	311-346	72,6	64	21	0,013	0,2
Σ	223	11	5	44	16	17	32	15	41	22	20						486	0.312	>4

The same set of analyses was done with more relaxed parameters. Alignment score was lowered to 2, 3, 5, monomer length range was expanded to 100-2000 bp and cutoff value for clustering was set to 60%, but there was no significant change in the output data. Clusters with more than 10 arrays were the same like in more stringent analyses; the only difference was a few extra, but more diverse, arrays per cluster.

Alignments of all monomers from each of the 9 selected clusters were downloaded in fasta format and imported in bioinformatics programs BioEdit 7.0.9.0 (Hall 1999) and Geneious 5.5.6 where they were further analyzed. All gaps from the TRDB alignment were deleted and new, ClustalW alignment (Thompson et al. 1997), was performed and consensus sequence for each of the cluster was defined (Figure 4.2.4.). Complete alignments of each of the clusters are shown in Supplementary Figure 4.2.2. Since tandemly repeated segments of genomes are still poorly assembled dot blot hybridization analyses was performed for each of extracted families in order to estimate actual genome content of 9 *in silico* found satDNAs. Furthermore, in order to determine the positions of new satDNAs related to (peri)centromeric regions two colored FISH was performed and for the purpose of validating a tandem repeats profile of new satDNAs Southern blot hybridization analyses were carried out. For the purpose of creating specific probes for each of the new satDNAs primers for amplifying each of the clusters (listed in Table 3.7.) were constructed based on consensus sequences. Products of PCR reactions with those primers are shown in Figure 4.2.5. Positions of the primers are marked in Supplementary Figure 4.2.3.

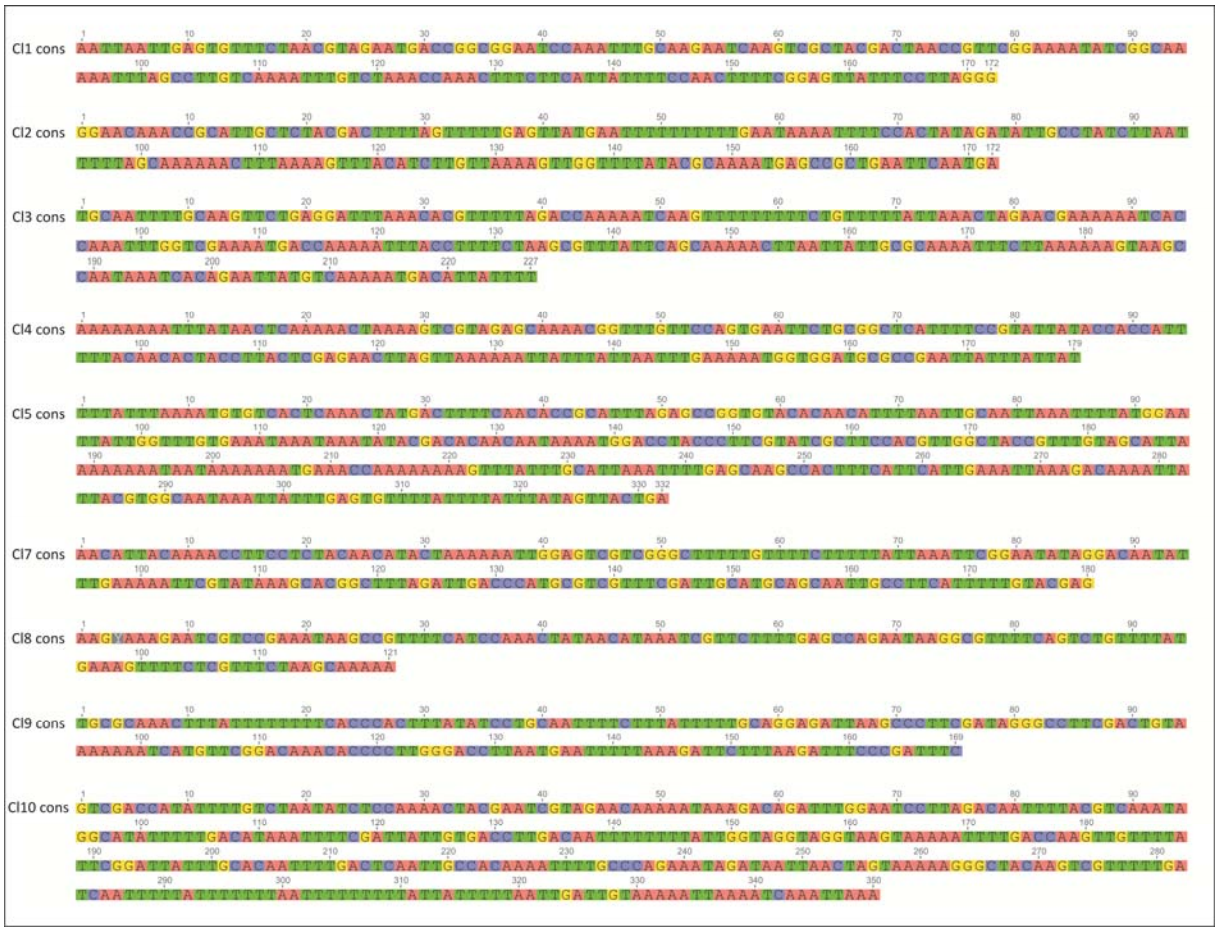


Figure 4.2.4. Consensus sequences of all nine clusters, Cl1 to Cl10.

Fragments obtained by PCR were used to transform bacterial cells. Several clones from each of the 9 transformation experiments were sequenced and one was chosen for oligonucleotide probe construction by PCR. Alignments of all cloned fragments are given in Supplementary Figure 4.2.3. in which clones chosen for probe construction are marked by black boxes.

Results obtained by dot blot (showed only for Cl5 and Cl7 satDNA in Fig. 4.2.6) revealed that Cl1 and Cl5 are the most abundant, each comprising about 1% the genome (Table 4.2.2.). Second category, with 0.5% abundance, are Cl2, Cl4 and Cl7. Other satDNAs comprise about 0.2% of the genome each. In summary, real abundance of all analyzed satDNAs is more than 4% of the genome which is about 10 times higher than in the assembled genome. Additional analyses of unassembled reads with TDRB, using the same parameters as for chromosomes, showed the highest proportion of Cl5, Cl7, Cl1 and Cl2 satellite DNAs in unassembled portion of the genome (Supplementary Figure 4.2.4.).

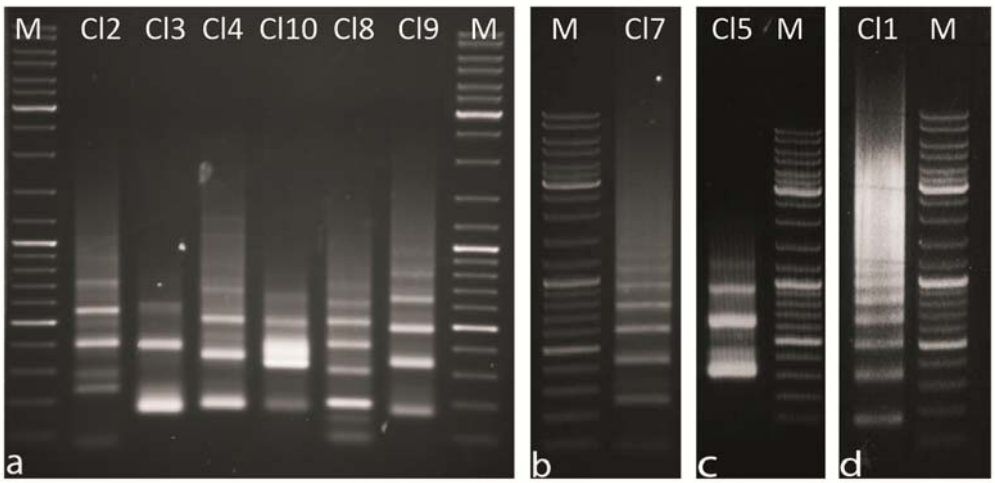


Figure 4.2.5. Electrophoretic separation of PCR products obtained by amplification of *T. castaneum* genomic DNA using primers specific for each of the new satDNAs. **a)** Primers for clusters 2, 3, 4, 10, 8 and 9; **b)** primers for cluster 7; **c)** primers for cluster 5; **d)** primers for cluster 1. M is the DNA ladder marker.

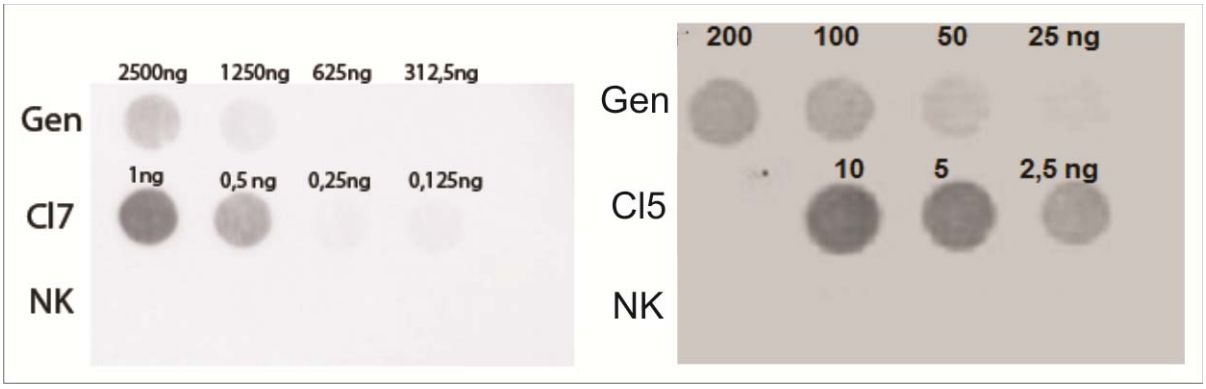


Figure. 4.2.6. Determination of the abundance of clusters 7 and 5 in total *T. castaneum* genomic DNA by dot blot hybridization.

Data obtained by dot blot are in accordance with FISH analyses, shown in Figure 4.2.7. Two colored FISH was used to determine positions of new satDNAs related to (peri)centromeric regions. TCAST satDNA has previously been characterized as the major satellite that encompasses (peri)centromeric regions of *T. castaneum* chromosomes (Ugarković et al. 1996). TCAST probe was labeled Cy3 and probes for each of the 9 clusters were biotin labeled. Because of small size of *T. castaneum* chromosomes and their condensation state which cause lower FISH sensitivity, specially in a case of low copy satDNA families, detailed mapping of newly satDNAs on chromosomes in meiotic prometaphase was not possible. For that reason chromosomes in mitotic prometaphase which enabled detection of centromere regions together with signals of newly detected satDNA were used. Signals obtained after FISH hybridization with Cl1 and Cl5 probe were significantly stronger

than for other families, as has already been shown by dot blot experiments. In general, FISH analyses showed localization of all nine satellites almost exclusively at noncentromeric chromosome regions with some overlapping signals in pericentromeric regions, especially in Cl5 FISH analysis. Also, cytogenetical analyses enabled detection of Yp chromosome, two metacentric chromosomes Ch3 and Ch2, while the remaining chromosomes are mostly telocentric.

Southern blot hybridization analyses were carried out for the most prominent satDNAs, Cl1, Cl2, Cl4, Cl5 and Cl6 (Fig. 4.2.7.), which each represent 0.5% or more of the genome, to validate a tandem repeats profile of the sequence sets generated *in silico* by TRF. Other low copied satDNAs, Cl3, Cl8, Cl9 and Cl10 were below the level of detection by Southern blot hybridization because they present 0.2% of the genome or less. Genomic DNA was digested completely using restriction enzymes which cut once in the most monomer sequences and with the once with recognition sites only in some monomers. These restriction enzymes were chosen because they produce clear n-mers in tandem organized sequences. Selection of REs for particular satDNA was based on alignment of all monomers of one cluster. Hybridization analyses for the most prominent satDNAs were performed separately with probe specific for each TRF satDNA family. In addition to a strong signal of predicted monomer size typical satellite ladder-like pattern showing individual n-mers units was observed in all hybridization analyses. No intermediate bands are observed in any of the ladders indicating a tandem arrangement of monomers in all detected families.

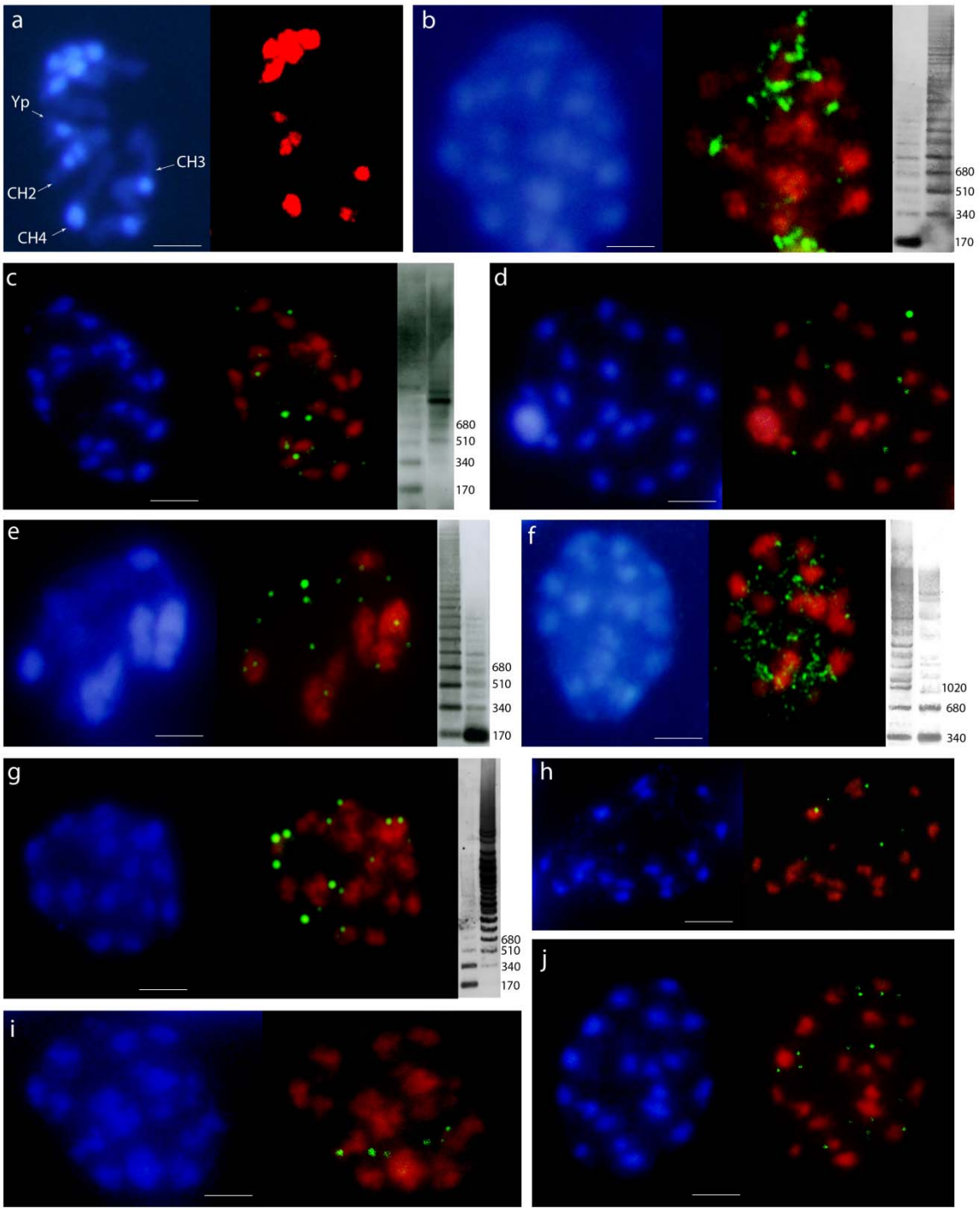


Figure 4.2.7. Fluorescence in situ hybridization of centromeric TCAST satDNA and satDNAs determined in this work by TRF analysis. Chromosomes are counter-stained with DAPI. The bar represents 1 µm. **a)** FISH showing centromeric TCAST satDNA (red signals) on *T. castaneum* chromosomes in meiotic prometaphase. Arrows point to chromosomes Ch2, Ch3, Ch4 and Yp. Two-colored FISH performed on chromosomes in mitotic prometaphase show localization of new satDNAs (green): **b)** Cluster 1, **c)** Cluster 2; **d)** Cluster 3; **e)** Cluster 4; **f)** Cluster 5; **g)** Cluster 7; **h)** Cluster 8; **i)** Cluster 9; **j)** Cluster 10. Aside to chromosome spreads Southern blot analyses of genomic DNA digested with restriction enzymes and hybridized with specific probes: Cl1 with HinfI and HaeIII (**b**), Cl2 with EcoRI and HaeIII (**c**), Cl4 with HaeIII and HinfI (**e**), Cl 5 with RsaI and DraI (**f**) and Cl7 with HaeIII and HinfI (**g**) are shown. Only satDNAs >1% of genomic DNA are presented.

4.2.2. Phylogenetic relationships among newly defined satDNAs

For revealing evolutionary trends of dominant non-centromeric satellite sequences in *T. castaneum* genome phylogenetic relationships between monomers from 9 extracted clusters were examined. Maximum likelihood (ML) trees based on Clustal W alignments (Supplementary Figure 4.2.2.) were obtained with the PhyML 3.0. software (Guindon and Gascuel 2003) using best-fit models calculated by the jModelTest 2.1.3. (Darriba et al. 2012). Truncated monomers from the beginning and from the end of the array were removed from alignments. Monomers were annotated with respect to chromosome of origin (1 to 10) and the original position of array on that chromosome. Trees are displayed and adjusted in FigTree 1.3.1. and CorelX3 softwares. Annotations and branches of all monomers from one chromosome are colored in the same color, identical for all trees. Trees for all clusters, with monomer annotations, are shown in Figures 4.2.8., 4.2.9., 4.2.10. and 4.2.11. Simplified tree forms, without monomer names and with added symbols for specific kind of distribution, defined further in the text, are shown in Figures 4.2.13., 4.2.14. and 4.2.15. for 6 selected clusters.

Colored arches and other symbols indicate: dominant chromosome-specific clusters of neighboring arrays (beige arches), chromosome specific clusters of arrays on distant position (lilac arches), clusters with arrays which come from non-homologous chromosomes (recent exchange, green arches) and dispersed monomers (diverse symbols). The tree topologies generally show strong clustering of repeats from the same array but the exception are dispersed short arrays (5-7 monomers) which is particularly evident in the Cl3 which is dominated by that kind of arrays.

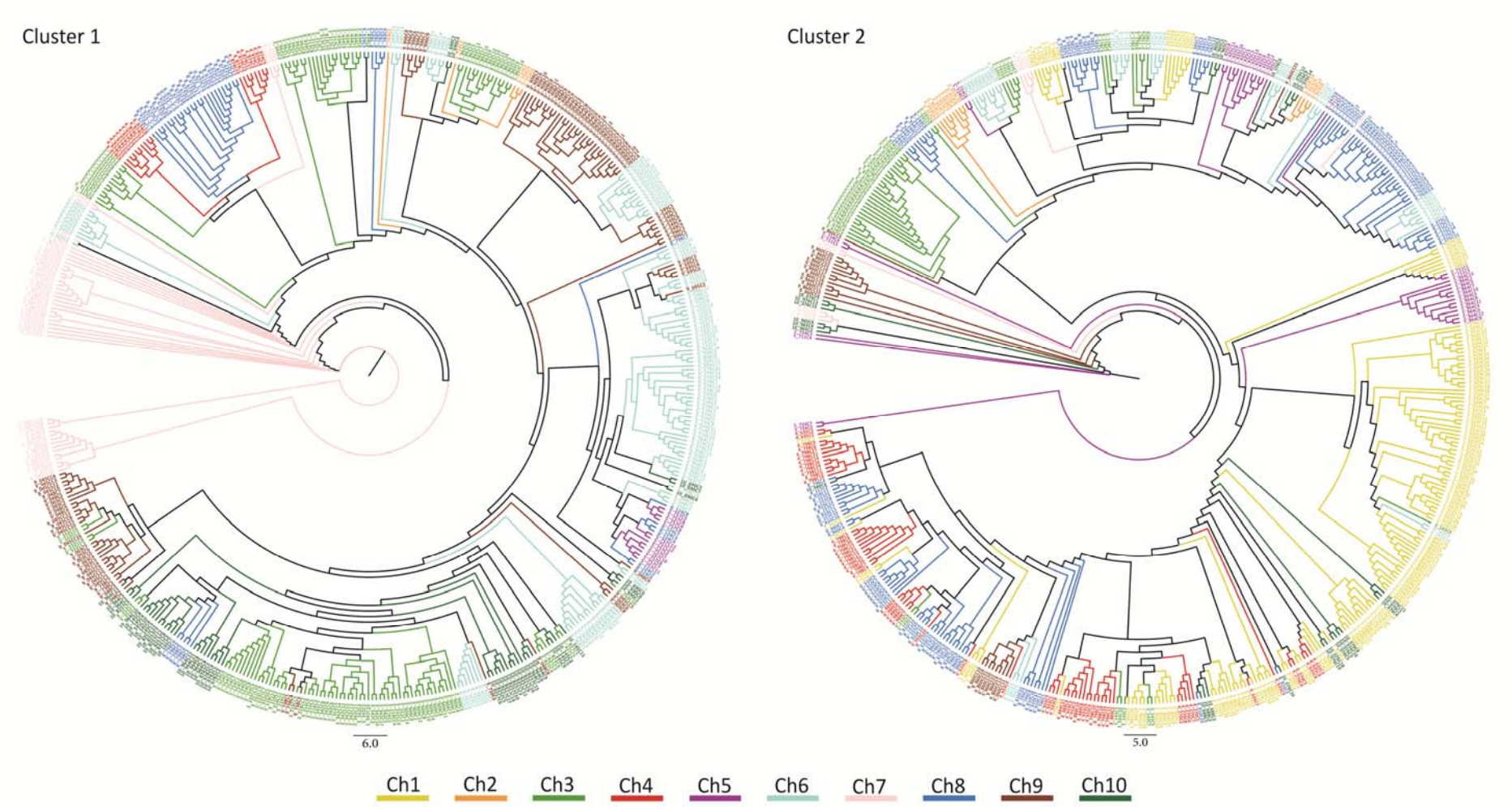


Figure 4.2.8. ML trees of Clusters 1 and 2. Arrays from one chromosome are colored in the same color.

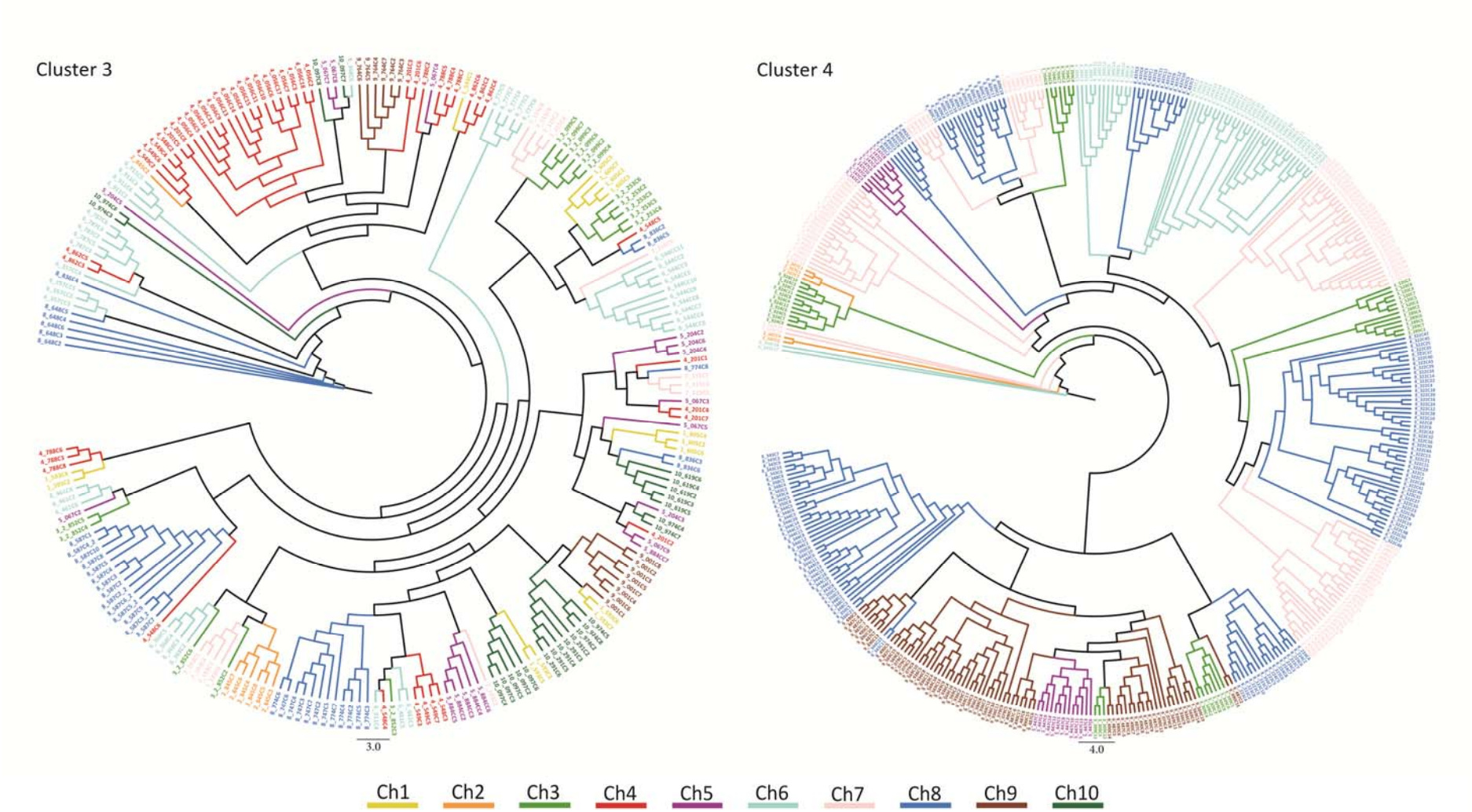


Figure 4.2.9. ML trees of Clusters 3 and 4. Arrays from one chromosome are colored in the same color.

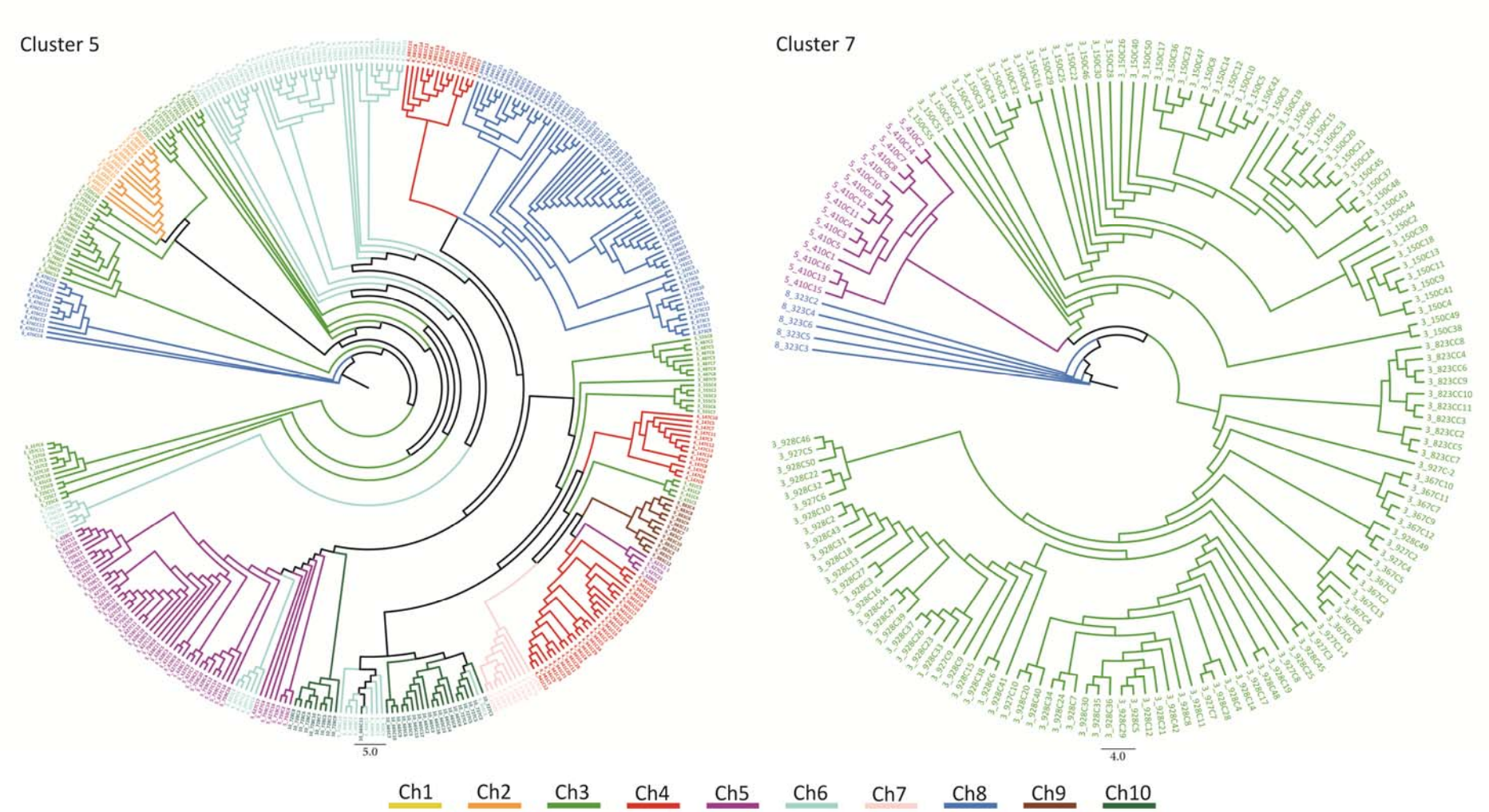


Figure 4.2.10. ML trees of Clusters 5 and 7. Arrays from one chromosome are colored in the same color.

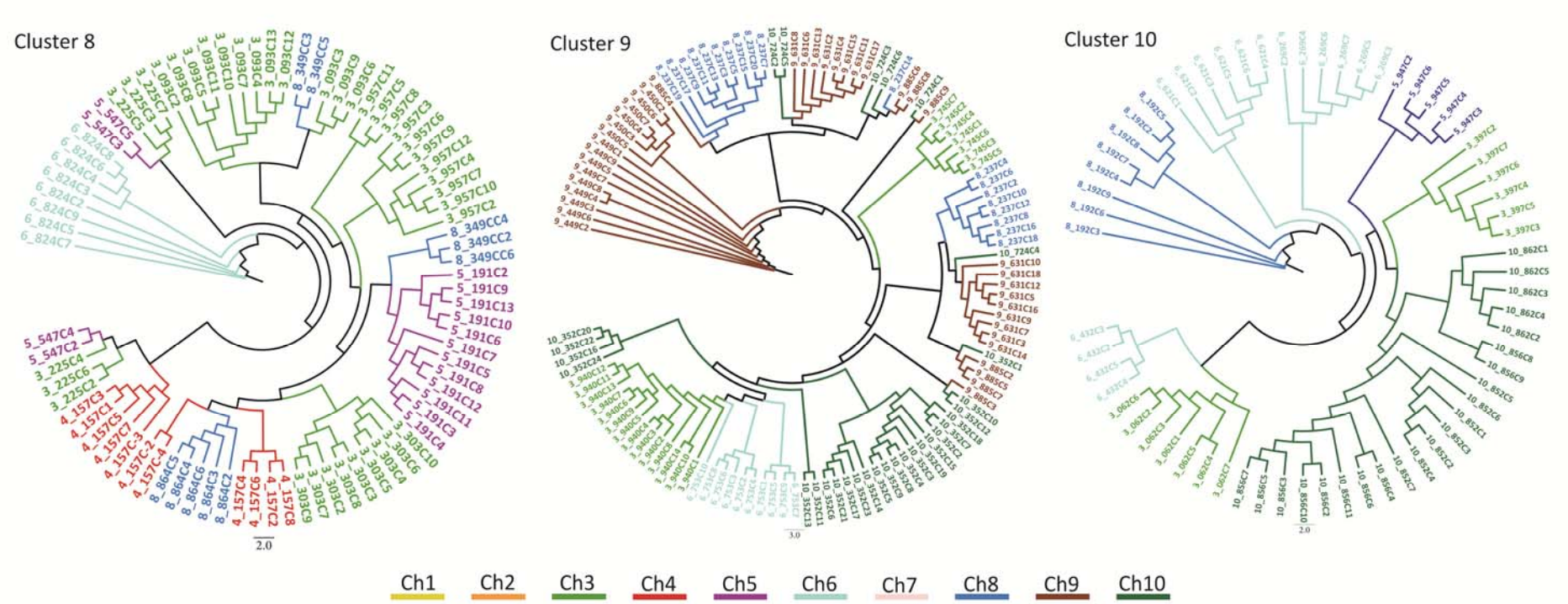


Figure 4.2.11. ML trees of Clusters 5 and 7. Arrays from one chromosome are colored in the same color.

Cluster 1

It is obvious that divergent Cl1 family (20%) is represented by both short and long arrays which is confirmed by FISH results (Fig. 4.2.7. b). However, the longest arrays are not represented in the assembled genome. The relationships are very diverse. Monomers from some long arrays tend to be clustered together while from others are very distant. There are 3 chromosome specific groups of distant arrays, 3 groups of arrays from different chromosomes while short arrays show scattered distribution. The cluster is distributed on every chromosome except Ch1 (sex chromosome) showing significant exchange between non-homologous and on chromosomes 2, 4 and 5 is present with only one, not to homogenous array.

Cluster 2

Low divergence between monomers (12%) is characteristic of this cluster. ML tree displays significant and relatively recent (short branches) dispersion of arrays between non-homologous chromosomes visible in two stages (indicated with two green arches). There is only one chromosome-specific cluster (Ch1=X chromosome) with mixed monomers (originated from tree arrays) located on distant chromosome positions. Cl2 tree shows exchange between sex-chromosome (Ch1=X) and autosomes.

Cluster 3

This family is also characterized by the low monomer divergence (11%) and short arrays, mainly comprising from 5 to 7 monomers. Cl3 tree displays clustering of monomers from long arrays while monomers derived from short arrays show scattered formation (see symbols on the tree). Cl3 tree also shows dynamic exchange of arrays between non homologous chromosomes.

Cluster 4

This family has a significant monomer divergence (21%) even within arrays, which is evident from the long branches. Tree mainly consists of long arrays whose monomers tend to be clustered together. There are two examples of chromosome specific clustering: arrays located close to each other (up to 20 kb) which are probably homogenized together (beige arches) and distant arrays that show intra chromosomal exchange (lilac arches). This family also shows significant exchange between non homologous chromosomes.

Cluster 5

Despite of the lowest overall divergence among monomers (9%) this cluster shows strong grouping of monomers from the same array (the average array length is about 15 monomers). Arrays from the same chromosome show slight tendency to group together even in the case when arrays are on distant locations (lilac arches). The tree also shows one recent interchromosomal exchange in a fraction of monomers.

Cluster 7

Tree from cluster 7 shows long arrays dominantly located on chromosome 3 and significant intrachromosomal exchange ("bar code" for Ch3). This almost exclusive localization of Cl7 family to Ch3 is confirmed by FISH experiments with Cl7 monomer as a probe on meiotic prometaphase chromosomes (Figure 4.2.12.). There are only two more arrays of this cluster (also visible as green dots on Figure 4.2.12), each one on distinct chromosome, deeply divided from one another and from arrays from Ch3, as can be seen from ML tree.

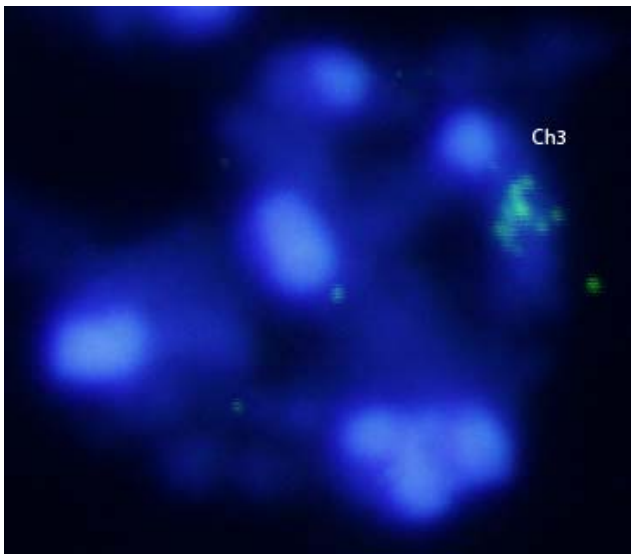


Figure 4.2.12. Overlapping of DAPI stained nucleus and hybridization with biotin labeled probe (green signal) for Cl7 monomer. Accumulation of Cl7 arrays is clearly visible on chromosome 3.

Trees for Cl8, Cl9 and Cl10 trees are shown only on Figure 4.2.11. because they follow similar evolutionary trends as described for other clusters and additional new features were not noted.

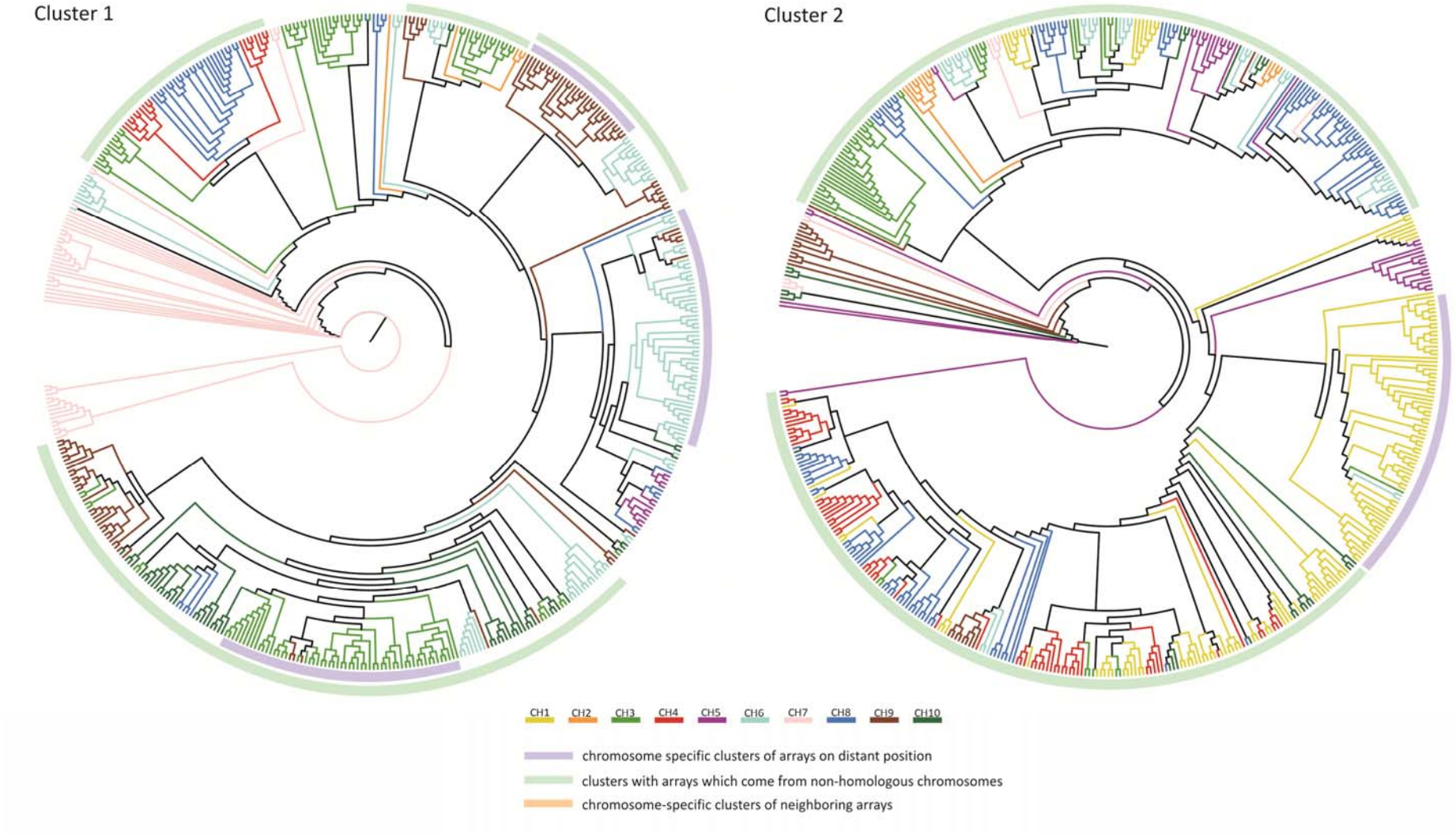


Figure 4.2.13. ML trees of clusters 1 and 2 with marked symbols for different types of distribution of satellite monomers. Arrays from one chromosome are colored in same color.

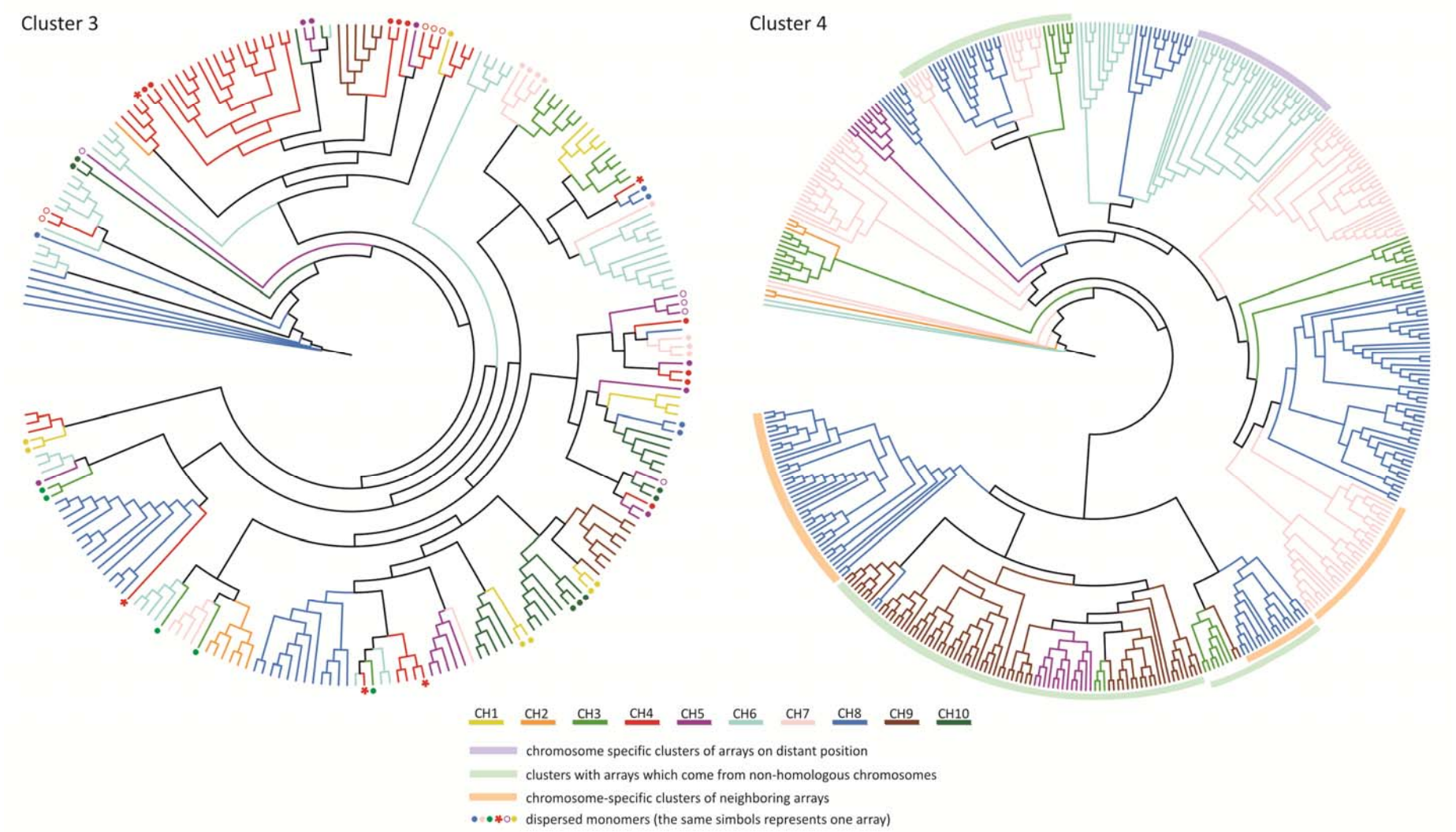


Figure 4.2.14. ML trees of clusters 3 and 4 with marked symbols for different types of distribution of satellite monomers. Arrays from one chromosome are colored in same color.

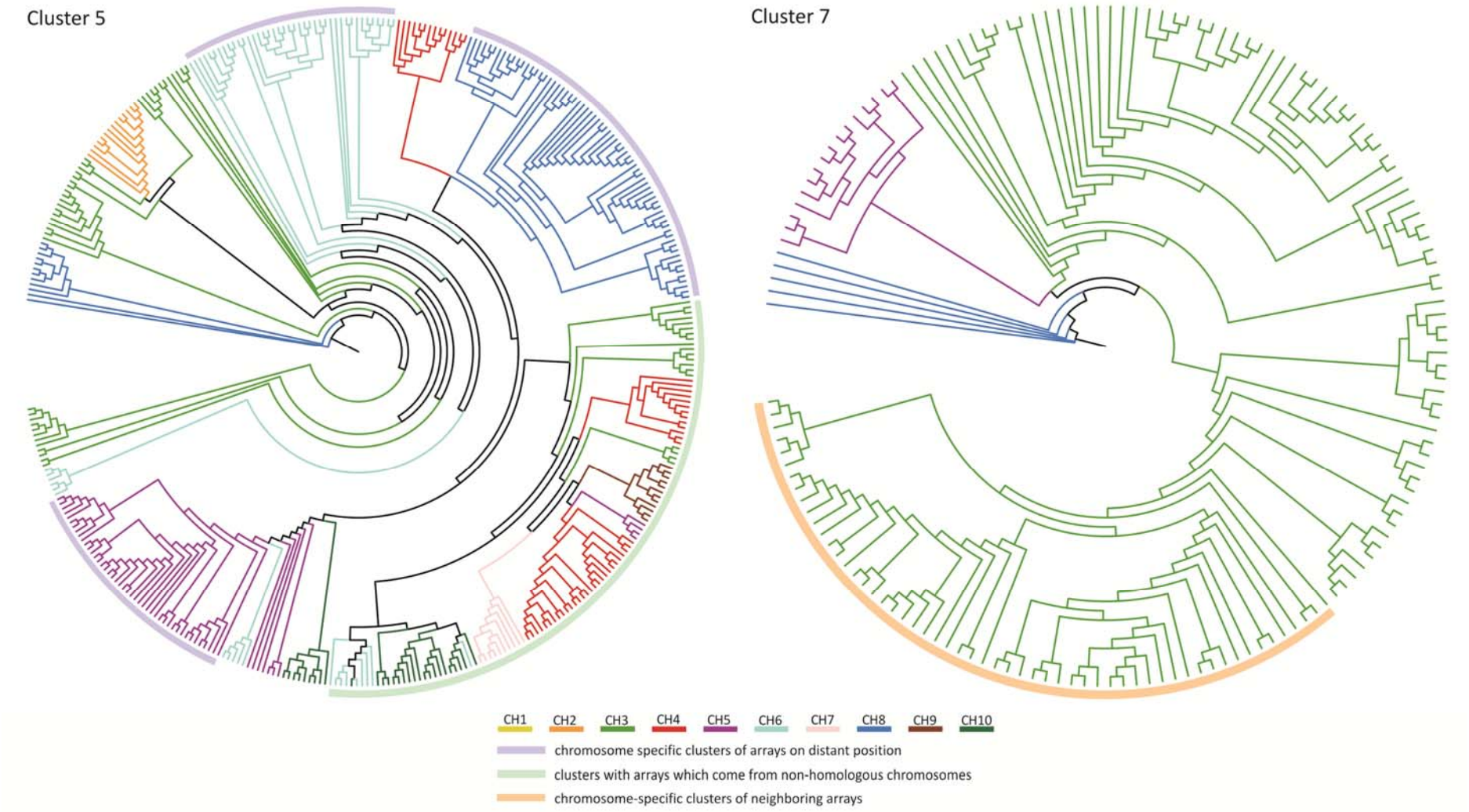


Figure 4.2.15. ML trees of clusters 5 and 7 with marked symbols for different types of distribution of satellite monomers. Arrays from one chromosome are colored in same color.

Summarized, the presented tree topologies show clustering of repeats from the same array which is particular characteristic of monomers belonging to long arrays. Short arrays generally do not show consistent clustering. Such patterns suggest homogenization mechanisms that occur at the array level and have a more dominant effect on long arrays until short ones probably have a limited homogenization possibility. General observation in all extracted satDNA families is that arrays (monomers) from the same chromosome are not clustered with significantly higher frequency with respect to arrays located at non-homologous chromosomes, especially taking into account that some grouped arrays from homologous chromosomes are located very near to each other on the chromosome that could imply the same homogenization effect. Interestingly, dominant clusters in Cl2, Cl3, Cl4 and Cl5 which include arrays from all non-homologous chromosomes, suggest extensive exchanges between non-homologous chromosomes in almost all analyzed satDNAs. Presence of several mixed clusters (heterologous chromosomes) in almost each analysed tree, allows a prediction of satDNA spread through several rounds of interchromosomal exchange and subsequent amplification. Those genome-wide expansion events, within an evolutionary short period of time, imply efficient mechanism of propagation of tandem repeats in non-centromeric genome regions, especially in cluster 5.

4.2.3. Mechanisms of propagation

In order to investigate putative mechanisms of satDNA propagation in non-centromeric part flanking regions of every array in all clusters were analysed. The hypothesis was if satDNAs are just passively carried by expansion of other DNA segments in the genome flanking regions, at least in some of them, should be mutually homologous. The 4kb of both (left and right) flanking region of all arrays were extracted and compared for each cluster separately. Flanking regions built predominantly of unspecified nucleotides (N) were excluded them from analyzes. Also, because of unspecified nucleotides on the distal side some flankings were shorter than 4kb but they were included in analyzes. This comparison showed no similarity in clusters flanking regions except for the clusters 2 and 5. Cluster 2 showed homology only in a small number of left flanking regions (7 out of 36 arrays) until the vast majority of Cluster 5 arrays (22 out of 28) could be grouped according to homologies in left and in right flanking regions. Flanking regions of Cl5 show high similarity with R66 and R140 repeated sequences from sequenced *T. castaneum* genome (class HighA)

obtained by Repeat Scout analyses in Wang et al. 2008 (Supplementary Figure 4.2.1.). Detailed analyses of previously extracted repeated sequences R66 and R140 recovered that they are composed of a part of Cl5 monomer together with part of flanking region. Alignments of R66-like and R140-like flanking regions are presented in Figure 4.2.16. and Figure 4.2.17. R66-like flanking regions show homology in the sequence length of about 1 kb while R140-like flanking regions have homology in 2kb (1kb is homologous in all while last 1kb stays homologous only in part of the sequences). Search analyses of RepBase with R66-R140-like flanking regions as query showed strech of 140 bp with a high homology (84%) to non-autonomous Tc1/Mariner transposon defined in *T. castaneum* thus indicates putative transposon nature of Cluster 5 flanking regions.

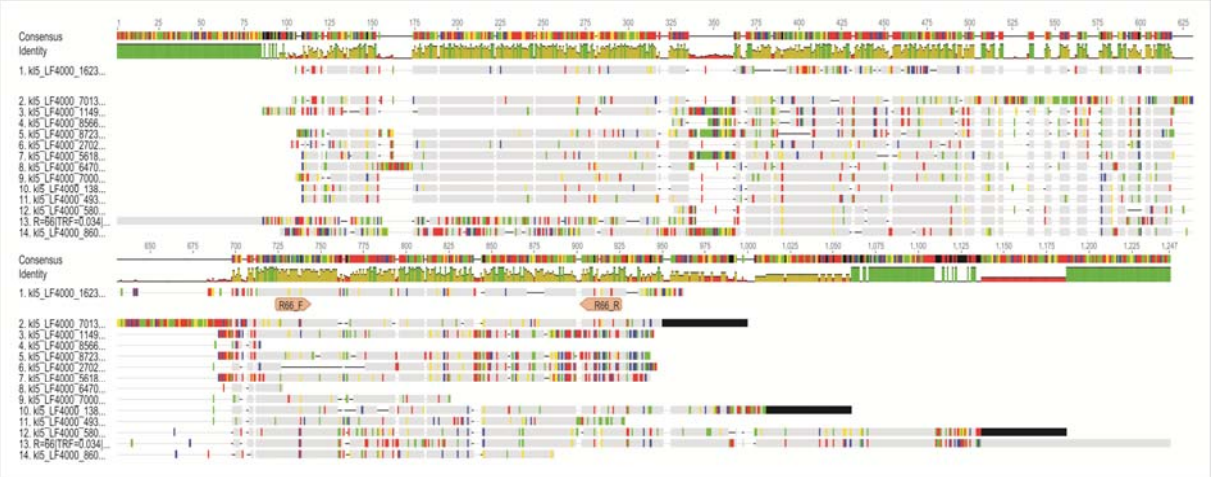


Figure 4.2.16. Alignment of R66-like sequences from flanking regions of cluster 5 with R=66|TRF=0.034|NSEG=0.471|HighA sequence from Wang et al. 2008. Positions identical to the first sequence are shown in gray, differences are shown in color and deletions are indicated with dash. Positions of R66_F and R66_R primers are marked.

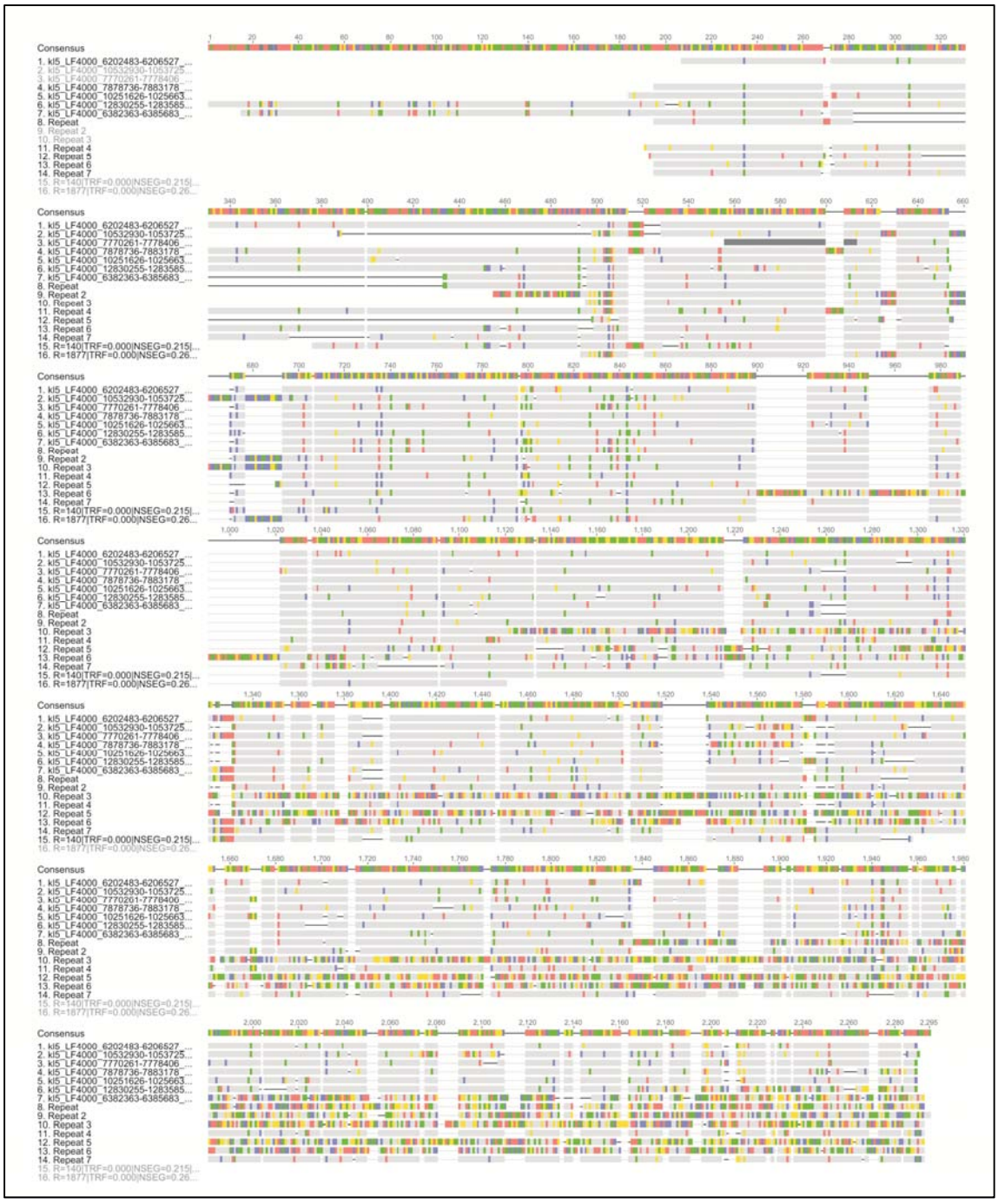


Figure 4.2.17. Alignment of R140-like sequences from flanking regions of cluster 5 with R R=140|TRF=0.000|NSEG=0.215|Mid and R=1877|TRF=0.000|NSEG=0.261|Mid sequence from Wang et al. 2008. Positions identical to the first sequence are shown in gray, differences are shown in color and deletions are indicated with dash.

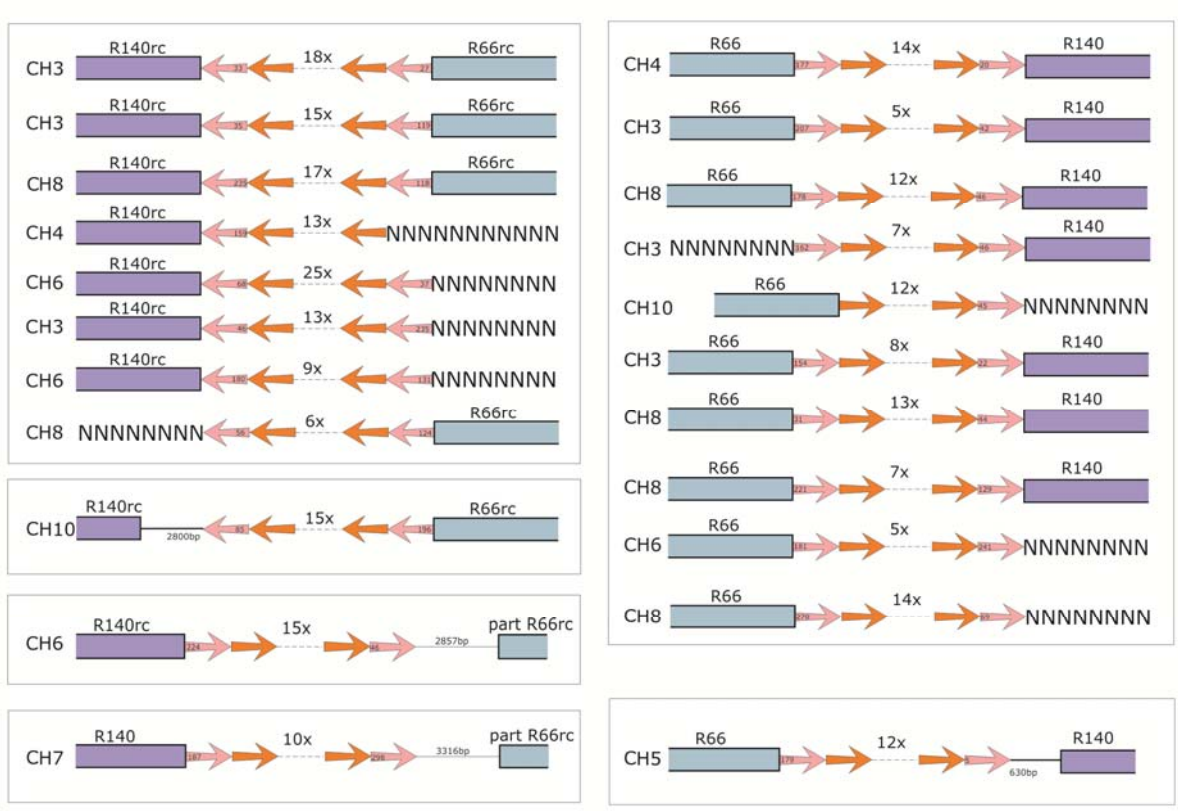


Figure 4.2.18. Schematic representation of different types of flanking regions and monomer arrays of cluster 5. Pink square is R140-like and gray square is R66-like region. Complete monomers are marked with orange and truncated monomers with lilac arrows (numbers inside truncated arrows mark their length). Direction of arrow is according to the consensus sequence direction. Numbers above dashed line indicates how many monomers are in that array. Strait line between array and R140-like or R566-like region marks unspecified unique sequence. N marks unspecified nucleotides. Left side is 5' and right 3' in the genome.

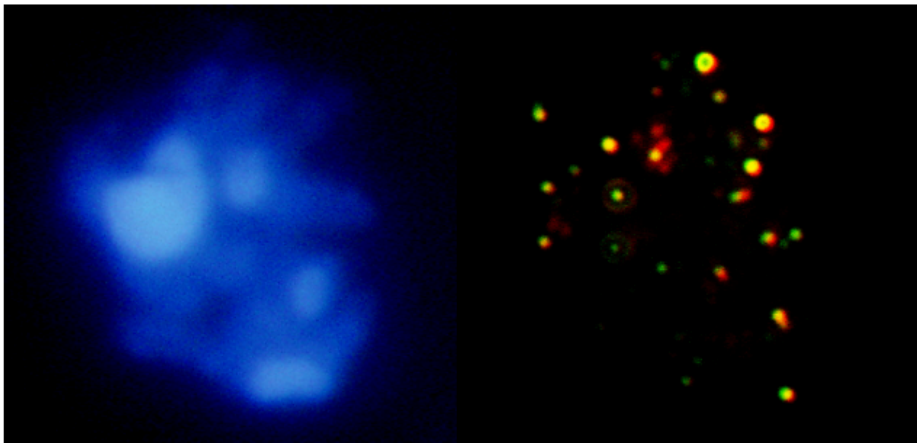


Figure 4.2.19. Two colour FISH with R66 biotin (green signal) and Cl5 monomer Cy3 (red signal) probe. Overlapping positions are yellow. On the left the same nucleus is counter-stained with DAPI.

Significant variations in the sequence length of junction regions were detected in both left and right flanking regions. Schematic representation of flanking regions with arrays of Cl5 family with respect to 5'-3' genome direction, monomer orientation and composition in arrays is presented in Figure 4.2.18. This figure shows the same orientation of monomers with respect to flanking regions in all arrays while elements (flanking+arrays) extend in both directions (5' - 3' and 3' - 5') in the genome.

Double FISH with Cl5 monomer and centromeric satDNA (Figure 4.2.6.) showed extreme expansion of the Cl5 family throughout the genome in comparison to all other families. In order to investigate the genomic location of Cl5 repeats within flanking regions we also performed double FISH experiments with probes from R66-like flanking regions (primers R66_F and R66_R) and Cl5 monomer (Figure 4.2.19.). Results obtained on meiotic promethaphase chromosome spreads produced mainly co-localizing hybridization signals although individual signals of flanking regions and satellites also can be seen. This is in accordance with bioinformatics analysis of arrays where some arrays have different flanking regions.

Diversity of flanking regions in other satDNAs indicates mechanism of self-propagation by insertion of satDNA repeats into different genome environment. For the purpose of checking if there is some part of the monomer sequence that is the preferential site of array insertion we extracted 30pb from the beginning and from the end of each array and assembled them with consensus sequence. We found no preferential site of array insertion for any of the clusters, as can be seen from the Supplementary Figures 4.2.5.-4.2.13. AT tracts can be noticed in the insertion position of several monomers but since these sequences (9 new satellite DNAs) are generally AT rich (Table 4.2.2.) we can’t say that high AT composition of these positions is really significant.

5. DISCUSSION

With this study the existence of satellite DNA library, made of five divergent satDNAs (1a, 1b, 1c, 1d, and 2a), in two recently separated species of root-knot nematodes *M. chitwoodi* and *M. fallax* (van der Beek and Karssen 1997, van Megen et al. 2009) was confirmed and a comprehensive analyses of all of them were performed. 2b satDNA being present only in the *M. chitwoodi* genome is a distinctive element of this satDNA library. This is in agreement with the theory that the presence of novel satDNAs in the library is accompanying the speciation processes (Meštrović et al. 2009). A search for 1a, 1b, 1c, 1d, 2a and 2b counterparts in other congeneric *Meloidogyne* species didn't give any results thus indicating that satDNAs described in this work are specific for *M. chitwoodi* and *M. fallax*. The remarkable characteristic of studied satDNAs is complex organization of repeat units in a form of simple arrays and higher order repeats (HOR). Simple arrays, composed of monomers or dimers, are highly homogenous with the dimers built of two highly divergent monomers. Comparable dimeric organization based on monomers of low sequence similarity (50–60%) was reported in the marmoset (New World monkeys) and it represents an ancient dimeric structure of alphoid sequences (Cellamare et al. 2009). In this work, complex HORs are formed of monomers of divergent satDNAs that range from ones sharing up to 86% sequence identity to apparently unrelated sequences (32% identity). While the first group can be considered as variants of a single satDNA, such as the 1b'H-1bH monomer pair, possible common evolutionary origin of the most divergent monomers is not clear. Such a complex organization of monomers is characteristic for alpha satDNA of human and great apes (Cellamare et al. 2009, Alkan et al. 2007). Alpha satDNA HORs are composed of monomers with relatively high mutual sequence similarity (75–88%) (Rudd and Willard 2004) as opposed to characterized nematode satDNAs. A major difference in organization of simple arrays can be also observed; while alpha satDNA exhibit sequence similarity comparable to that of monomers in alpha HORs (Rudd and Willard 2004), simple arrays of *M. fallax* and *M. chitwoodi* are highly homogenous (94– 97% sequence similarity). Phylogenetic analyses of alpha satDNA monomers in primates and human categorized HOR and monomeric forms as phylogenetically distinct and suggested evolution of both forms from ancestral arrays of monomeric repeats (Rudd et al. 2006). Similar analysis in *M. chitwoodi* and *M. fallax* revealed clustering of HOR units with those from simple arrays, indicating continuous shuffling of monomers between HORs and simple arrays. The only exception is grouping of 1aH and 1aM monomers, in accordance with array affiliation. This result

suggests that mechanisms in addition to unequal crossover over and gene conversion (Dover 1986, Talbert and Henikoff 2010) are probably involved in formation of HORs (see below).

In spite to generally low level of sequence identity (32–64%) among studied satDNAs and in no relation to the organizational pattern in which they are found, examined monomers share two conserved segments, named Box 1 and Box 2. Box 1 is a conserved 17 bp- long segment characteristic for all analyzed satDNAs. This particular motif is observed even in the divergent 2b satDNA, found only in homogeneous monomeric arrays of *M. chitwoodi.* One single deleted nucleotide was found in Box 1 of 1bH and 1b'H monomers which, curiously, appear exclusively as HOR- included elements. This raises the speculative possibility that conserved Box 1 participates in the formation of homogenous simple arrays. It was already proposed that abundant satDNAs may have been selected for amplification because of their ability to bind nuclear proteins (Csink and Henikoff 1998). Interestingly, conserved Box 1 shows significant homology with the human CENP-B box, with identity in 10–12 out of 17 nucleotides. The CENP-B box is a well-described sequence motif of human alpha satDNA which represents a binding site for the CENP-B protein in a subset of alpha satellite HORs (Masumoto et al. 1993). It has been proposed that the CENP-B protein participates in human centromere assembly (Masumoto et al. 1993) but normal chromosome segregation in a mouse CENP-B protein null mutant and absence of CENP-B binding sites at the centromeres of human and mouse Y chromosome make its exact function unclear (Earnshaw et al. 1991, Fowler et al. 2000). DNA sequence motifs similar to the CENP-B box were found in diverse mammalian species, although their satDNA sequences are completely unrelated among themselves and with the alpha satDNA (Kipling et al. 1995, Alkan et al. 2011). For example, seven divergent horse satDNAs exibit CENP B box variants with identity in 9–12 out of 17 nucleotide of human CENP B box (Alkan et al. 2011). Presence of motifs similar to the CENP-B box has also been detected in a number of satDNAs from diverse species outside mammals (López and Edström 1998). In examined nematode species, homology of Box 1 with the human CENP-B box is in the same range found for the CENP-B box in diverse mammalian species (Alkan et al. 2011, Fantaccione et al. 2005). Exceptional feature of the nematode CENP-B box-like motif is significant conservation in the six divergent satDNAs which emphasized it as the most prominent example of the CENP-B box-like sequence out of mammals.

Mechanisms of genetic exchange of satDNAs are hard to study because of repetitive nature of satDNAs arrays. However, our experimental system composed of complex HORs and their counterparts in simple arrays offers a convenient model in which ‘‘beginning’’ and ‘‘end’’ of monomers can be precisely defined. Detailed analyses of *Meloidogyne* satDNA arrays led to observation that junctions between monomers are always located in conserved motifs. Box 1 is found at sites of insertion of the complete 2a monomer into highly divergent 1d and 1c monomers, while in turn, the corresponding segment of equivalent length in 1d and 1c, limited with Box 1, has been extruded (Figure 5.1.). This rearrangement event indicates novel cut-and-paste mechanism that involves the 17 bp-long CENP-B box-like motif and, probably, is related to mechanisms of transposition. It has been already hypothesized that the CENP-B box, in addition to its putative centromeric role, might have a function in satDNA sequence rearrangements (Kipling and Warburton 1997). This assumption is based on similarity of the CENP-B protein and transposases of the pogo family (Casola et al. 2008). Accordingly, the CENP-B box might trigger illegitimate recombination in centromeric areas, in an epigenetically controlled process (Jaco et al. 2008). Highly conserved CENP-B protein homologs were detected in many mammalian species, but not in other metazoans (Casola et al. 2008). In contrast, transposase-derived proteins related to the CENP-B and with putative ability to interact with satDNAs have been detected in diverse invertebrate and vertebrate species (Casola et al. 2008).

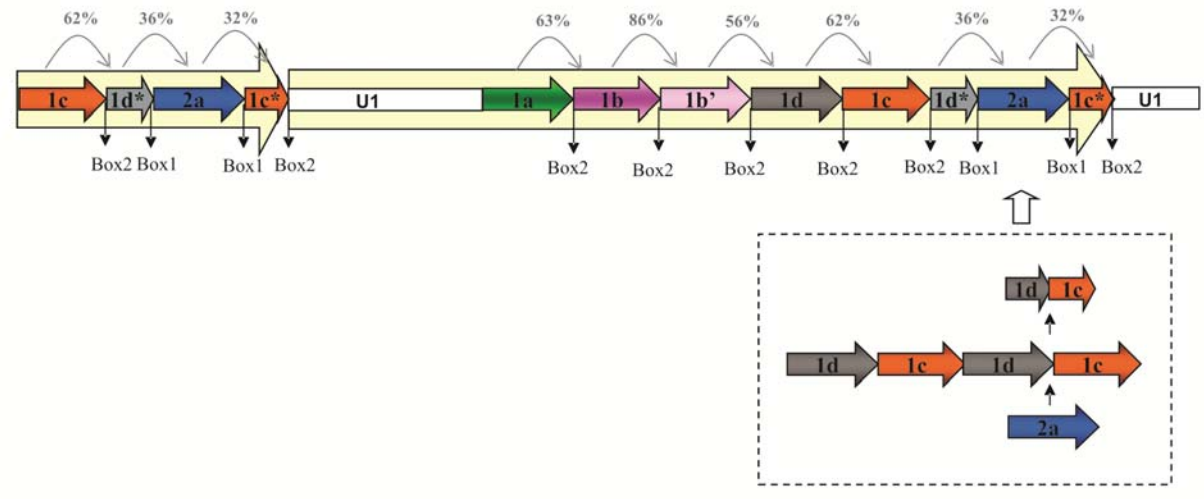


Figure 5.1. The scheme in the frame represents outcome of the proposed cut-and-paste mechanism of 2a insertion in HOR array.

In support, a search in the genome sequence of related species *M. incognita* (Abad et al. 2008) allowed identification of an EST-supported gene encoding a protein with both CENP-B/Tc5 transposase DNA binding domains (Minc05185) (unpublished data) as well as the existence of different repetitive sequences that contain the CENP B box- like motif identical as that observed in this work. The conserved Box 2 is a sequence motif composed of A/T/C tracts, found as a 20 bp- long transition region of all group 1 monomers in HORs. This indicates that homopolymeric tracts which have been found as a common feature of many satellites (Plohl et al. 2008), participate in sequence recombination events in *Meloidogyne*. Since divergent monomers are involved, a mechanism of illegitimate recombination mediated by Box 2 can be assumed. Illegitimate recombination was previously proposed as a mechanism responsible for interspersion of long arrays generating abrupt switches between nonhomologous satDNAs in *Drosophila* (Kuhn et al. 2009). While switches between unrelated arrays in *Drosophila* were detected as relatively rare events, our results nominate Box 2 as promoter of recombination acting frequently on DNA fragments of near monomer size. The minimal observed junction length of about 20 bp in both Box 1 and Box 2 is in accordance with the length of recombination breakpoints in human alpha-satellite (Warburton et al. 1993). In support to this, the role in satDNA shuffling can be assumed by presence of different conserved regions of similar length, as observed in the MEL 172 satDNA family identified in several *Meloidogyne* species (Meštrović et al. 2006) and in other, such as *Arabidopsis* (Hall et al. 2005).

Genome-wide annotation and study of satellite-rich regions from reference assemblies of complex genomes is a challange. Due to the long arrays composed of nearly identical monomer units, satellite DNA remain the most poorly mapped areas of the genomes. However, satDNA annotation, especially in euchromatin regions, is important not only for filling gaps in assembled genomes but also for performing evolutionary studies and studying mechanisms involved in emergence and expansion of these sequences in order to better understand the impact of abundant satDNAs on genome organization and expression. Satellite-rich whole-genome assembly of coleopteran *Tribolium castaneum* based on whole genome shotgun (WGS) reads as well as Fosmid and BAC end sequences offers an exceptional platform for genome-wide analyses of satellite DNA repeats.

To build a database of putative satDNAs TRF algorithm on assembled chromosomes of the *T. castaneum* genome assembly was first applied. Further, arrays with repeats in the

range of 100-500 bp and with monomer number equal or higher than 5 were selected. The total amount of those arrays was 1.63 Mb, which constitutes about 1.04 % of the assembled genome. Arrays are not uniformly distributed between chromosomes showing higher density in five chromosomes, the result being in agreement with the previous report obtained on analysis of overall repeat fraction (HighA and TE classes) (Wang et al 2008). It is important to note that even arrays defined here as long (≥5 copies) are dispersed in non-centromeric and thus potentially euchromatic region on each chromosome.

Long gaps in the assembled genome projects are likely to represent regions of highly repetitive DNA that may have not been sequenced or assembled. In support to this, FISH and quantitative analysis of newly characterized satDNAs (Cluster 1 to Cluster 10) found in the assembled genome recovered existence of significantly higher amounts of these non-centromeric satDNAs. This is particularly obvious in Cluster 1 FISH analysis, where highly abundant signals suggest presence of long unassembled arrays. Presence of only dispersed arrays with few copies of TCAST centromeric satDNA in the assembled genome indicates complete lack of centromeric region. Although anomalies in sequencing and/or assembly coverage between heterochromatic and euchromatic regions can results in discrepancy of the distribution profiles, our results definitively indicate presence of a significant portion of long arrays of tandem repeats out of centromeric regions in the genome.

Chromosomal distribution of (Cluster 1 to Cluster 10) satDNAs is clearly distinct from that of centromeric TCAST satDNA. Based on bioinformatic and FISH analyses, all centromeric TCAST subfamilies are almost exclusively located in centromeric regions. Previous data also revealed only few short arrays of TCAST-like elements embedded in euchromatic genome portions and some of them were identified within a complex unit that resembles a DNA transposon (Brajković et al. 2012). Similar distribution profiles with large blocks of tandem repeats in the pericentromeric regions and short arrays of up to 6 or 12 monomers located in the euchromatin were defined for 1.688 (Kuhn et al. 2012) and *Rsp* satDNA (Larracuente 2014), respectively, in the *Drosophila melanogaster* genome. Authors suggest that limited number of 1.688 satDNA repeats in euchromatic arrays could be selectively constrained by their possible role as gene regulators (Kuhn et al. 2012). In humans, besides alpha satDNA, prominent arrays of classical satDNAs (e.g. satII, satIII, gamma) are exclusively located in pericentromeric regions of many chromosomes (Warburton et al. 2008, Lee et al. 2000). Distribution profile of satDNAs in *T. castaneum*

suggest *certain kind of* chromosome compartmentation with limited transfer of satDNA between the centromeric region and euchromatin.

Structural analysis of tandem repeats in assembled *T. castaneum* genome revealed correlation between the monomer number in arrays and the monomer length. There is an obvious tendency of predominance of 170 and 360 bp long monomers if number of monomers in an array increases. In accordance, in this range are monomer lengths of satDNAs (Cluster 1 to Cluster 9) detected as the most abundant genome components in non-centromeric regions, as well as monomer length in all 5 subfamilies of centromeric TCAST satDNA. Recent data on human and plants revealed the periodicity of CenH3 nucleosomes to be exactly the same as the monomer length of the satDNA, which is therefore in phase with the sequence (Hasson et al. 2013; Zhang et al. 2013). Current understanding of this phenomenon is given by the hypothesis that links preferred monomer length in centromeric satDNAs and the length of DNA wrapped around 1 or 2 nucleosomes as a requirement that may facilitate regular phasing of nucleosomes in the centromere (Heslop-Harrison and Schwarzacher 2013). Further structural analysis of 9 non-centromeric satDNAs revealed periodical distribution of A or T >4 tracts, in the same manner as observed in TCAST and centromeric satDNAs in several other *Tribolium* species (Mravinac et al. 2004). In this regard, periodically distributed tracts of As and/or Ts, present in many centromeric satDNAs of tenebrionid beetles, define the sequence-induced curvature of DNA helix axis and could facilitate the tight packing of DNA in centromeric heterochromatin (Barceló et al. 1998). Results obtained in this study did not suggest any differences in structural features as monomer lengths preferences and nucleotide pattern between centromeric and non-centromeric satDNAs. We hypothesize that similarly as in the centromeric region, monomer length and A or T >4 tracts could be equally important for expansion of tandem repeats in non-centromeric parts and formation of putative micro-heterochromatic regions within euchromatin. It has been proposed that the satellite monomer length longer of two nucleosomes is rare because longer sequences are unlikely for nucleosome stabilizing. Overall analysis of TR in *T. castaneum* assembled genome is also in accordance with this, showing that proportion of arrays with over 380 bp long monomers dramatically decreases. As already known *T. castaneum* has a higher prevalence of satellite DNA in the genome in comparison with *Drosophila* sp. and our results confirm a presence a long arrays in

euchromatic regions. In this sense, it can be hypothesized that *T. castaneum* genome is more flexible with respect to accumulation of satellite DNA in euchromatic regions.

Within this study, the most prominent non-centromeric satDNAs in *T. castaneum* genome assembly were further analysed in detail in order to define evolutionary trends of repeats and mechanisms of dispersion through the genome. It was done by combining bioinformatics and experimental approach. To address these issues the availability of whole-genome assembly of *T. castaneum* genome was the most important step in making possible the examination of repeat sequence evolution at the chromosomal and at the repeat-array level. Phylogenetic analyses of monomers show similar evolutionary scheme for all analyzed families. Clustering of repeats from the same array is observed mainly for long arrays while monomers that originate from shorter arrays (5-7 monomers) often show dispersed formation in phylogenetic trees. These data suggest that homogenization mechanisms are more effective on long arrays than on short ones. In addition, phylogenetic analyses recovered extensive exchanges between non-homologous chromosomes in almost all analyzed satDNAs. This trend is particularly significant in Cl2, Cl3, Cl4 and Cl5 families where the dominant groups include arrays from all chromosomes. Presence of several mixed groups (non homologous chromosomes) in almost each analyzed tree allows a prediction of satDNA spread through several rounds of interchromosomal exchange and subsequent amplification. Those genome-wide expansion events within an evolutionary short period of time imply efficient mechanism of propagation of tandem repeats in non-centromeric genome regions, especially in Cluster 5. Previous studies that were dealing with the evolutionary trends of satellite sequences in the genome are mostly made on centromeric DNA. The conclusion drawn from these analyzes is that intrahomologous exchange is more frequent than interchromosomal. The most prominent example is the human alpha satellite DNA whose higher order (HOR) units show difference in their monomer composition and length, what makes them chromosome-specific (Rudd and Willard 2004). Further, in order to recover the mechanism involved in expansion of satDNAs thought the *T. castaneum* genome we analyzed flanking regions of arrays. We recovered homologous left and right flanking regions in extremely dispearsed satDNA family Cl5. Search analyses of RepBase database with flanking regions as query showed a partial homology with putative non-autonomous Tc1/Mariner. This implies that repeat sequences of Cl5 were distributed by a certain transposition activity over various chromosomal locations. Interspersed repeats are mainly

represented by transposable elements (TEs), until satDNAs usually reside in centromeric compartments of genome. TEs have effective mechanisms of proliferation and movement throughout the genome. A recent study in which internal tandem repeats were found in some TEs provokes a hypothesis that onset and spread of tandem repeats can be linked to processes of transposition (Šatović and Plohl 2013). In plants, a hypervariable region of one LTR-retrotransposon was found expanded into tandem repeats of a satDNA in the pea (*Pisum sativum*) genome (Macas et al. 2009). Similarly, *Zea mays* centromeres became enriched in tandem repeats derived from LTRs and untranslated regions of two unrelated centromere-specific retrotransposons, what probably happened in two independent evolutionary events (Sharma et al. 2013). In contrast, diversity of flanking regions in other satDNAs indicates mechanism of self-propagation by insertion of satDNA repeats into different genome environment.

In summary, satDNA annotation and characterization specially in non-centromeric/euchromatin regions is important not only for filling gaps in assembled genomes but also for understanding of evolutionary trends and mechanisms involved in emergence and expansion of these sequences. Importantly, the suggested approach may help to overall understanding of composition, organization and sequences dynamic in complex genomes.

6. CONCLUSIONS

Our analysis of evolutionary trends of satDNAs in recently separated *Meloidogyne* species revealed:

1. Complex organization of monomers in two *Meloidogyne* species, characterized by highly homogenous simple arrays and by higher order repeats (HORs), composed of highly divergent monomers were disclosed.

2. Despite sequence differences in five analysed satDNAs, two conserved motifs were recovered. Box 1 turned out to be highly similar to the CENP-B box of human alpha satDNA.

3. The onset of this organizational pattern was mediated by conserved Box 1 and Box 2 sequence motifs and the two mechanisms are envisaged in this process: satDNA transposition and illegitimate recombination.

4. HORs can represent a template from which monomers with conserved CENP-B box-like segments can be amplified and form high copy number arrays.

Genome-wide analyses of satellite DNAs in assembled *T. castaneum* genome offered important insights into the evolutionary trends of tandems repeats from a whole-genome perspective:

1. In contrast of previous findings in genus *Drosophila* our analyzes showed that non-centromeric regions of the genome contain significant portion of tandem repeats organized in long arrays.

2. Presence of several mixed clusters (non homologous chromosomes) in almost each analyzed phylogenetic tree allows a prediction of satDNA spread through several rounds of interchromosomal exchange and subsequent amplification.

3. Those genome-wide expansion events within an evolutionary short period of time imply efficient mechanism of propagation of tandem repeats in non-centromeric genome regions.

4. The finding of homologous flanking regions with transposable elements feature in extremely dispersed satDNA family imply a putative role of these regions in expansion of tandem repeats through the genome.

In summary, this PhD thesis represents the comprehensive characterization of satellite DNAs in different model organisms combining bioinformatical and experimental approaches and provides important insights into mechanisms involved in the genesis, evolution and spread of satellite DNAs in complex genomes.

7. REFERENCES

Abad P, Gouzy J, Aury J-M, et al. 2008. Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita. Nat. Biotechnol 26:909–915.

Alexandrov I, Kazakov A, Tumeneva I, Shepelev V, Yurov Y. 2001. Alpha- satellite DNA of primates: old and new families. Chromosoma 110:253–266.

Alkan C, Cardone MF, Catacchio CR, et al. 2011. Genome-wide characterization of centromeric satellites from multiple mammalian genomes. Genome Res. 21:137–145.

Alkan C, Ventura M, Archidiacono N, Rocchi M, Sahinalp SC, Eichler EE. 2007. Organization and evolution of primate centromeric DNA from whole-genome shotgun sequence data. PLoS Comput Biol 3:1807–1818.

Altemose N, Miga KH, Maggioni M, Willard HF. 2014. Genomic characterization of large heterochromatic gaps in the human genome assembly. PLoS Comput. Biol. 10:1–14.

Barceló F, Gutiérrez F, Barjau I, Portugal J. 1998. A theoretical perusal of the satellite DNA curvature in tenebrionid beetles. J. Biomol. Struct. Dyn. 16:41–50.

Van der Beek JG, Karssen G. 1997. Interspecific Hybridization of Meiotic Parthenogenetic Meloidogyne chitwoodi and M. fallax. Phytopathology 87:1061–1066.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27:573–580.

Bird DM, Williamson VM, Abad P, McCarter J, Danchin EGJ, Castagnone-Sereno P, Opperman CH. 2009. The genomes of root-knot nematodes. Annu Rev Phytopathol 47:333–351.

Del Bosque M, Navajas- Pérez R, Panero J, Fernández- González, A Garrido- Ramos M. 2011. A satellite DNA evolutionary analysis in the North American endemic dioecious plant Rumex hastatulus (Polygonaceae). Genome 54:253–260.

Brajković J, Feliciello I, Bruvo-Mađarić B, Ugarković D. 2012. Satellite DNA-Like Elements Associated With Genes Within Euchromatin of the Beetle Tribolium castaneum. G3 Genes|Genomes|Genetics 2:931–941.

Bulazel K, Metcalfe C, Ferreri GC, Yu J, Eldridge MDB, O'Neill RJ. 2006. Cytogenetic and molecular evaluation of centromere-associated DNA sequences from a marsupial (Macropodidae: Macropus rufogriseus) X chromosome. Genetics 172:1129–1137.

Cafasso D, Cozzolino S, De Luca P, Chinali G. 2003. An unusual satellite DNA from Zamia paucijuga (Cycadales) characterised by two different organisa- tions of the repetitive unit in the plant genome. Gene 311:71–79.

Casola C, Hucks D, Feschotte C. 2008. Convergent domestication of pogo-like transposases into centromere-binding proteins in fission yeast and mammals. Mol Biol Evol 25:29–41.

Castagnone-Sereno P, Leroy H, Semblat JP, Leroy F, Abad P, Zijlstra C. 1998. Unusual and strongly structured sequence variation in a complex satellite DNA family from the nematode Meloidogyne chitwoodi. J Mol Evol 46:225–233.

Castagnone-Sereno P, Semblat JP, Leroy F, Abad P. 1998. A new AluI satellite DNA in the root-knot nematode Meloidogyne fallax: relationships with satellites from the sympatric species M. hapla and M. chitwoodi. Mol Biol Evol 15:1115–1122.

Cellamare A, Catacchio CR, Alkan C, et al. 2009. New insights into centromere organization and evolution from the white-cheeked gibbon and marmoset. Mol Biol Evol 26:1889–1900.

Cheng Z, Dong F, Langdon T, Ouyang S, Buell CR, Gu M, Blattner FR, Jiang J. 2002. Functional Rice Centromeres Are Marked by a Satellite Repeat and a Centromere-Specific Retrotransposon. 14:1691–1704.

Csink AK, Henikoff S. 1998. Something from nothing: the evolution and utility of satellite repeats. Trends Genet 14:200–204.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods 9:772. A

Dernburg AF. 2001. Here, there, and everywhere: kinetochore function on holocentric chromosomes. J Cell Biol 153:F33–F38.

Dover G. 1982. Molecular drive: a cohesive mode of species evolution. Nature 299:111–117.

Dover GA. 1986. Molecular drive in multigene families: How biological novelties arise, spread and are assimilated. Trends Genet 2:159–165.

Earnshaw WC, Bernat RL, Cooke CA, Rothfield NF. 1991. Role of the centromere/kinetochore in cell cycle control. Cold Spring Harb. Symp. Quant. Biol. 56:675–685.

Earnshaw WC, Tomkiel JE. 1992. Centromere and kinetochore structure. Curr. Biol. 2:156.

Fantaccione S, Pontecorvo G, Zampella V. 2005. Molecular characterization of the first satellite DNA with CENP-B and CDEIII motifs in the bat Pipistrellus kuhli. FEBS Lett. 579:2519–2527.

Feliciello I, Chinali G, Ugarković Đ. 2011. Structure and population dynamics of the major satellite DNA in the red flour beetle Tribolium castaneum. Genetica 139:999–1008.

Fowler KJ, Hudson DF, Salamonsen L a, Edmondson SR, Earle E, Sibson MC, Choo KH. 2000. Uterine dysfunction and genetic modifiers in centromere protein B-deficient mice. Genome Res 10:30–41.

Fry K, Salser W. 1977. Nucleotide sequences of HS- alpha satellite DNA from kangaroo rat Dipodomys ordii and characterization of similar sequences in other rodents. Cell 12:1069–1084.

Gaffney PM, Pierce JC, Mackinley AG, Titchen DA, Glenn WK. 2003. Pearl, a novel family of putative transposable elements in bivalve mollusks. J. Mol. Evol. 56:308–316.

Gassmann R, Rechtsteiner A, Yuen KW, et al. 2013. An inverse relationship to germline transcription defines centromeric chromatin in C. elegans. Nature 484:534–537.

Gelfand Y, Rodriguez A, Benson G. 2006. TRDB--the Tandem Repeats Database. Nucleic Acids Res. 35:D80–D87.

Goff SA, Ricke D, Lan T, et al. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science 296:92–100.

Guindon S, Gascuel O. 2003. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. Syst. Biol. 52:696–704.

Hall SE, Kettler G, Preuss D. 2003. Centromere satellites from Arabidopsis populations : maintenance of conserved and variable domains. Genome Res 13:195–205.

Hall SE, Luo S, Hall AE, Preuss D. 2005. Differential rates of local and global homogenization in centromere satellites from Arabidopsis relatives. Genetics 170:1913–1927.

Hall TA. 1999. BioEdit- a user-friendly biological sequence alignment editor and analysis program for Windows 95-98-NT.pdf. Nucleic Acids Symp Ser. 41:95–98.

Hasson D, Panchenko T, Salimian KJ, Salman MU, Sekulic N, Alonso A, Warburton PE, Black BE. 2013. The octamer is the major form of CENP-A nucleosomes at human centromeres. Nat. Struct. Mol. Biol. 20:687–695.

Heckmann S, Macas J, Kumke K, et al. 2013. The holocentric species Luzula elegans shows interplay between centromere and large-scale genome organization. Plant J 73:555–565.

Henikoff S, Ahmad K, Malik HS. 2001. The centromere paradox: stable inheritance with rapidly evolving DNA. Science 293:1098–1102.

Heslop-Harrison JSP, Schwarzacher T. 2013. Nucleosomes and centromeric DNA packaging. Proc. Natl. Acad. Sci. U. S. A. 110:19974–19975.

Hillier LW, Miller RD, Baird SE, Chinwalla A, Fulton L a, Koboldt DC, Waterston RH. 2007. Comparison of C. elegans and C. briggsae genome sequences reveals extensive conservation of chromosome organization and synteny. PLoS Biol. 5:1603–1616.

Jaco I, Canela A, Vera E, Blasco M a. 2008. Centromere mitotic recombination in mammalian cells. J Cell Biol 181:885–892.

Jonstrup AT, Thomsen T, Wang Y, Knudsen BR, Koch J, Andersen AH. 2008. Hairpin structures formed by alpha satellite DNA of human centromeres are cleaved by human topoisomerase IIalpha. Nucleic Acids Res. 36:6165–6174.

Juan C, Vazquez P, Rubio J, Pettipierre E, Godfrey M. 1993. Presence of highly repetitive DNA sequences in Tribolium flour-beetles. Heredity 70:1–8.

Jurka J, Kapitonov V V, Pavlicek a, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. 110:462–467.

Kapitonov V V, Jurka J. 2003. Molecular paleontology of transposable elements in the Drosophila melanogaster genome. Proc. Natl. Acad. Sci. U. S. A. 100:6569–6574.

Kejnovsky E, Kubat Z, Macas J, Hobza R, Mracek J, Vyskot B. 2006. Retand: a novel family of gypsy-like retrotransposons harboring an amplified tandem repeat. Mol. Genet. Genomics 276:254–263.

Kim HS, Murphy T, Xia J, et al. 2010. BeetleBase in 2010: revisions to provide comprehensive genomic information for Tribolium castaneum. Nucleic Acids Res. 38:D437–D442.

King LM, Cummings MP. 1997. Satellite DNA Repeat Sequence Variation is Low in Three Species of Burying Beetles in the Genus Nicrophorus (Coleoptera: Silphidae). Mol Biol Evol 14:1088–1095.

Kipling D, Mitchell AR, Masumoto H, Wilson HE, Nicol L, Cooke HJ. 1995. CENP-B binds a novel centromeric sequence in the Asian mouse Mus caroli. Mol Cell Biol 15:4009–4020.

Kipling D, Warburton PE. 1997. Centromeres, CENP-B and Tigger too. Trends Genet 13:141–145.

Komissarov AS, Gavrilova E V, Demin SJ, Ishov AM, Podgornaya OI. 2011. Tandemly repeated DNA families in the mouse genome. BMC Genomics 12:531.

Koukalova B, Moraes A, Renny- Byfield S, Matyasek R, Leitch A, Kovarik A. 2010. Fall and rise of satellite repeats in allopolyploids of Nicotiana over c. 5 million years. New Phytol 186:148–160.

Kuhn GCS, Küttler H, Moreira-Filho O, Heslop-Harrison JS. 2012. The 1 . 688 repetitive DNA of Drosophila: Concerted evolution at different genomic scales and association with genes. Mol Biol Evol 29:7–11.

Kuhn GCS, Teo CH, Schwarzacher T, Heslop-Harrison JS. 2009. Evolutionary dynamics and sites of illegitimate recombination revealed in the interspersion and sequence junctions of two nonhomologous satellite DNAs in cactophilic Drosophila species. Heredity 102:453–464.

Kumar S, Tamura K, Nei M. 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Br. Bioinform 5:150–163.

Lander ES, Linton LM, Birren B, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Larracuente AM. 2014. The organization and evolution of the Responder satellite in species of the Drosophila melanogaster group: dynamic evolution of a target of meiotic drive. BMC Evol. Biol. 14:1–12.

Lee C, Critcher R, Zhang JG, Mills W, Farr CJ. 2000. Distribution of gamma satellite DNA on the human X and Y chromosomes suggests that it is not required for mitotic centromere function. Chromosoma 109:381–389.

Lee H-RR, Neumann P, Macas J, Jiang J. 2006. Transcription and evolutionary dynamics of the centromeric satellite repeat CentO in rice. Mol Biol Evol 23:2505–2520.

Lin C, Li Y. 2006. Chromosomal distribution and organization of three cervid satellite DNAs in Chinese water deer (Hydropotes inermis). Cytogenet Genome Res. 114:147–154.

López CC, Edström JE. 1998. Interspersed centromeric element with a CENP-B box-like motif in Chironomus pallidivittatus. Nucleic Acids Res 26:4168–4172.

Ma J, Jackson SA. 2006. Retrotransposon accumulation and satellite amplification mediated by segmental duplication facilitate centromere expansion in rice. Genome Res 16:251–259.

Macas J, Koblízková A, Navrátilová A, Neumann P. 2009. Hypervariable 3' UTR region of plant LTR-retrotransposons as a source of novel satellite repeats. Gene 448:198–206.

Mahtani MM, Willard HF. 1998. Physical and Genetic Mapping of the Human X Chromosome Centromere : Repression of Recombination Physical and Genetic Mapping of the Human X Chromosome Centromere : Repression of Recombination. Genome Res 8:100–110.

Martínez-Balbás A, Rodríguez-Campos A, García- Ramírez M, Sainz J, Carrera P. 1990. Satellite DNAs contain sequences that induced curvature. Biochemistry 29:2342–2348.

Masumoto H, Yoda K, Ikeno M, Kitagawa K, Muro Y, Okazaki T. 1993. Properties of CENP-B and its target sequence in a satellite DNA. In: Baldev KV, editor. In: Chromosome Segregation and Aneuploidy. Vol. 72. NATO ASI Ser. H 72. p. 31–43.

Van Megen H, van den Elsen S, Holterman M, Karssen G, Mooyman P, Bongres T, Holovachov O, Bakker J, Helder J. 2009. A phylogenetic tree of nematodes based on about 1200 full-length small subunit ribosomal DNA sequences. Nematology 11:927–950.

Melters DP, Bradnam KR, Young H a, et al. 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol 14:R10.

Meštrović N, Castagnone-Sereno P, Plohl M. 2006. Interplay of selective pressure and stochastic events directs evolution of the MEL172 satellite DNA library in root-knot nematodes. Mol Biol Evol 23:2316–2325.

Meštrović N, Plohl M, Castagnone-Sereno P. 2009. Relevance of satellite DNA genomic distribution in phylogenetic analysis: a case study with root-knot nematodes of the genus Meloidogyne. Mol Phylogenet Evol 50:204–208.

Meštrović N, Plohl M, Mravinac B, Ugarković Đ. 1998. Evolution of satellite DNAs from the genus Palorus - experimental evidence for the library hypothesis. Mol Biol Evol 15:1062–1068.

Miller WJ, Nagel A, Bachmann J, Bachmann L. 2000. Evolutionary dynamics of the SGM transposon family in the Drosophila obscura species group. Mol. Biol. Evol.] 17:1597–1609.

Mola L, Papeschi A. 2006. Holokinetic chromosomes at a glance. BAG J Basic Appl Genet 17:17–33.

Mravinac B, Plohl M, Meštrović N, Ugarković Đ. 2002. Sequence of PRAT satellite DNA “frozen” in some Coleopteran species. J Mol Evol 54:774–783.

Mravinac B, Plohl M, Ugarković Đ. 2004. Conserved patterns in the evolution of Tribolium satellite DNAs.pdf. Gene 332:169–177.

Mravinac B, Plohl M. 2007. Satellite DNA junctions identify the potential origin of new repetitive elements in the beetle Tribolium madens. Gene 394:45–52.

Mravinac B, Plohl M. 2010. Parallelism in evolution of highly repetitive DNAs in sibling species. Mol Biol Evol 27:1857–1867.

Mravinac B, Ugarković E, Franjević D, Plohl M. 2005. Long inversely oriented subunits form a complex monomer of Tribolium brevicornis satellite DNA. J Mol Evol 60:513–525.

Okamoto Y, Nakano M, Ohzeki J, Larionov V, Masumoto H. 2007. A minimal CENP-A core is required for nucleation and maintenance of a functional human centromere. EMBO J. 26:1279–1291.

Opperman CH, Bird DM, Williamson VM, et al. 2008. Sequence and genetic map of Meloidogyne hapla: A compact nematode genome for plant parasitism. Proc Natl Acad Sci 105:14802–14807.

Paar V, Gluncić M, Rosandić M, Basar I, Vlahović I. 2011. Intragene higher order repeats in neuroblastoma breakpoint family genes distinguish humans from chimpanzees. Mol. Biol. Evol. 28:1877–1892.

Perelman P, Johnson WE, Roos C, et al. 2011. A molecular phylogeny of living primates. PLoS Genet. 7:1–17.

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

Plohl M, Meštrović N, Mravinac B. 2012. Satellite DNA Evolution. In: Repetitive DNA. p. 126–152.

Plohl M, Petrović V, Luchetti A, Ricci A, Satović E, Passamonti M, Mantovani B. 2010. Long-term conservation vs high sequence divergence: the case of an extraordinarily old satellite DNA in bivalve mollusks. Heredity 104:543–551.

Prasad AB, Allard MW, Green ED. 2008. Confirming the phylogeny of mammals by use of large comparative sequence data sets. Mol. Biol. Evol. 25:1795–1808.

Radic MZ, Saghbinil M, Elton TS, Reeves R, Hamkalo BA. 1992. Hoechst 33258 , distamycin A , and high mobility group protein I (HMG-I) compete for binding to mouse satellite DNA. 1:602–608.

Richards S, Gibbs R a, Weinstock GM, et al. 2008. The genome of the model beetle and pest Tribolium castaneum. Nature 452:949–955.

Rudd MK, Willard HF. 2004. Analysis of the centromeric regions of the human genome assembly. Trends Genet 20:529–533.

Rudd MK, Wray GA, Willard HF. 2006. The evolutionary dynamics of α-satellite.pdf. Genome Res 16:88–96.

Schueler MG, Sullivan B a. 2006. Structural and functional dynamics of human centromeric chromatin. Annu. Rev. Genomics Hum. Genet. 7:301–313.

Sharma A, Wolfgruber TK, Presting GG. 2013. Tandem repeats derived from centromeric retrotransposons. BMC Genomics 14:142.

Smith GP. 1976. Evolution of repeated DNA sequences by unequal crossover. Science (80-.). 191:528–535.

Stam M, Belele C, Ramakrishna W, Dorweiler JE, Bennetzen JL, Chandler VL. 2002. The Regulatory Regions Required for B Paramutation and Expression Are Located Far Upstream of the Maize b1 Transcribed Sequences. Genetics 162:917–930.

Stephan W. 1986. Recombination and the evolution of satellite DNA. Genet Res 47:167–174.

Swofford DL. 2002. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer Assoc. Sunderland, Massachusetts.:1–142.

Šatović E, Plohl M. 2013. Tandem repeat-containing MITE elements in the clam Donax trunculus. Genome Biol. Evol. 5:2549–2559.

Talbert PB, Henikoff S. 2010. Centromeres convert but don't cross. PLoS Biol 8:1–5.

Tares S, Cornuet J-M, Abad P. 1993. Characterization of an Unusually Conserved AluI Highly Reiterated DNA Sequence Family From the Honeybee, Apis mellifera. Genetics 134:1195–1204.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876–4882.

Ugarković Đ, Podnar M, Plohl M. 1996. Satellite DNA of the Red Flour Beetle Tribolium castaneum-Comparative Study of Satellites from the Genus Tribolium. Mol Biol Evol 13:1059–1066.

Wang S, Lorenzen MD, Beeman RW, Brown SJ. 2008. Analysis of repetitive DNA distribution patterns in the Tribolium castaneum genome. Genome Biol. 9:14.

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

Warburton PE, Waye JS, Willard HF. 1993. Nonrandom localization of recombination events in human alpha satellite repeat unit variants: implications for higher-order structural characteristics within centromeric heterochromatin. Mol Cell Biol [Internet] 13:6520–6529.

Willard H, Waye J. 1987. Hierarchical order in chromosome- specific human alpha satellite DNA. Trends Genet 3:192–198.

Zhang T, Talbert PB, Zhang W, Wu Y, Yang Z, Henikoff JG, Henikoff S, Jiang J. 2013. The CentO satellite confers translational and rotational phasing on cenH3 nucleosomes in rice centromeres. Proc. Natl. Acad. Sci. U. S. A. 110:E4875–E4883.

Zijlstra C. 2000. Identification of Meloidogyne chitwoodi, M. fallax and M. hapla based on SCAR-PCR: a powerful way of enabling reliable identification of populations or individuals that share common traits. Eur J Plant Pathol 106:283–290.

Žinić SD, Ugarković Đ, Cornudella L, Plohl M. 2000. A novel interspersed type of organization of satellite DNAs in Tribolium madens heterochromatin. Chromosom. Res 8:201–212.

8. SUMMARY

Tandemly arrayed non-coding sequences or satellite DNAs (satDNAs) are rapidly evolving segments of eukaryotic genomes, preferentially located on centromeres. However, the mechanisms involved in the genesis, evolution and spread of satellite DNAs in complex genomes are only partially elucidate. In order to better understand mechanisms related to evolutionary trends of satellite DNA model organisms nematode *Meloidogyne* spp and beetle *Tribolim castaneum* were selected. The specific aims of this study were: (**i**) to characterize of satDNAs library shared by recently separated nematode species *M. fallax* and *M. chitwoodi* **(ii)** to performe structural, organizational and phylogenetic analyzes of sequences in the common satDNA library of *Meloidogyne* species in order to investigate mechanisms of satDNA genesis **(iii)** to analyse of the structure and distribution of satDNAs in the recently sequenced genome of *T. castaneum* **(iv)** to perform extensive bioinformatics and experimental analyses of non centromeric satDNAs in *T. castaneum* genome and reveal the evolutionary forces that govern the distribution and dynamics of non-centromeric satellite.

Structural sequence analyses of satDNAs in the library of *Meloidogyne* species disclosed complex organization patterns of monomers in the form of simple and higher-order repeat (HOR) arrays. Despite sequence differences between five satDNAs, two conserved motifs were recovered. One of them turned out to be highly similar to the CENP-B box of human alpha satDNA, identical in 10–12 out of 17 nucleotides. Analyses of monomer junction regions in complex HORs highlighted the role of short motifs in rearrangements, even among highly divergent sequences. Two mechanisms are proposed to be involved in this process, i.e., putative transposition-related cut-and-paste insertions and/or illegitimate recombination.

In addition, genome-wide identification of new satDNAs in *T. castaneum* by Tandem Repeat Finder and double Fluorescence in situ hybridization cofirmed their non-centromeric localization. Characterized satDNA arrays are almost evenly distributed within the putative heterochromatic and euchromatic regions of chromosomes. Phylogenetic studies of monomers in newly defined satDNAs showed extensive exchange between homologous as well as non-homologous chromosomes suggesting efficient propagation mechanism of tandem repeats in non-centromeric regions. The finding of homologous flanking regions with transposable elements feature in extremely dispersed satDNA family imply a putative role of these regions in expansion of tandem repeats through the genome.

9. SAŽETAK

Uzastopno ponovljene nekodirajuće sekvence, odnosno satelitne DNA (satDNA), su brzo evoluirajući dio eukariotskog genoma, preferencijalno smještene u području centromere. Mehanizmi njihovog nastanka, evolucije i širenja u kompleksnim genomima još uvijek nisu potpuno razjašnjeni. Kako su satDNA glavne komponente peri/centromernih područja kromosoma najveći je broj studija napravljen upravo na ovoj frakciji satelitnih DNA. Iako je centromerna funkcija vrlo konzervirana, satDNA iznimno variraju u sekvenci i udjelu u genomu već i među blisko srodnim vrstama. Na osnovu komparativnih studija centromernih DNA te mapiranja centromernih regija nekih modelnih organizama zaključeno je da centromerna funkcija nije uvjetovana određenom sekvencom, ali da ovo područje preferira uzastopno ponovljene DNA. Smatra se da satDNA sudjeluju u organizaciji i evoluciji centromernog područja te da njihova varijabilnost stimulira reproduktivnu izolaciju pa tako i specijaciju. Sekvenciranje velikog broja genoma i nedavni razvoj novih bioinformatičkih platformi za analizu ponavljajućih sekvenci omogućio je analize satDNA profila na razini cijelog genoma ukazujući na prisutnost ovih regija i u eukromatinu. Nedavne studije satDNA u ovim područjima pokazale su njihovu ulogu u moduliranju genske regulacije, vezu s nekim genskim bolestima te ulogu u akumulaciji razlika koje mogu imati za posljedicu promjenu fenotipa. Na osnovu gore navedenog razvidno je da satDNA ne oblikuju samo centromerno područje nego i ostale djelove genoma tako da studije satDNA i u necentromernim područjima predstavljaju važan korak naprijed u razumijevanju organizacije i funkcioniranja genoma u cjelini.

U svrhu boljeg razumijevanja evolucijskih trendova i mehanizama formiranja te širenja satelitnih DNA razmatranih u ovom radu, odabrani su sljedeći modelni i organizmi: dvije vrste oblića iz roda *Meloidogyne* i kornjaš *Tribolium castaneum*. Postojanje satelitne biblioteke u genomima nedavno odvojenih vrsta roda *Meloidogyne* predstavlja idealan sustav za istraživanje mehanizama formiranja satDNA te otkrivanja evolucijski preferiranih oblika sekvenci. S druge strane model organizam *T. castaneum,* čiji je sekvencirani genom javno dostupan, nudi mogućnost ekstenzivnog istraživanja satDNA na razini cijelog genoma te tako i mogućnost istraživanja mehanizama uključenih u širenje ovih sekvenci u genomu. Specifični ciljevi ovog istraživanja su: **i)** karakterizacija biblioteke satelitnih DNA koju dijele nedavno odvojene vrste oblića *M. chitwoodi* i *M. fallax* **ii)** strukturna, organizacijska i filogenetska analiza sekvenci satelitnih DNA iz zajedničke biblioteke vrsta *M. chitwoodi* i *M. fallax* u svrhu otkrivanja mehanizama nastanka satelitnih DNA **iii)** analiza strukture i

distribucije satelitnih DNA u sekvenciranom genomu kornjaša *T. castanaum* **iv)** temeljite bioinformatičke i eksperimentalne analize ne-centromernih satelitnih DNA u vrsti *T. castaneum* u svrhu otkrivanja evolucijskih trendova koje upravljaju širenjem i dinamikom ne-centromernih satelitnih DNA.

Analiza sekvenci i strukturnih obilježja 5 satelitnih DNA u biblioteci vrsta *Meloidogyne chitwoodi* i *M. fallax* otkrila je njihovu nisku međusobnu sličnost (32–64%) te kompleksnu organizaciju satelitnih monomera iz biblioteke u obliku jednostavnih nizova i jedinica višeg stupnja organizacije, tzv. HOR-ova (od engl. higher-order repeat), usporedivu s organizacijom alfa satelita kod ljudi i ostalih primata. Za razliku od alfa satelita, filogenetska analiza satelitnih DNA kod oblića je pokazala grupiranje monomera iz jednostavnih nizova i iz jedinica višeg stupnja organizacije što ukazuje na stalnu izmjenu monomera između ova dva tipa organizacije. Unatoč činjenici da je pet proučavanih satelitnih DNA međusobno različito, pronađena su dva dijela sekvence (motiva) očuvana kod svih. Jedan od njih, nazvan Box 1, pokazuje visoku sličnost sa 17 parova baza dugim CENP-box-om humanog alfa satelita, konkretno u 10 do 12 parova baza. Dodatna zanimljiva činjenica je da je prijelaz između različitih monomera unutar HOR jedinice točno na mjestu tog motiva, tj. Box-a 1. Drugi 20 parova baza dug očuvani motiv, nazvan Box 2, nalazi se također prelasku između monomera podgrupe 1. Ova opažanja naglašavaju važnost kratkih konzerviranih odsječaka DNA u procesima genomskih rearanžmana, čak i među divergentnim sekvencama. Predložena su dva organizma odgovorna za uočene rearanžmane; *cut and paste* insercija povezana s mehanizmom transpozicije i/ili ilegitimna rekombinacija. Mogućnost sudjelovanja CENP-B box-u sličnih motiva kod oblića u transpozicijskim procesima i dokazana sličnost humanog CENP-B proteina s transpozazom *pogo* porodice ukazuje na moguću novu ulogu CENP-B box-a u sekvenci DNA osim već dokazane u vezivanju centromernih proteina.

Pretraživanjem cijelog sekvenciranog genoma (sastavljenog od 10 kromosoma i nesastavljenih dijelova) kornjaša *T. castaneum* i identifikacijom novih satelitnih DNA uz pomoć algoritma Tandem Repeat Finder-a i dvobojne flourescencijske hibridizacije *in situ* pronađeno je 9 novih, do sada neopisanih, satelitnih DNA. Usporedbom njihove lokalizacije u odnosu na od prije poznati centormerni satelit TCAST, koji čini čak 35% genima, pokazana je ne-centromerna lokalizacija svake od njih. Opisani nizovi satelitnih DNA skoro su jednakomjerno raspoređeni u heterokromatinskom i eukromatinskom području pojedinih kromosoma. Najviše ih ima na kromosomima 3, 6, 8, 9 i 10 što je u skladu s radom iz 2008

(Wang et al. 2008) u kojem je RepeatScout metodom otkriveno da je čak 26% genoma sastavljeno od uzastopnih ponavljanja. Našom analizom ustanovljeno je da 9 opisanih satelitnih DNA čine preko 4% genoma s tim da su pojedine porodice zastupljene od ispod 0.2 pa do preko 1%. Dužina monomera novo otkrivenih satelitnih DNA je oko 170 pb što je u skladu s veličinom monomera kod većine do sada opisanih centromernih satelita. Ovaj se trend objašnjava činjenicom da se dužina poklapa s duljinom nukleosomalne DNA te ukazuje na slične strukturne karakteristike centomernih i ne-centromernih satDNA. Filogenetske analize monomera pojedinačnih novoopisanih satDNA ukazuju na značajnu izmjenu sekvenci među homolognim, ali i ne-homolognim kromosomima, ukazujući na učinkovite mehanizme širenja uzastopno ponovljenih sekvenci u ne-centromernom području, pogotovo kod nekih porodica. Proces homogenizacije učinkovitiji je kod dugih nizova dok je za kratke (5 do 7 monomera) karakteristično raspršenost po filogenetskom stablu. Postojanje homolognih rubnih regija sličnih transponirajućim elementima kod iznimno raširene porodice satelitne DNA ukazuje na moguću ulogu ovih regija u širenju uzastopnih ponavljanja po genomu.

10. CURRICULUM VITAE

EUROPEAN CURRICULUM VITAE FORMAT

PERSONAL INFORMATION

Surname(s) / First name(s)	Pavlek / Martina
Address(es)	Laboratory for structure and function of heteochromatin, Ruđer Bošković Institute, Bijenička 54, 10002 Zagreb
Telephone(s)	+ 385 1 457 1322
Fax(es)	+ 385 1 4561 177
E-mail(s), Web address(s)	mpavlek@irb.hr
Nationality(-ies)	Croatian
Date of birth	October 6 th, 1981
Identification number from Records of Scientific Workers	315522

WORK EXPERIENCE

• Dates (from – to)	08. 05. 2009. - present
Name and address of employer	Ruđer Bošković Institute, Bijenička 54, 10000 Zagreb
Type of business or sector	Scientific research
Occupation or position held	Doctoral student, assistant
Main activities and responsibilities	Scientific research in natural science: biology; molecular biology; molecular genetics; structure, organization, function and evolution of repetitive sequences

EDUCATION

Date	2010. - present
Place of education	Zagreb
Name and type of organisation providing education	Ph.D. studies in Molecular Biosciences (University of Osijek, University of Dubrovnik and Rudjer Boskovic Institute)
Title or qualification awarded	
Date	2000-2006
Place of education	Zagreb
Name and type of organisation providing education	Faculty of Science, University of Zagreb
Title or qualification awarded	B.Sc., ecology

PERSONAL SKILLS AND COMPETENCIES

Mother tongue: Croatian

Other language(s)

Self-assessment	Understanding		Speaking		Writing
European level (*)	Listening	Reading	Spoken interaction	Spoken production	
English	C2	C2	C1	C1	C1
German	A2	B2	A2	A1	A2

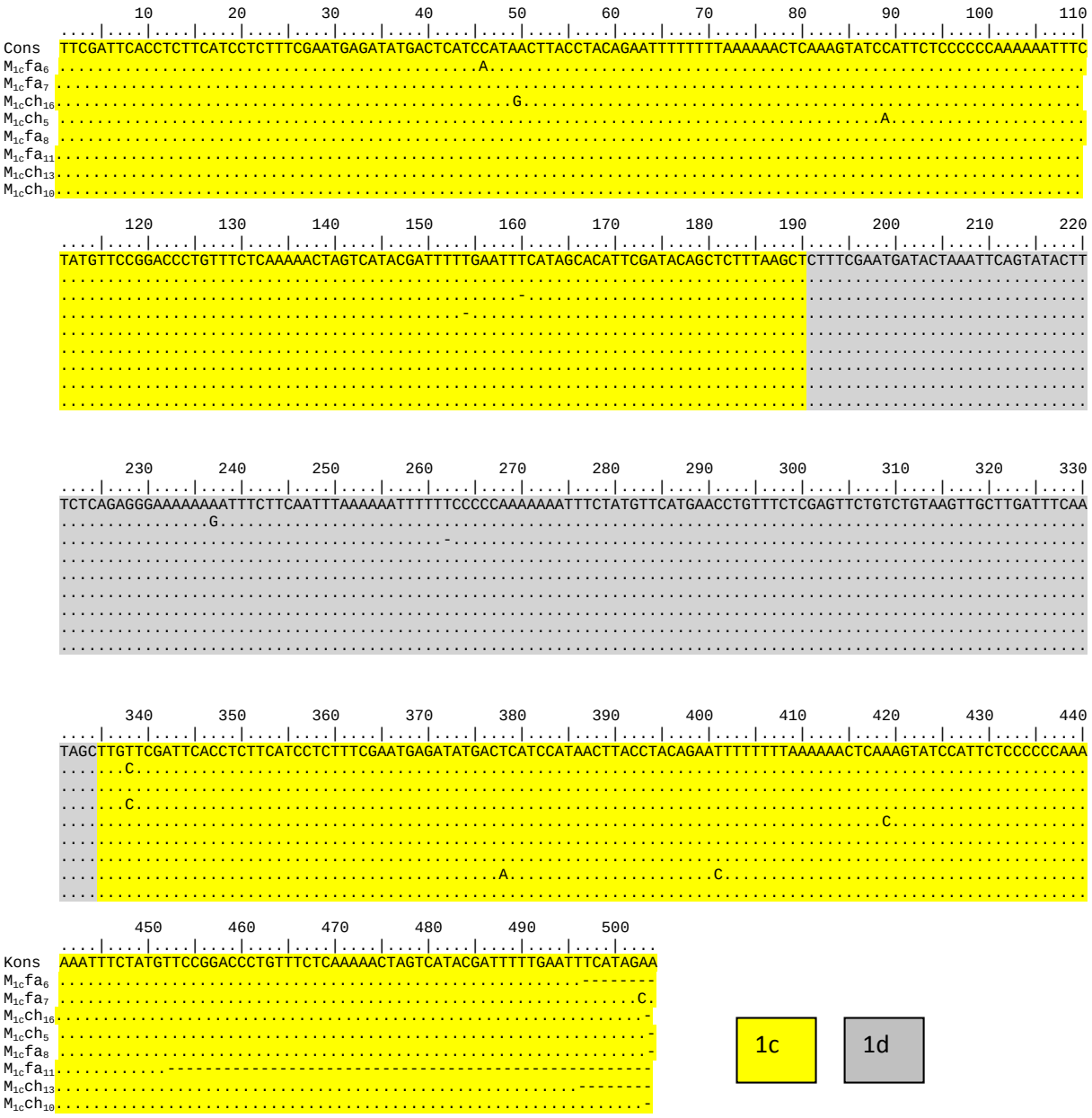
(*) Common European Framework of Reference (CEF) level

PARTICIPATION IN SCIENTIFIC AND SPECIALIZED PROJECTS	Adris Foundation “Development of DNA markers for identifying commercially important mollusk species Ruditapes decussatus in the Adriatic Sea” (team member), 2013-2014. Ministry of Science, Education and Sports “Evolution, properties and functional interactions of satellite DNA sequences” (team member), 2009-2013. "Biodiversity of subterranean fauna of Karlovac County" financed by European Union through PHARE 2006 Program, 2009. "Conservation of Eunapius subterraneus, the only subterranean freshwater sponge in the world", 2003-2008. "KEC – Karst Ecosystem Conservation Project", 2003-2007. "Dinaric Alps rare habitats and species conservation project Croatia", 2003-2006. "Conservation of the Croatian subterranean fauna through inventarisation, mapping, education and popularisation", 2000. – 2010. "Subterranean conservation of the lost cave systems of the Dinaric Arc", 2011-2012
PARTICIPATION IN SCIENTIFIC MEETINGS	2014. The 28th European Congress of Arachnology, Torino, Italy, poster presentation: Katušić L., Ozimec R., Pavlek M., Majer M., Drakšić M., Čukušić A. & Kolundžić E.: A lot or not? A review of 200 years of spider research in Croatia 2012. 21st International Conference on Subterranean Biology, Košice, Slovakia, poster presentation: Ćukušić A, Pavlek M, Ozimec R: Diversity and distribution of cave dwelling spiders of the families Nesticidae and Agelenidae (Araneae) in Croatia 2011. SIEEC2222.- Symposium internationale entomofaunisticum Europae centralis, Varaždin, oral presentation: M. Pavlek, M. Plohl, N. Meštrović: Highly repeated non-coding DNA in Tribolium castaneum (Coleoptera) genome 2011. SIEEC2222.- Symposium internationale entomofaunisticum Europae centralis, Varaždin, poster presentation: Ćukušić, Anđela; Pavlek, Martina. Faunistics, ecology and biogeography of the cave-dwelling spiders of the families Nesticidae and Agelenidae (Araneae) in Croatia. 2011. 18th International Chromosome Conference, Manchester, UK, poster presentation: Pavlek, M., Plohl, M., Mestrovic, N.: Highly repeated non-coding DNA in Tribolium castaneum (Coleoptera) genome 2010. Society for Experimental Biology Annual Main Meeting; oral presentation: Meštrović, N., Castagnone-Sereno, P., Pavlek, M., Car, A., Plohl, M.: How satellite DNAs in the “library” are created? 2010. Nemagenics: Exploiting genomics to understand plant-nematode interactions; poster presentation: Meštrović, N., Castagnone-Sereno, P., Pavlek, M., Car, A., Plohl, M.: Complex organization of satellite DNA library in the root-knot nematodes Meloidogyne chitwoodi and M. fallax. 2009. First student Phd symposium ″The Architecture of Life″; poster presentation: Meštrović, N., Pavlek, M., Žižek, M., Plohl, M.: Analysis of satellite DANs in the sequenced genome of Tribolium castaneum (Coleoptera) 2009. 10th Croatian Biological Congress; poster presentation: Meštrović, N., Pavlek, M., Žižek, M., Plohl, M.: Analysis of satellite DANs in the sequenced genome of Tribolium castaneum (Coleoptera) 2009. Embo young scientists forum; poster presentation: Meštrović, N., Pavlek, M., Žižek, M., Plohl, M.: Analysis of satellite DANs in the sequenced genome of Tribolium castaneum (Coleoptera) 2008. 50 years of molecular biology in Croatia; poster presentation: Meštrović, N., Pavlek, M., Žižek, M., Plohl, M.: Analysis of satellite DANs in the sequenced genome of Tribolium castaneum (Coleoptera) 2008. ESF Exploratory Workshop on Heterochromatin structure and function from repetitive DNA sequences to epigenetics

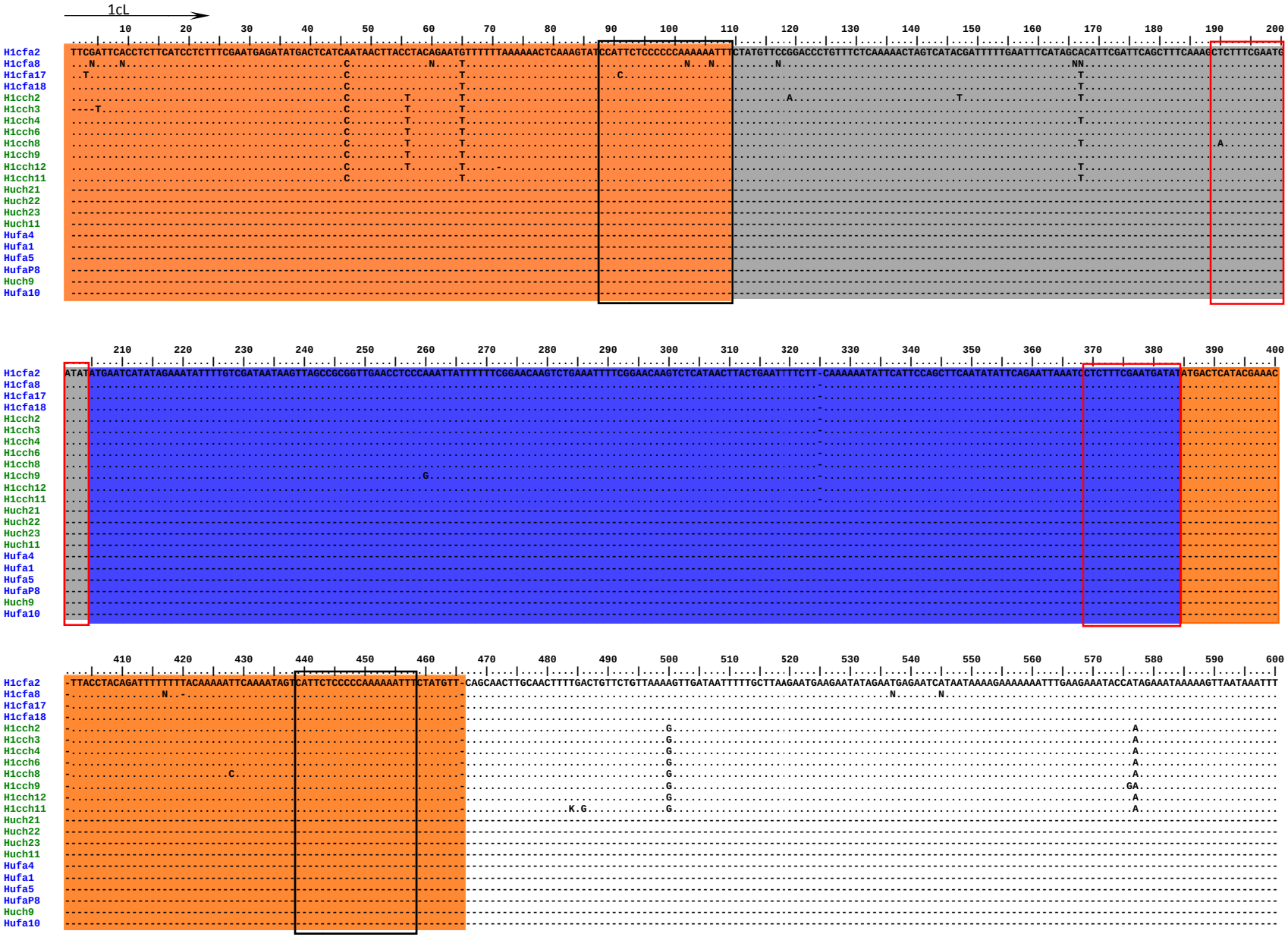
TECHNICAL SKILLS AND COMPETENCIES	2014. 3rd Workshop on the Application of Next Generation Sequencing to Repetitive DNA Analysis in Plants, České Budějovice, Czech Republic 2011. Methodological Course in Molecular Biology and Medicine „Molecular Phylogeny“, Zagreb, 2011. One month training at INRA institute in Antibes, France 2009. Practical Course „Introduction to Bioinformatics“, Zagreb Expirience in field work, specially in research of caves and underground fauna, specialist for cave spiders taxonomy Nucleic acids characterization; electrophoretic techniques; hybridization techniques (Southern), molecular cloning, library screening, PCR techniques, molecular cytogenetic techniques (multicolored FISH, PRINS), computer analyses of DNA sequences (phylogenetic analyses, structure predictions, sequence assembly)
PUBLICATIONS	Meštrović, Pavlek, Car, Castagnone-Sereno, Abad, Plohl (2013) Conserved DNA Motifs, Including the CENP-B Box-like, Are Possible Promoters of Satellite DNA Array Rearrangements in Nematodes. PLoS ONE. 8(6): e6732 Jalžić, Branko; Bedek, Jana; Bilandžija, Helena; Bregović, Petra; Cvitanović, Hrvoje; Čuković, Tamara; Ćukušić, Anđela; Dražina, Tvrtko; Đud, Lana; Gottstein, Sanja; Hmura, Dajana; Kljaković-Gašpić, Fanica; Komerički, Ana; Kutleša, Petra; Lukić, Marko; Malenica, Marta; Miculinić, Kazimir; Ozimec, Roman; Pavlek, Martina; Raguž, Nikolina; Slapnik, Rajko; Štamol, Vesna. 2013: The cave type localities Atlas of Croatian fauna, Volume 2. Zagreb. Marguš, Drago; Barišić, Teo; Bedek, Jana; Dražina, Tvrtko; Gracin, Joso; Hamidović, Daniela; Jalžić, Branko; Komerički, Ana; Lukić, Marko; Marguš, Marija; Menđušić, M.; Miculinić, Kazimir; Mihelčić, G.; Ozimec, Roman; Pavlek, Martina. 2012: Tajne podzemlja. Šibenik. Jalžić, B.; Bedek, J., Bilandžija, H.; Cvitanović, H.; Dražina, T.; Gottstein, S.; Kljaković Gašpić, F.; Lukić, M.; Ozimec, R.; Pavlek, M.; Slapnik, R.; Štamol, V. 2010: The cave type localities Atlas of Croatian fauna, Volume 1. Zagreb. Ozimec, R.; Bedek, J.; Gottstein, S.; Jalžić, B.; Slapnik, R.; Štamol, V.; Bilandžija, H.; Dražina, T.; Kletečki, E.; Komerički, A.; Lukić, M.; Pavlek, M. 2009: Red book of cave fauna of Croatia. Zagreb: Ministry of Culture of the Republic of Croatia, The State Institute for Nature Protection. Pavlek, M. et Ozimec, R. 2009: New cave-dwelling species of genus Troglohyphantes (Araneae, Linyphiidae) for Croatian fauna. Natura Croatica 18 (1), 29-37.
CURRENT RESEARCH INTEREST	Mapping and analyzes of satellite DNAs in *Tribolium castaneum* (Coleoptera, Tenebrionidae) and root-knot nematodes of the genus *Meloidogine* (Nematode). Biology, ecology, phylogeny, taxonomy and evolution of Dinaric cave spiders. Research and protection of cave fauna and cave habitats in general.
MEMBERSHIP IN SOCIETIES	2003 on Croatian Biospeleological Society 2003 on Croatian Mountaineering Society “Željezničar”, Department of Speleology 2010 on Society for Experimental Biology (London) 2014 on European Society of Arachnology

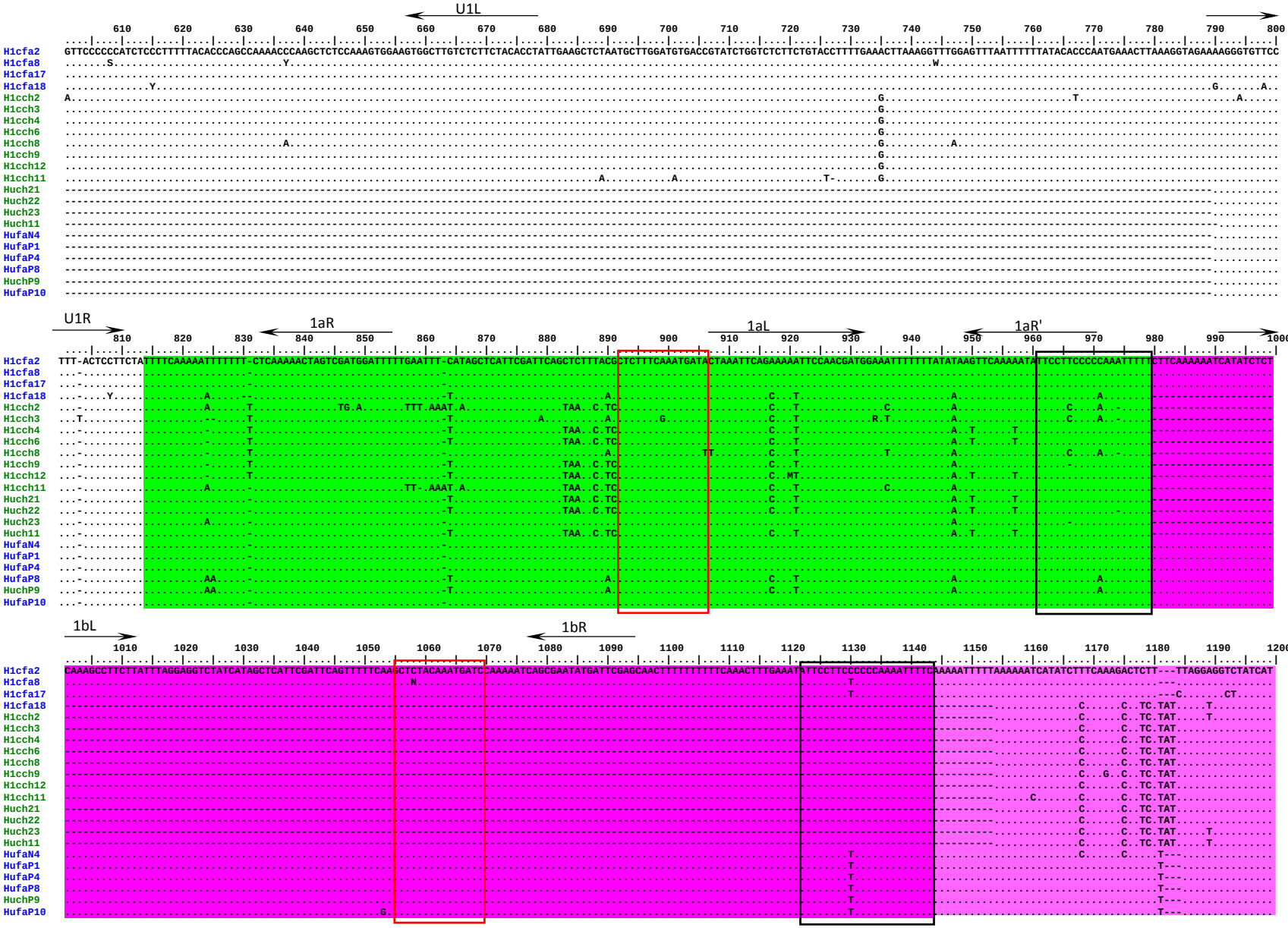
11. SUPPLEMENTARY MATERIAL

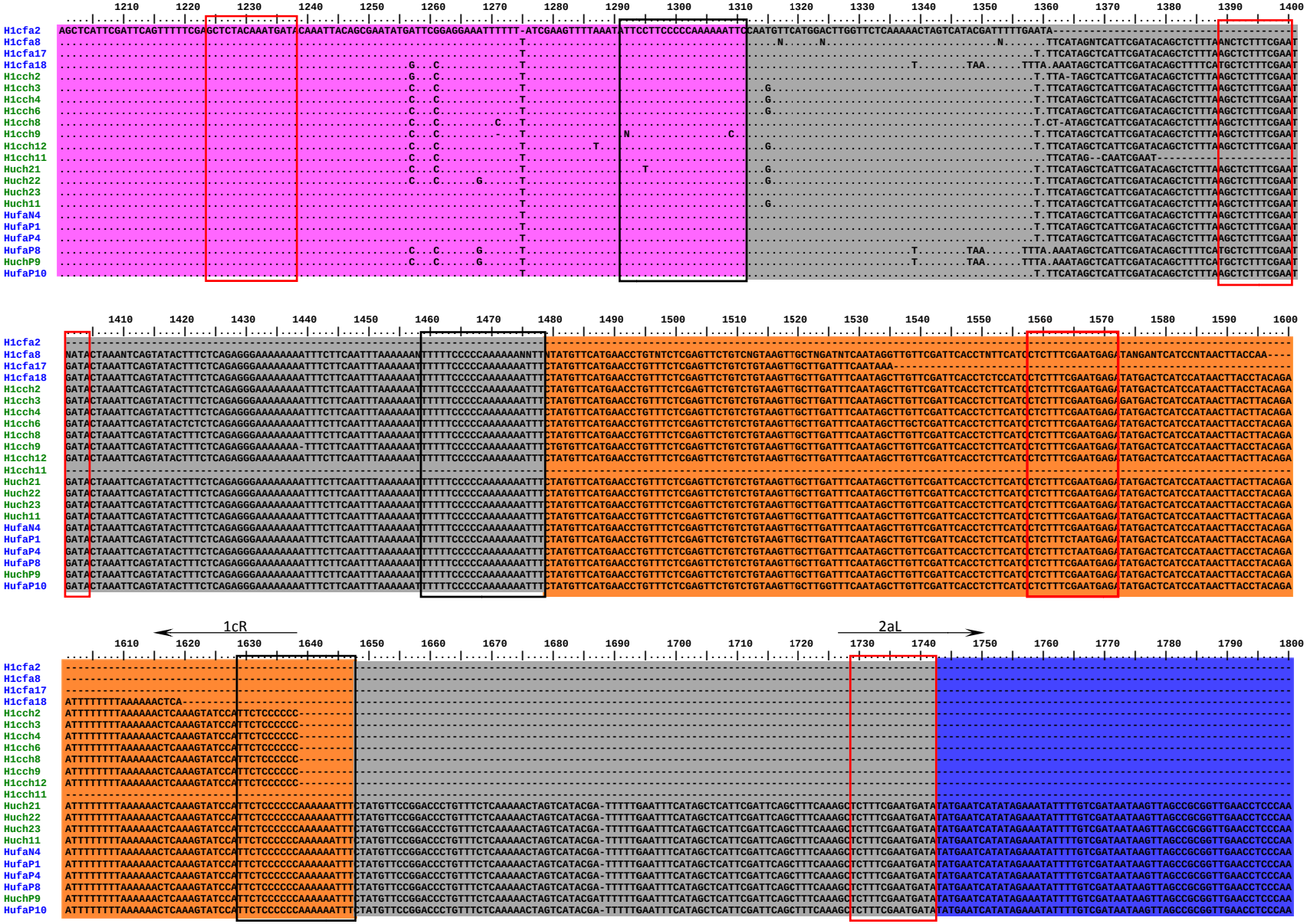
Supplementary Figure 4.1.1. Alignment of nucleotide sequences obtained by random cloning of trimer after PCR amplification of *M. fallax* and *M. chitwoodi* genome with primers specific for 1c satellite DNA. Satellite trimers from *M. fallax* are marked fa(n) and from *M. chitwoodi* ch(n). All trimers are compared to the first sequence. Positions identical to the first sequence are shown with dot and deletions are indicated with dash.

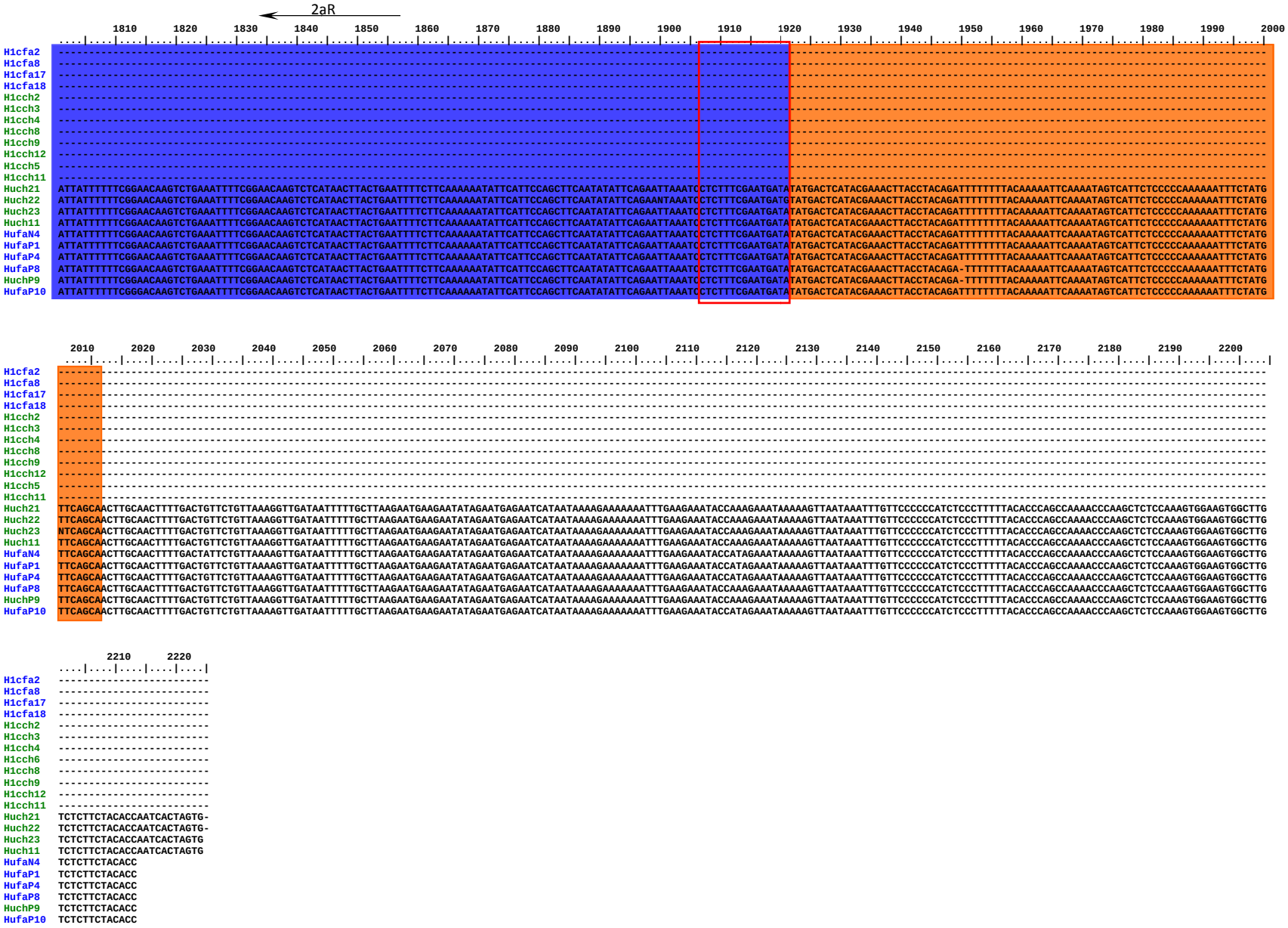


Supplementary Fig. 4.1.2. Alignment of HORs from *M. fallax* (clone names in blue) and *M. chitwoodi* (clone names in green). H1cfa(n) and H1cch(n) represent fragments amplified with 1c primers. Hufa(n) and Huch(n) are amplified with primers specific for U1 sequence. All primer positions are marked above sequences and primers are listed in Table 3.6. SatDNA monomers are indicated in different colors; 1c, 1d, 2a, 1a, 1b and 1b'. Unlabeled part of the HOR is U1 sequence. Red boxes indicate Box 1, and black boxes represent Box 2.

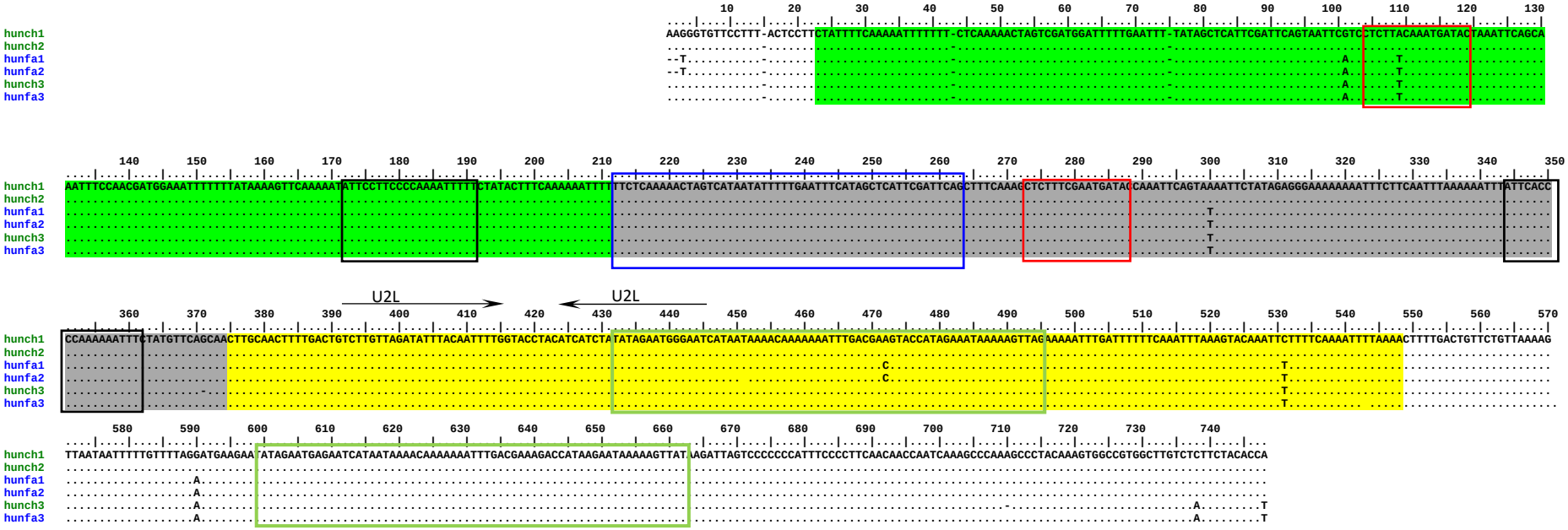

1cL
H1cfa2
H1cfa8
H1cfa17
H1cfa18
H1cch2
H1cch3
H1cch4
H1cch6
H1cch8
H1cch9
H1cch12
H1cch11
Huch21
Huch22
Huch23
Huch11
Hufa4
Hufa1
Hufa5
HufaP8
Huch9
Hufa10


U1L
U1R
1aR
1aL
1aR'
1bL
1bR
H1cfa2
H1cfa8
H1cfa17
H1cfa18
H1cch2
H1cch3
H1cch4
H1cch6
H1cch8
H1cch9
H1cch12
H1cch11
Huch21
Huch22
Huch23
Huch11
HufaN4
HufaP1
HufaP4
HufaP8
HuchP9
HufaP10

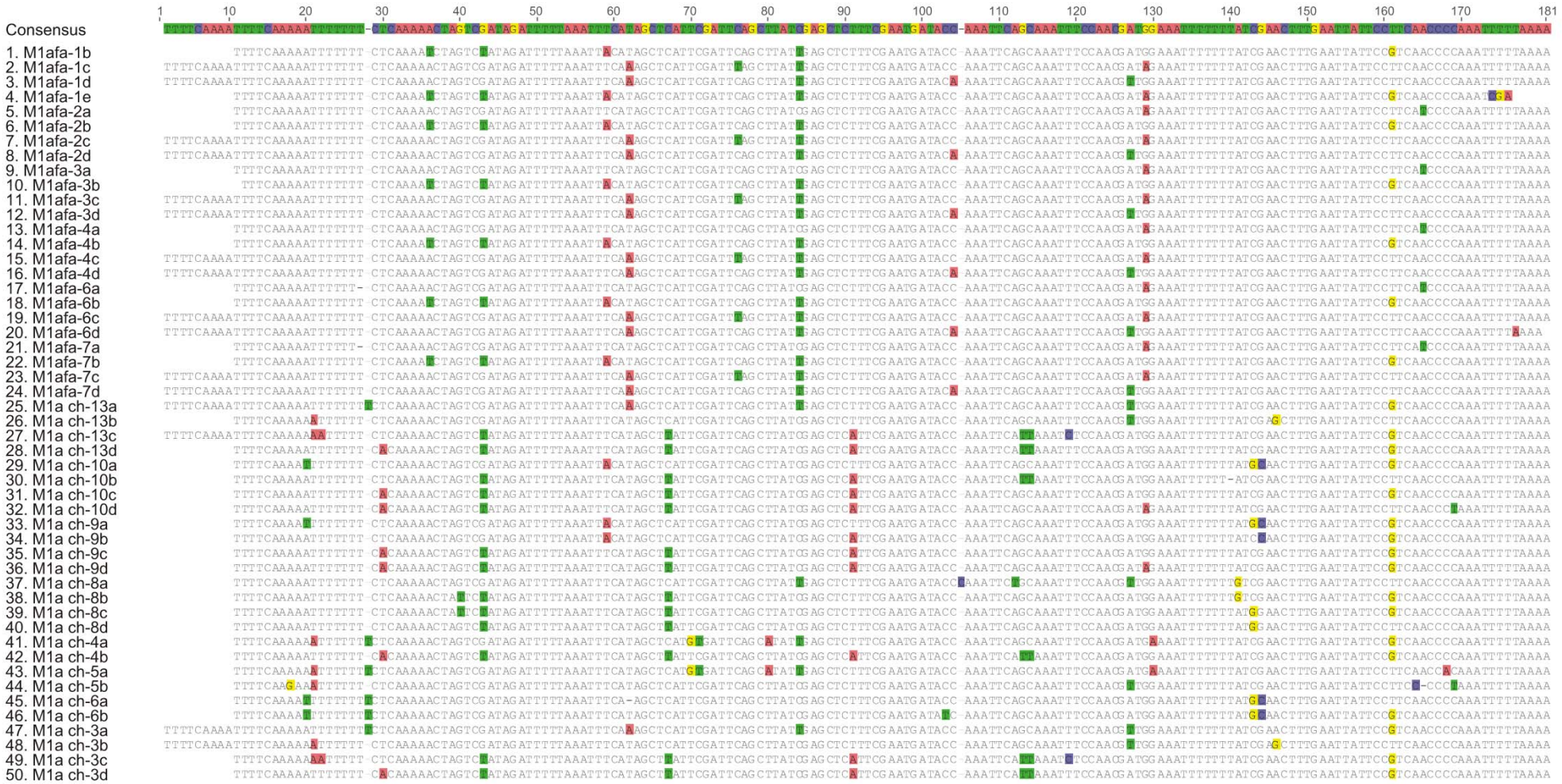

1cR
2aL


2aR
H1cfa2
H1cfa8
H1cfa17
H1cfa18
H1cch2
H1cch3
H1cch4
H1cch8
H1cch9
H1cch12
H1cch5
H1cch11
Huch21
Huch22
Huch23
Huch11
HufaN4
HufaP1
HufaP4
HufaP8
HuchP9
HufaP10

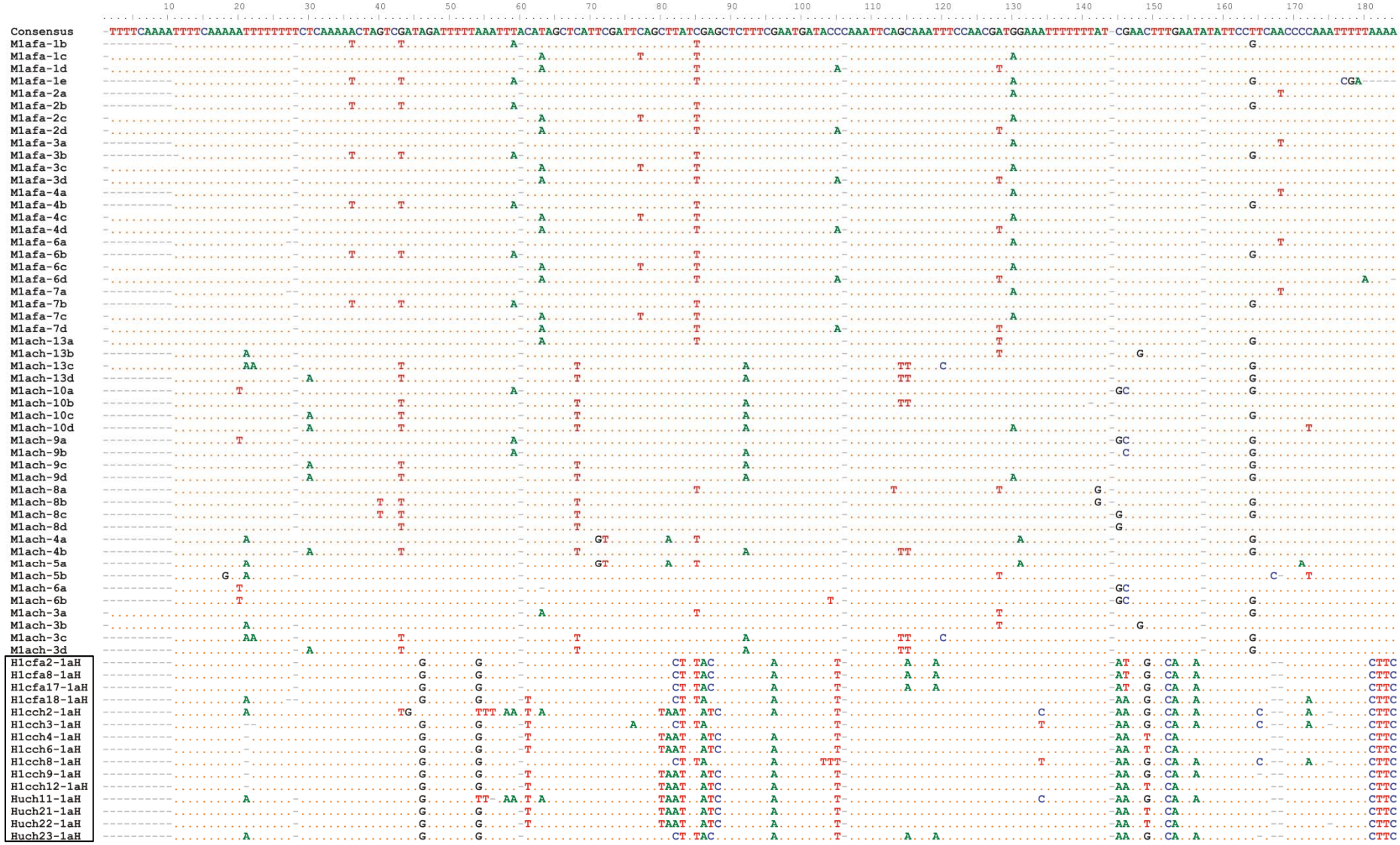
Supplementary Fig. 4.1.3. Alignment of complex fragments from *M. fallax* (clone names in blue) and *M. chitwoodi* (clone names in green) amplified with primers specific for U1 sequence. Sequences are indicated in different colors: 1a monomer (green), 1d monomer (grey) and U2 sequence (yellow). Unlabeled part belongs to U1 sequence. Blue box represents overlapping region of 1a and 1d monomers. Box 1 is indicated in red and Box 2 in black square. Green boxes represent perfectly conserved fragment common for U1 and U2 sequences. Primer positions for U2 are indicated above sequences.



Supplementary Fig. 4.1.4. Alignment of 14 1aM sequences obtained cut to monomers.

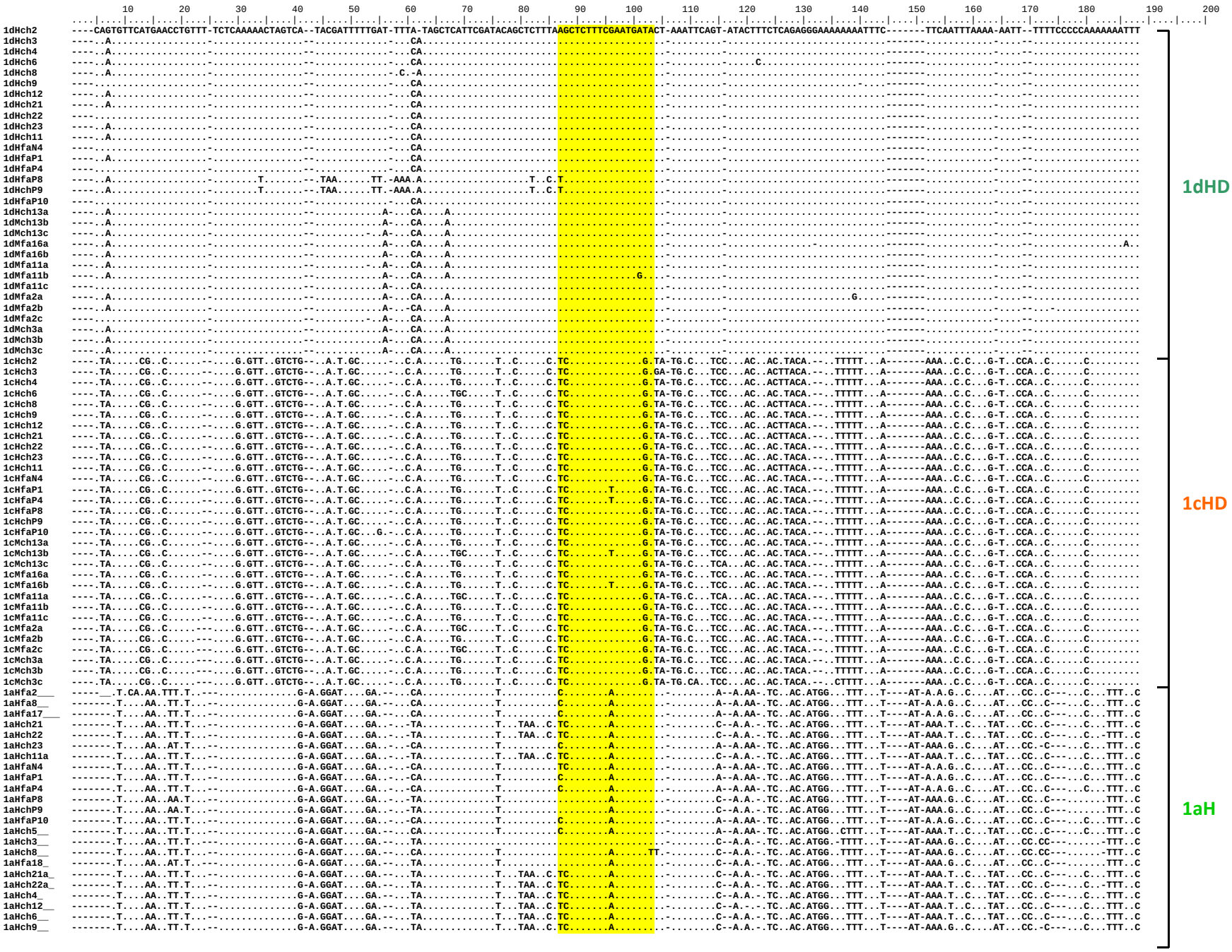


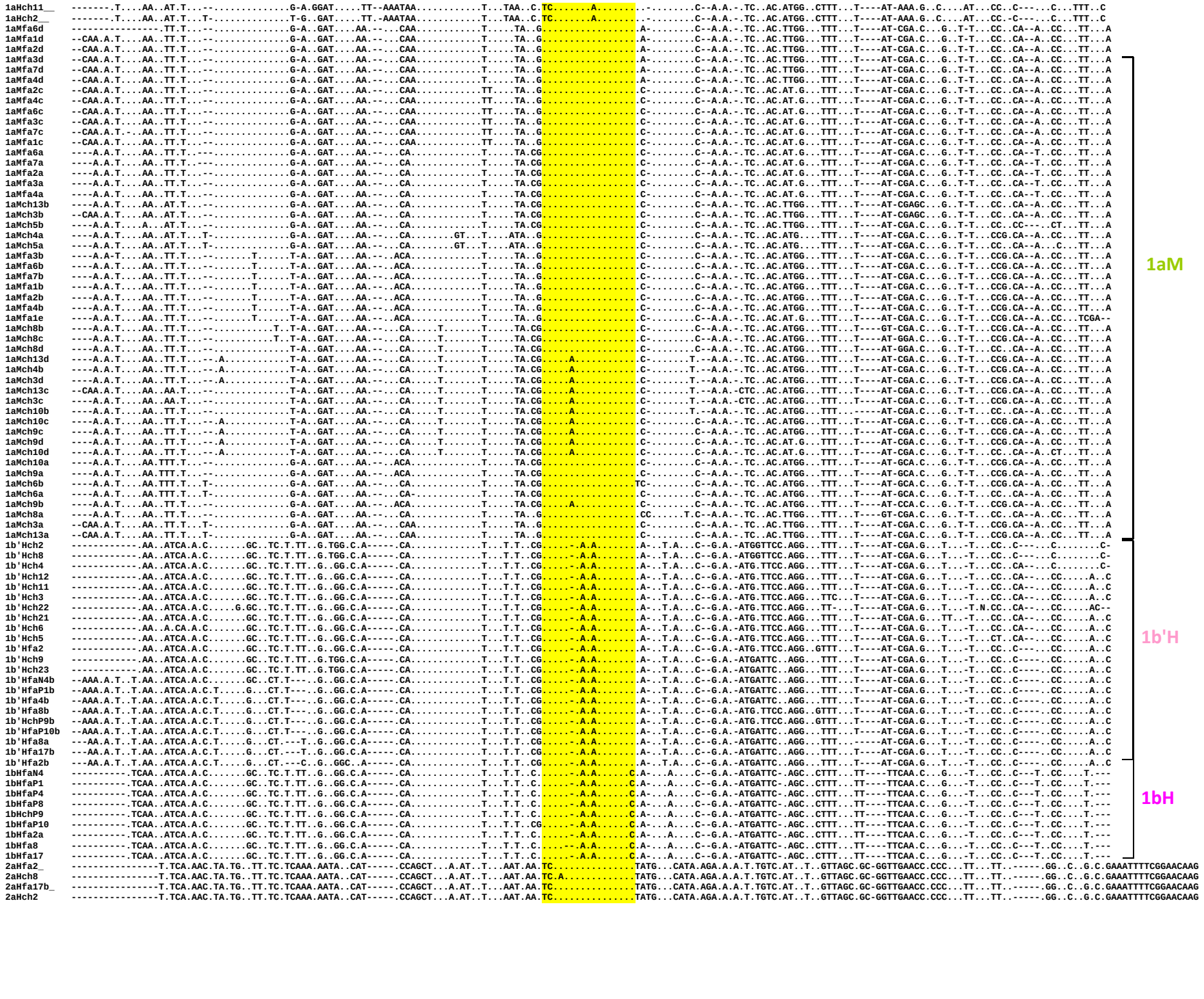
Supplementary Fig. 4.1.5. Alignment of 1aM and 1aH sequences. 1aH marked with black box.

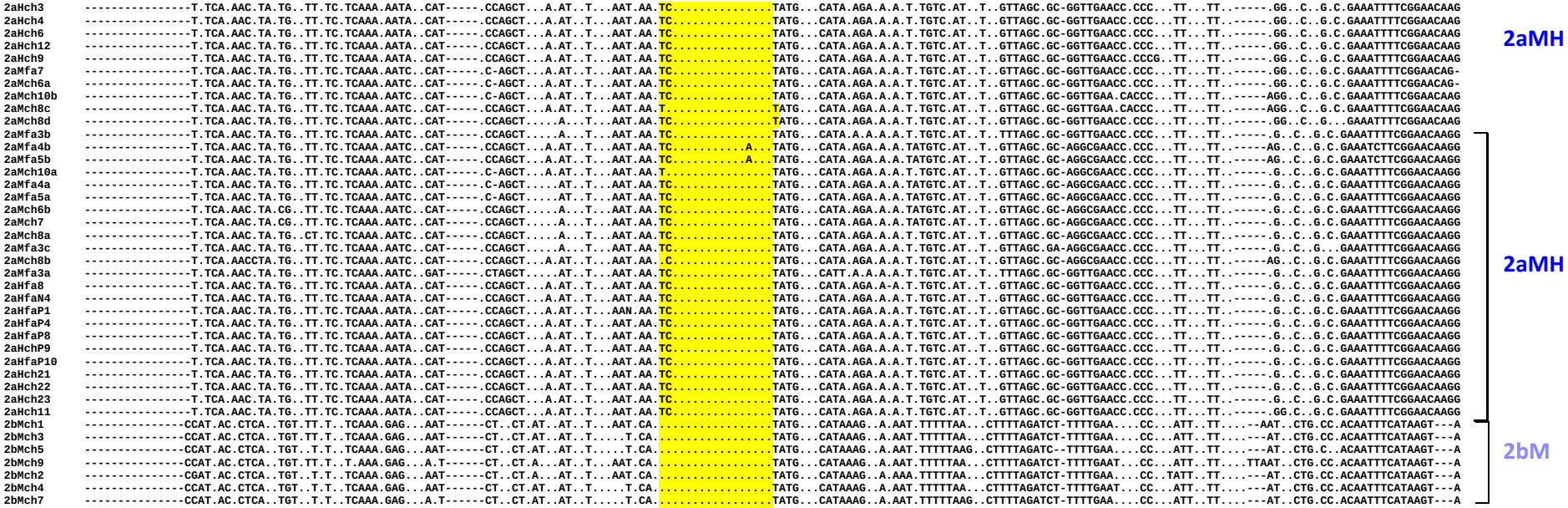


Supplementary Fig. 4.1.6. Alignment of 1a, 1b, 1b', 1c, 1d, 2a and 2b monomers from *M. fallax* and *M. chitwoodi*. Monomers are extracted from monomeric and HOR arrays using KSA algorithm (Rosandić et al. 2003). All monomers are compared with first sequence and positions identical to the first sequence are shown with dot. Monomer group are indicated on the right side. Box 1 is shaded with yellow. Detail descriptions of satellite monomers are indicated below:

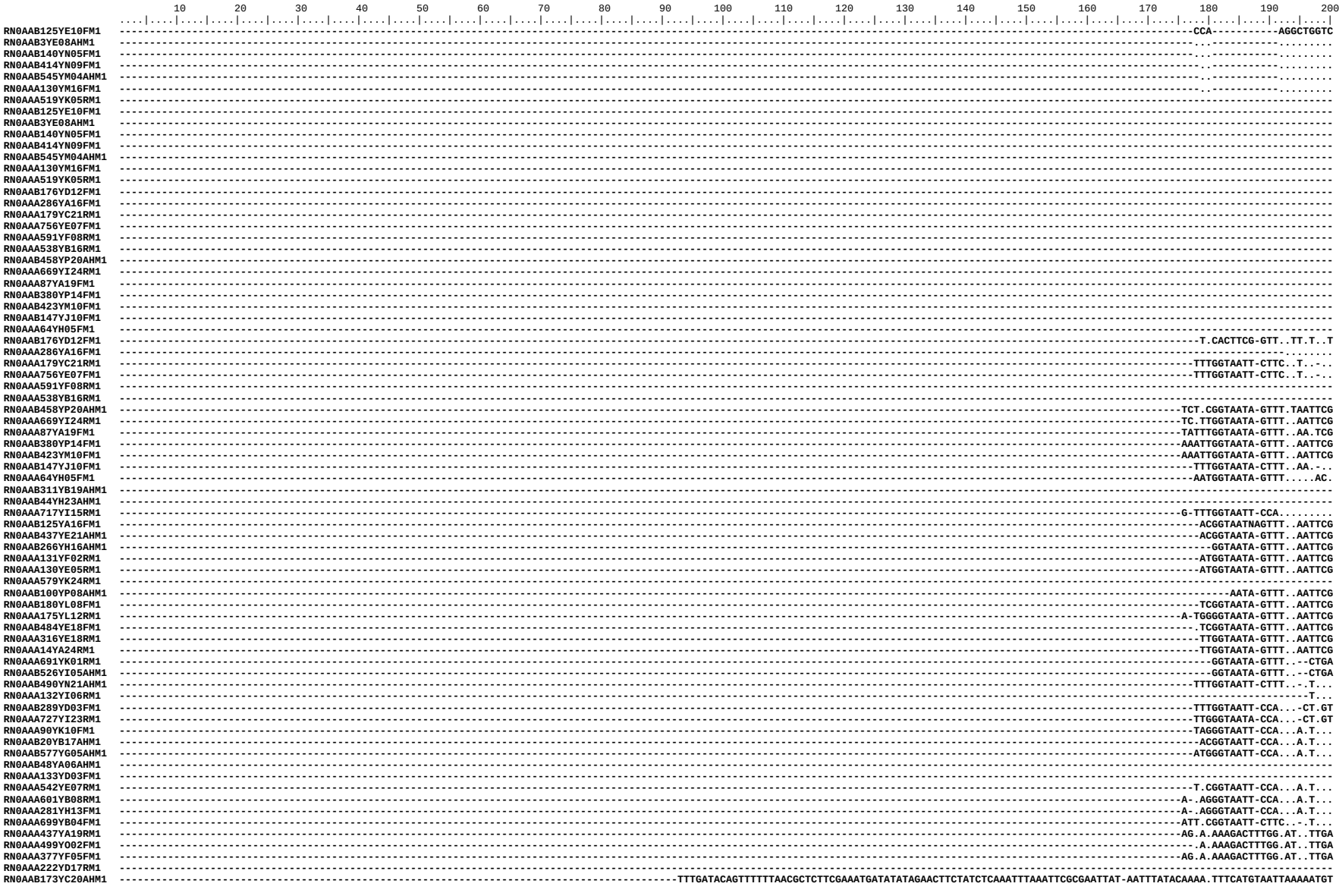
1cHch$_n$-1c monomers from chitwoodi HOR arrays ($H_{1c}ch_n$, H_uch_n)
1cHfa$_n$ -1c monomers from fallax HOR arrays ($H_{1c}fa_n$, H_ufa_n)
1cMch$_n$-1c monomers from chitwoodi monomeric arrays ($M_{1c}ch_n$)
1cMfa$_n$ -1c monomers from fallax monomeric arrays ($M_{1c}fa_n$)
1dHch$_n$-1d monomers from chitwoodi HOR arrays ($H_{1c}ch_n$, H_uch_n)
1dHfa$_n$ -1d monomers from fallax HOR arrays ($H_{1c}fa_n$, H_ufa_n)
1dMch$_n$-1d monomers from chitwoodi monomeric arrays ($M_{1c}ch_n$)
1dMfa$_n$ -1d monomers from fallax monomeric arrays ($M_{1c}fa_n$)
1aHch$_n$-1a monomers from chitwoodi HOR arrays ($H_{1c}ch_n$, H_uch_n)
1aHfa$_n$ -1a monomers from fallax HOR arrays ($H_{1c}fa_n$, H_ufa_n)
1aMch$_n$-1a monomers from chitwoodi monomeric arrays ($M_{1a}ch_n$)
1aMfa$_n$ -1a monomers from fallax monomeric arrays ($M_{1a}fa_n$)
1bHch$_n$-1b monomers from chitwoodi HOR arrays ($H_{1c}ch_n$, H_uch_n)
1bHfa$_n$ -1b monomers from fallax HOR arrays ($H_{1c}fa_n$, H_ufa_n)
1b'Hch$_n$-1b monomers from chitwoodi HOR arrays ($H_{1c}ch_n$, H_uch_n)
1b'Hfa$_n$ -1b monomers from fallax HOR arrays ($H_{1c}fa_n$, H_ufa_n)
2aHch$_n$-2a monomers from chitwoodi HOR arrays ($H_{1c}ch_n$, H_uch_n)
2aHfa$_n$ -2a monomers from fallax HOR arrays ($H_{1c}fa_n$, H_ufa_n)
2aMch$_n$-2a monomers from chitwoodi monomeric arrays ($M_{2a}ch_n$)
2aMfa$_n$ -2a monomers from fallax monomeric arrays ($M_{2a}fa_n$)
2bMch$_n$-2b monomers from chitwoodi monomeric arrays ($M_{2b}ch_n$)

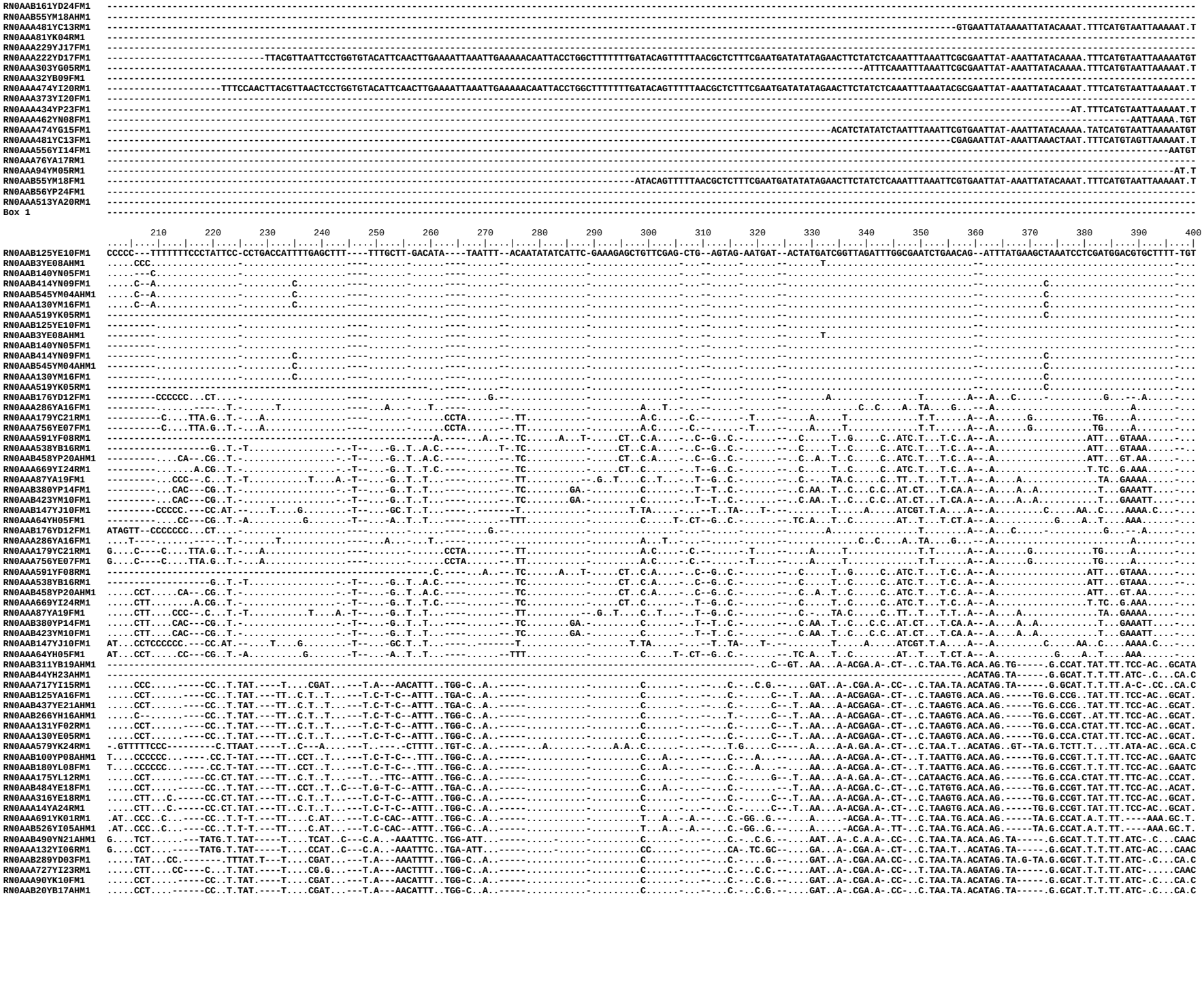

Box 1
1dHD
1cHD
1aH

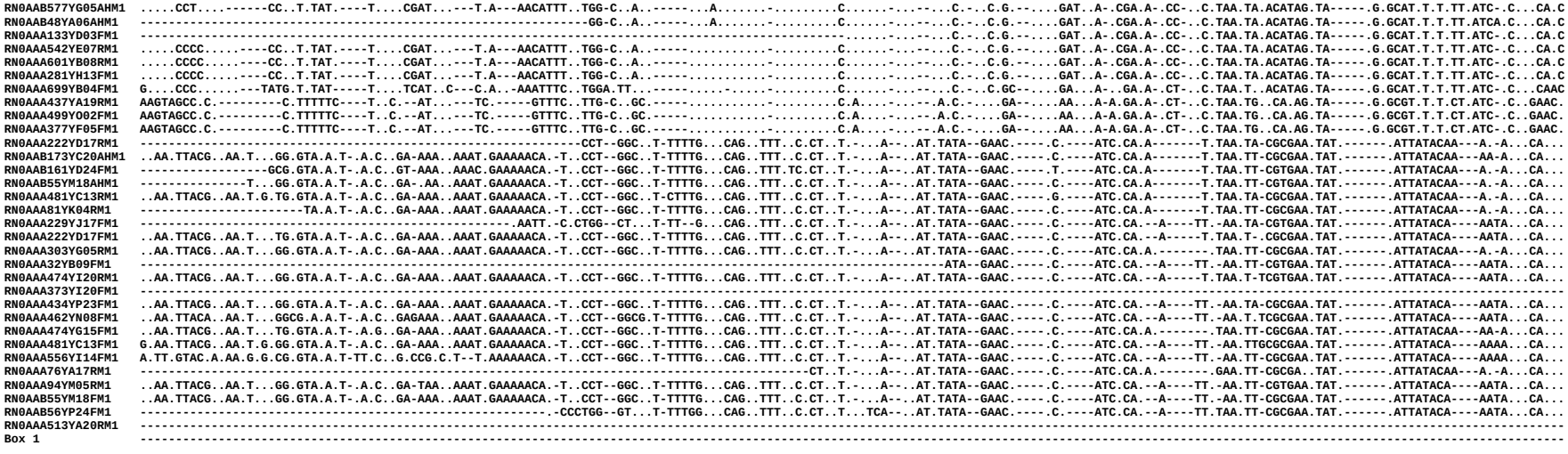

1aM
1b'H
1bH

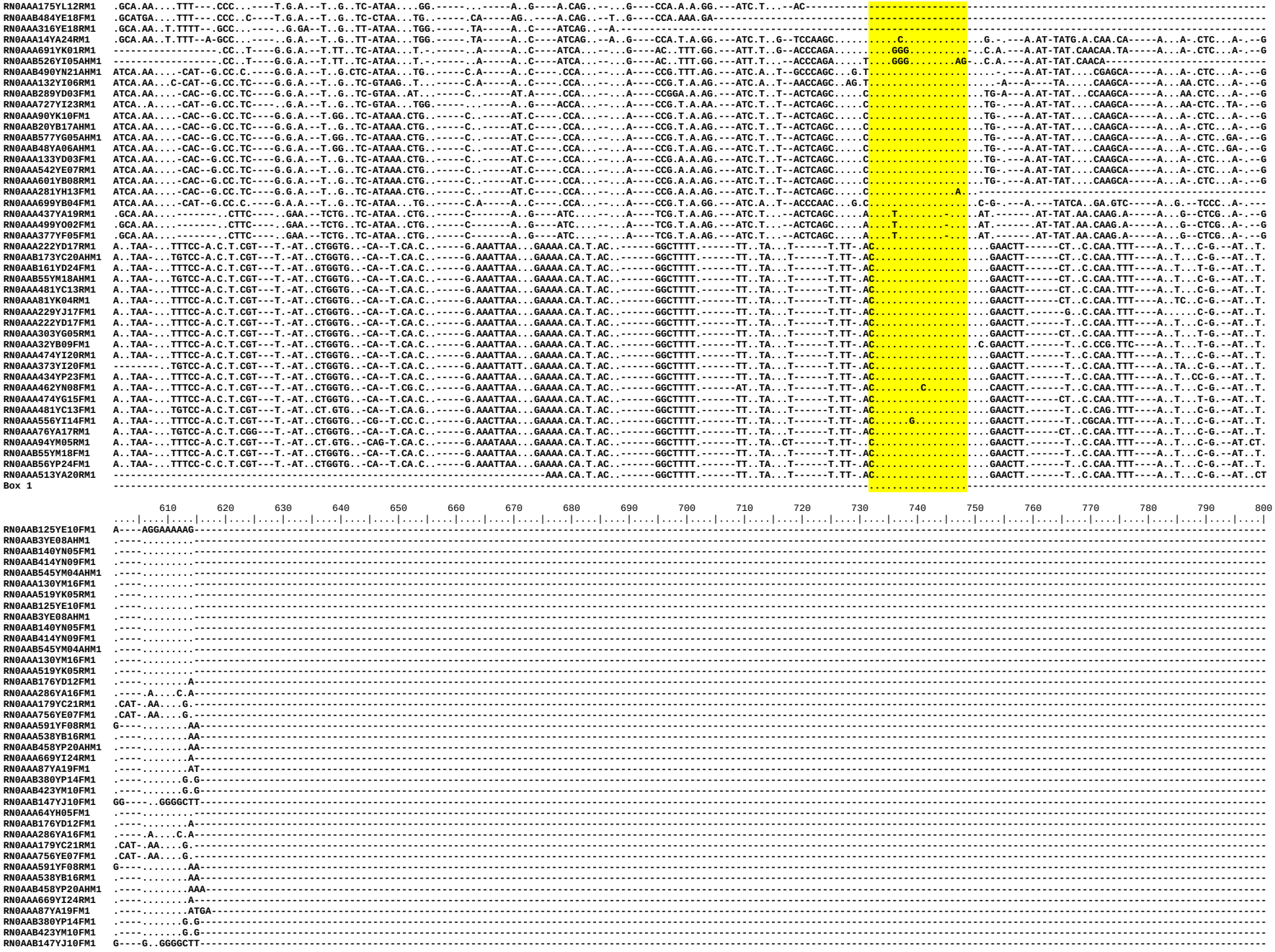

2aMH
2aMH
2bM

Supplementary Fig 4.1.7. Alignment of Box 1-containing sequences from unassembled reads of *M. incogita* sequenced genome. Sequences are compared with first sequence and positions identical to the first sequence are shown with dot. Box 1 is shaded with yellow.




RN0AAB161YD24FM1
RN0AAB55YM18AHM1
RN0AAA481YC13RM1
RN0AAA81YK04RM1
RN0AAA229YJ17FM1
RN0AAA222YD17FM1
RN0AAA303YG05RM1
RN0AAA32YB09FM1
RN0AAA474YI20RM1
RN0AAA373YI20FM1
RN0AAA434YP23FM1
RN0AAA462YN08FM1
RN0AAA474YG15FM1
RN0AAA481YC13FM1
RN0AAA556YI14FM1
RN0AAA76YA17RM1
RN0AAA94YM05RM1
RN0AAB55YM18FM1
RN0AAB56YP24FM1
RN0AAA513YA20RM1
Box 1
210
220
230
240
250
260
270
280
290
300
310
320
330
340
350
360
370
380
390
400
RN0AAB125YE10FM1
RN0AAB3YE08AHM1
RN0AAB140YN05FM1
RN0AAB414YN09FM1
RN0AAB545YM04AHM1
RN0AAA130YM16FM1
RN0AAA519YK05RM1
RN0AAB125YE10FM1
RN0AAB3YE08AHM1
RN0AAB140YN05FM1
RN0AAB414YN09FM1
RN0AAB545YM04AHM1
RN0AAA130YM16FM1
RN0AAA519YK05RM1
RN0AAB176YD12FM1
RN0AAA286YA16FM1
RN0AAA179YC21RM1
RN0AAA756YE07FM1
RN0AAA591YF08RM1
RN0AAA538YB16RM1
RN0AAB458YP20AHM1
RN0AAA669YI24RM1
RN0AAA87YA19FM1
RN0AAB380YP14FM1
RN0AAB423YM10FM1
RN0AAB147YJ10FM1
RN0AAA64YH05FM1
RN0AAB176YD12FM1
RN0AAA286YA16FM1
RN0AAA179YC21RM1
RN0AAA756YE07FM1
RN0AAA591YF08RM1
RN0AAA538YB16RM1
RN0AAB458YP20AHM1
RN0AAA669YI24RM1
RN0AAA87YA19FM1
RN0AAB380YP14FM1
RN0AAB423YM10FM1
RN0AAB147YJ10FM1
RN0AAA64YH05FM1
RN0AAB311YB19AHM1
RN0AAB44YH23AHM1
RN0AAA717YI15RM1
RN0AAB125YA16FM1
RN0AAB437YE21AHM1
RN0AAB266YH16AHM1
RN0AAA131YF02RM1
RN0AAA130YE05RM1
RN0AAA579YK24RM1
RN0AAB100YP08AHM1
RN0AAB180YL08FM1
RN0AAA175YL12RM1
RN0AAB484YE18FM1
RN0AAA316YE18RM1
RN0AAA14YA24RM1
RN0AAA691YK01RM1
RN0AAB526YI05AHM1
RN0AAB490YN21AHM1
RN0AAA132YI06RM1
RN0AAB289YD03FM1
RN0AAA727YI23RM1
RN0AAA90YK10FM1
RN0AAB20YB17AHM1


RN0AAB577YG05AHM1
RN0AAB48YA06AHM1
RN0AAA133YD03FM1
RN0AAA542YE07RM1
RN0AAA601YB08RM1
RN0AAA281YH13FM1
RN0AAA699YB04FM1
RN0AAA437YA19RM1
RN0AAA499YO02FM1
RN0AAA377YF05FM1
RN0AAA222YD17RM1
RN0AAB173YC20AHM1
RN0AAB161YD24FM1
RN0AAB55YM18AHM1
RN0AAA481YC13RM1
RN0AAA81YK04RM1
RN0AAA229YJ17FM1
RN0AAA222YD17FM1
RN0AAA303YG05RM1
RN0AAA32YB09FM1
RN0AAA474YI20RM1
RN0AAA373YI20FM1
RN0AAA434YP23FM1
RN0AAA462YN08FM1
RN0AAA474YG15FM1
RN0AAA481YC13FM1
RN0AAA556YI14FM1
RN0AAA76YA17RM1
RN0AAA94YM05RM1
RN0AAB55YM18FM1
RN0AAB56YP24FM1
RN0AAA513YA20RM1
Box 1
410
420
430
440
450
460
470
480
490
500
510
520
530
540
550
560
570
580
590
600
RN0AAB125YE10FM1
RN0AAB3YE08AHM1
RN0AAB140YN05FM1
RN0AAB414YN09FM1
RN0AAB545YM04AHM1
RN0AAA130YM16FM1
RN0AAA519YK05RM1
RN0AAB125YE10FM1
RN0AAB3YE08AHM1
RN0AAB140YN05FM1
RN0AAB414YN09FM1
RN0AAB545YM04AHM1
RN0AAA130YM16FM1
RN0AAA519YK05RM1
RN0AAB176YD12FM1
RN0AAA286YA16FM1
RN0AAA179YC21RM1
RN0AAA756YE07FM1
RN0AAA591YF08RM1
RN0AAA538YB16RM1
RN0AAB458YP20AHM1
RN0AAA669YI24RM1
RN0AAA87YA19FM1
RN0AAB380YP14FM1
RN0AAB423YM10FM1
RN0AAB147YJ10FM1
RN0AAA64YH05FM1
RN0AAB176YD12FM1
RN0AAA286YA16FM1
RN0AAA179YC21RM1
RN0AAA756YE07FM1
RN0AAA591YF08RM1
RN0AAA538YB16RM1
RN0AAB458YP20AHM1
RN0AAA669YI24RM1
RN0AAA87YA19FM1
RN0AAB380YP14FM1
RN0AAB423YM10FM1
RN0AAB147YJ10FM1
RN0AAA64YH05FM1
RN0AAB311YB19AHM1
RN0AAB44YH23AHM1
RN0AAA717YI15RM1
RN0AAB125YA16FM1
RN0AAB437YE21AHM1
RN0AAB266YH16AHM1
RN0AAA131YF02RM1
RN0AAA130YE05RM1
RN0AAA579YK24RM1
RN0AAB100YP08AHM1
RN0AAB180YL08FM1

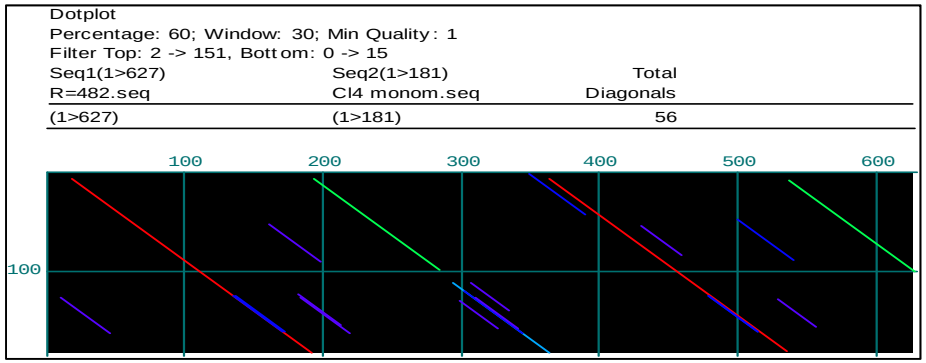

RN0AAA175YL12RM1
RN0AAB484YE18FM1
RN0AAA316YE18RM1
RN0AAA14YA24RM1
RN0AAA691YK01RM1
RN0AAB526YI05AHM1
RN0AAB490YN21AHM1
RN0AAA132YI06RM1
RN0AAB289YD03FM1
RN0AAA727YI23RM1
RN0AAA90YK10FM1
RN0AAB20YB17AHM1
RN0AAB577YG05AHM1
RN0AAB48YA06AHM1
RN0AAA133YD03FM1
RN0AAA542YE07RM1
RN0AAA601YB08RM1
RN0AAA281YH13FM1
RN0AAA699YB04FM1
RN0AAA437YA19RM1
RN0AAA499YO02FM1
RN0AAA377YF05FM1
RN0AAA222YD17RM1
RN0AAB173YC20AHM1
RN0AAB161YD24FM1
RN0AAB55YM18AHM1
RN0AAA481YC13RM1
RN0AAA81YK04RM1
RN0AAA229YJ17FM1
RN0AAA222YD17FM1
RN0AAA303YG05RM1
RN0AAA32YB09FM1
RN0AAA474YI20RM1
RN0AAA373YI20FM1
RN0AAA434YP23FM1
RN0AAA462YN08FM1
RN0AAA474YG15FM1
RN0AAA481YC13FM1
RN0AAA556YI14FM1
RN0AAA76YA17RM1
RN0AAA94YM05RM1
RN0AAB55YM18FM1
RN0AAB56YP24FM1
RN0AAA513YA20RM1
Box 1
610
620
630
640
650
660
670
680
690
700
710
720
730
740
750
760
770
780
790
800
RN0AAB125YE10FM1
RN0AAB3YE08AHM1
RN0AAB140YN05FM1
RN0AAB414YN09FM1
RN0AAB545YM04AHM1
RN0AAA130YM16FM1
RN0AAA519YK05RM1
RN0AAB125YE10FM1
RN0AAB3YE08AHM1
RN0AAB140YN05FM1
RN0AAB414YN09FM1
RN0AAB545YM04AHM1
RN0AAA130YM16FM1
RN0AAA519YK05RM1
RN0AAB176YD12FM1
RN0AAA286YA16FM1
RN0AAA179YC21RM1
RN0AAA756YE07FM1
RN0AAA591YF08RM1
RN0AAA538YB16RM1
RN0AAB458YP20AHM1
RN0AAA669YI24RM1
RN0AAA87YA19FM1
RN0AAB380YP14FM1
RN0AAB423YM10FM1
RN0AAB147YJ10FM1
RN0AAA64YH05FM1
RN0AAB176YD12FM1
RN0AAA286YA16FM1
RN0AAA179YC21RM1
RN0AAA756YE07FM1
RN0AAA591YF08RM1
RN0AAA538YB16RM1
RN0AAB458YP20AHM1
RN0AAA669YI24RM1
RN0AAA87YA19FM1
RN0AAB380YP14FM1
RN0AAB423YM10FM1
RN0AAB147YJ10FM1

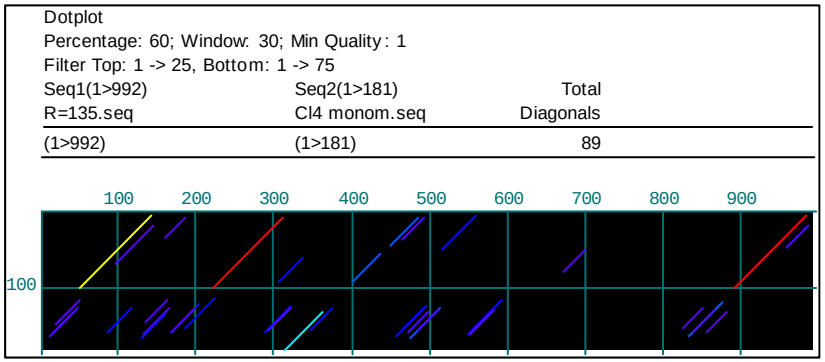

RN0AAA64YH05FM1
RN0AAB311YB19AHM1
RN0AAB44YH23AHM1
RN0AAA717YI15RM1
RN0AAB125YA16FM1
RN0AAB437YE21AHM1
RN0AAB266YH16AHM1
RN0AAA131YF02RM1
RN0AAA130YE05RM1
RN0AAA579YK24RM1
RN0AAB100YP08AHM1
RN0AAB180YL08FM1
RN0AAA175YL12RM1
RN0AAB484YE18FM1
RN0AAA316YE18RM1
RN0AAA14YA24RM1
RN0AAA691YK01RM1
RN0AAB526YI05AHM1
RN0AAB490YN21AHM1
RN0AAA132YI06RM1
RN0AAB289YD03FM1
RN0AAA727YI23RM1
RN0AAA90YK10FM1
RN0AAB20YB17AHM1
RN0AAB577YG05AHM1
RN0AAB48YA06AHM1
RN0AAA133YD03FM1
RN0AAA542YE07RM1
RN0AAA601YB08RM1
RN0AAA281YH13FM1
RN0AAA699YB04FM1
RN0AAA437YA19RM1
RN0AAA499YO02FM1
RN0AAA377YF05FM1
RN0AAA222YD17RM1
RN0AAB173YC20AHM1
RN0AAB161YD24FM1
RN0AAB55YM18AHM1
RN0AAA481YC13RM1
RN0AAA81YK04RM1
RN0AAA229YJ17FM1
RN0AAA222YD17FM1
RN0AAA303YG05RM1
RN0AAA32YB09FM1
RN0AAA474YI20RM1
RN0AAA373YI20FM1
RN0AAA434YP23FM1
RN0AAA462YN08FM1
RN0AAA474YG15FM1
RN0AAA481YC13FM1
RN0AAA556YI14FM1
RN0AAA76YA17RM1
RN0AAA94YM05RM1
RN0AAB55YM18FM1
RN0AAB56YP24FM1
RN0AAA513YA20RM1
Box 1

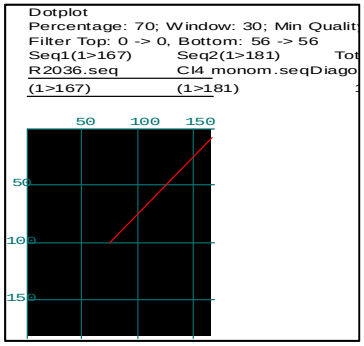
Supplementary Figure 4.2.1. LOCAL BLAST search of newly TRF satDNA families with RepeatScout library repeat families published in Wang et al 2008. **a)** results for search with Cl4 monomer **b**) results for search with Cl5 monomer

a) results for search with Cl4 monomer

R=482|TRF=0.000|NSEG=0.429|HighB	151	7e-038
R=135|TRF=0.000|NSEG=0.444|HighB	151	7e-038
R=0|TRF=0.080|NSEG=0.370|HighB-dispearsed dimer	151	7e-038
R=2036|TRF=0.000|NSEG=0.341|Low-part of monomer+ flanking	143	2e-035
R=1766|TRF=0.000|NSEG=0.306|Mid- part of monomer+flanking	127	1e-030
R=692 (RR=689. TRF=0.000 NSEG=0.388part of monomer+flanking	127	1e-030
R=3699|TRF=0.000|NSEG=0.242|Low	125	4e-030

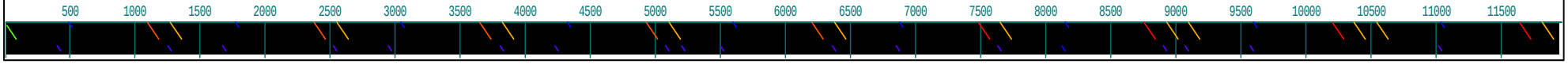






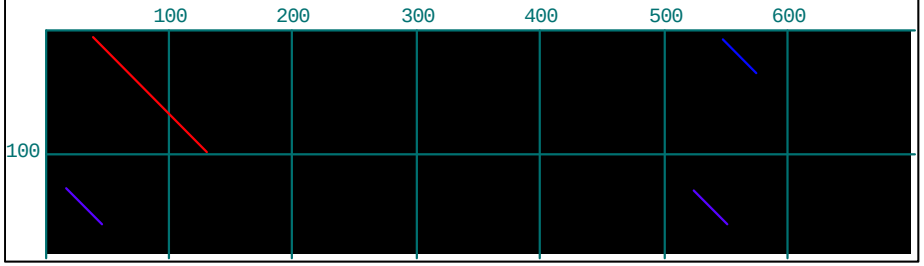
Dotplot
Percentage: 70; Window: 30; Min Quality: 1
Filter Top: 1 -> 73, Bottom: 1 -> 14

Seq1(1>11954)	Seq2(1>181)	Total
R=0.seq	Cl4 monom.seq	Diagonals
(1>11954)	(1>181)	95



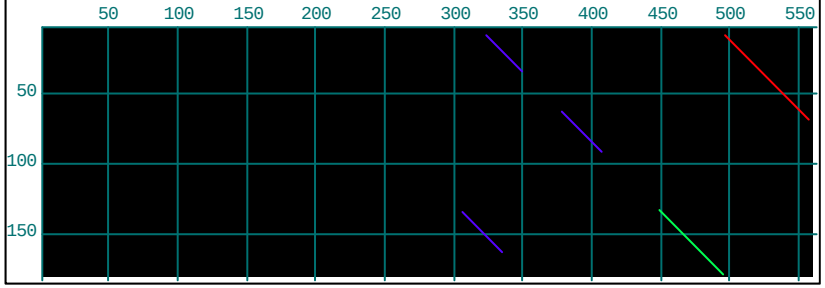
Dotplot
Percentage: 70; Window: 30; Min Quality: 1
Filter Top: 1 -> 62, Bottom: 2 -> 3

Seq1(1>698)	Seq2(1>181)	Total
R=1766.seq	Cl4 monom.seq	Diagonals
(1>698)	(1>181)	6



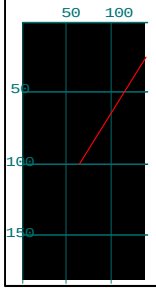
Dotplot
Percentage: 70; Window: 30; Min Quality: 1
Filter Top: 1 -> 41, Bottom: 5 -> 5

Seq1(1>560)	Seq2(1>181)	Total
R=692.seq	Cl4 monom.seq	Diagonals
(1>560)	(1>181)	6



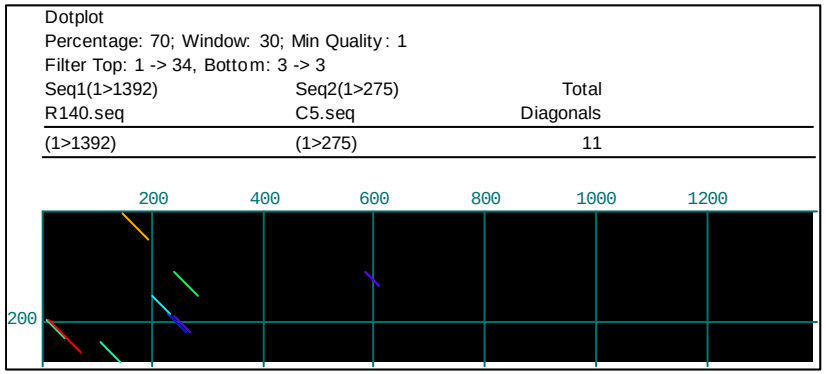
Dotplot
Percentage: 70; Window: 30; Min
Filter Top: 0 -> 0, Bottom: 39 -> 3

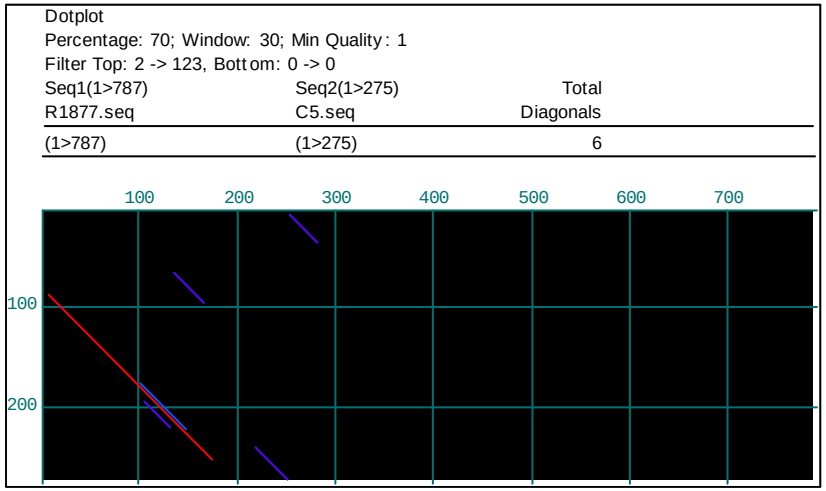
Seq1(1>139)	Seq2(1>181)	To
R3699.seq	Cl4 monom.seq	Diage
(1>139)	(1>181)	

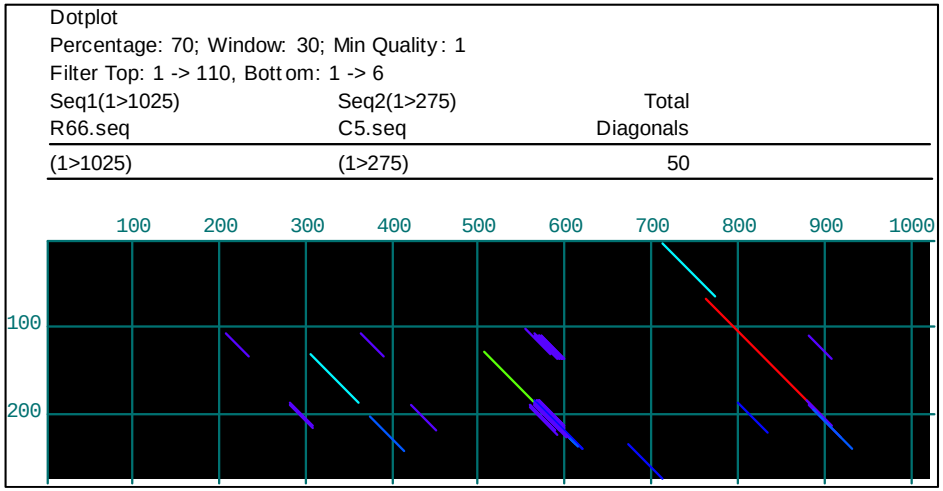


b) results for search with Cl5 monomer

R=1877|TRF=0.000|NSEG=0.261|Mid	172	4e-044
R=66|TRF=0.034|NSEG=0.471|HighA	131	1e-031
R=140|TRF=0.000|NSEG=0.215|Mid	92	1e-019

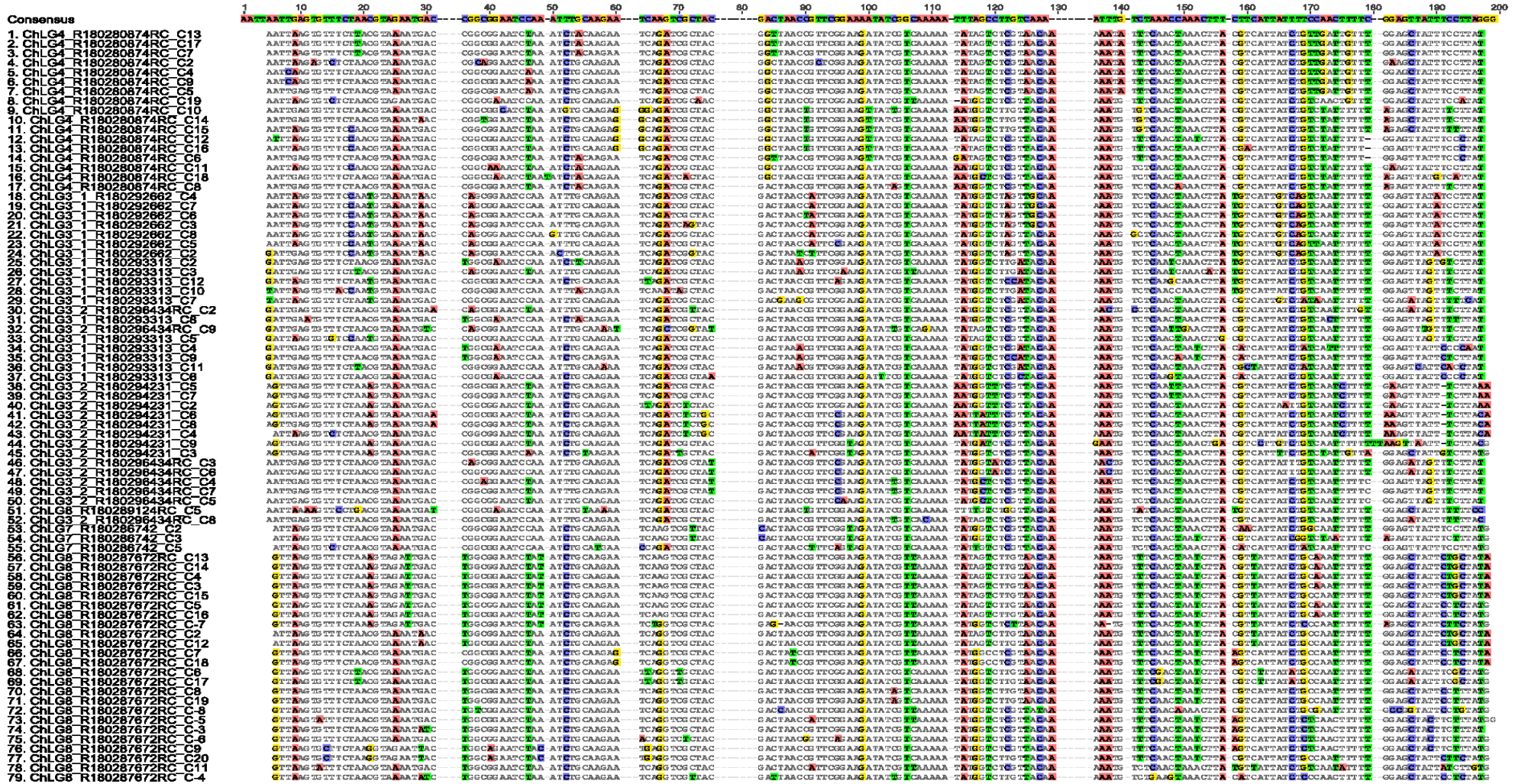


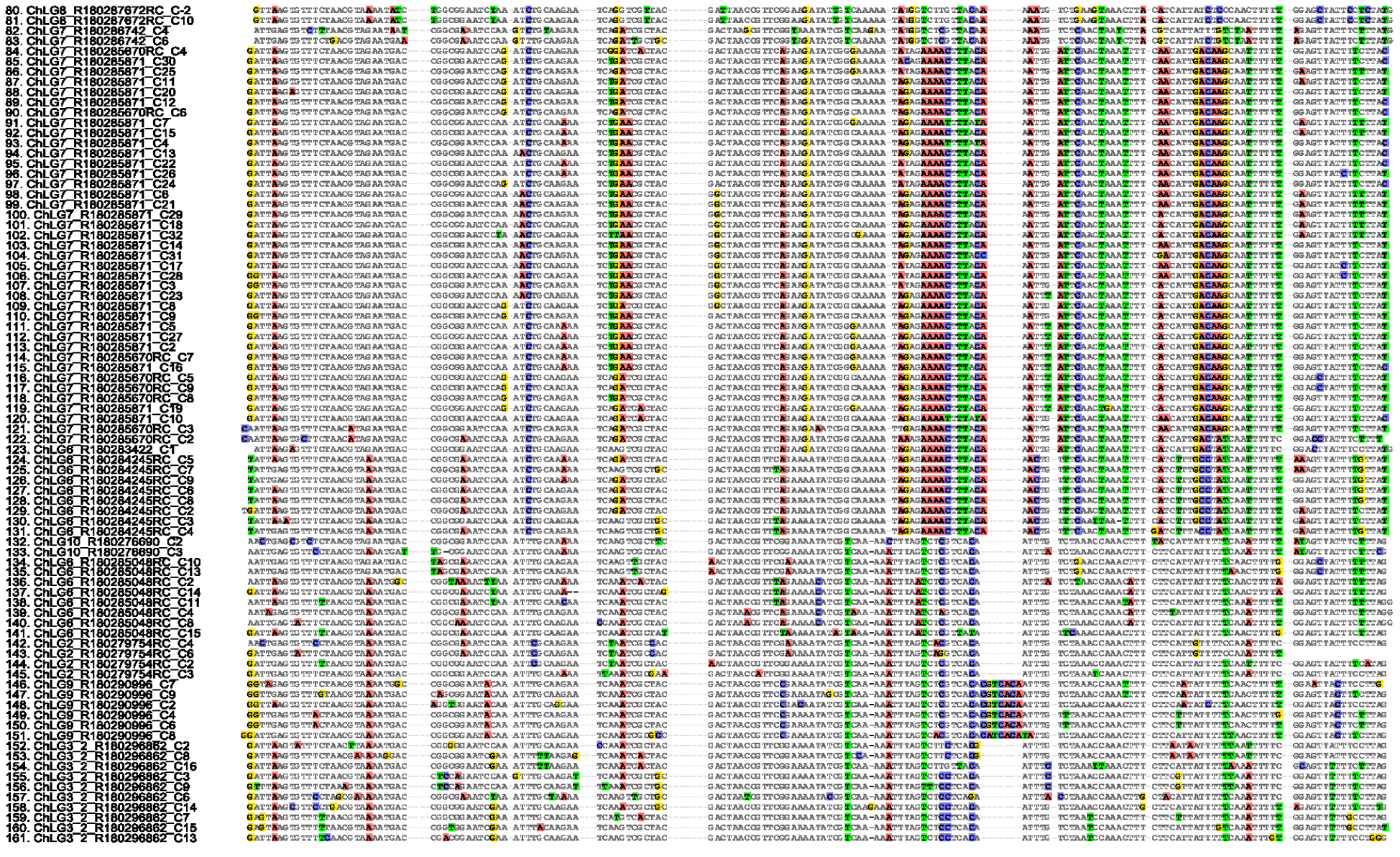


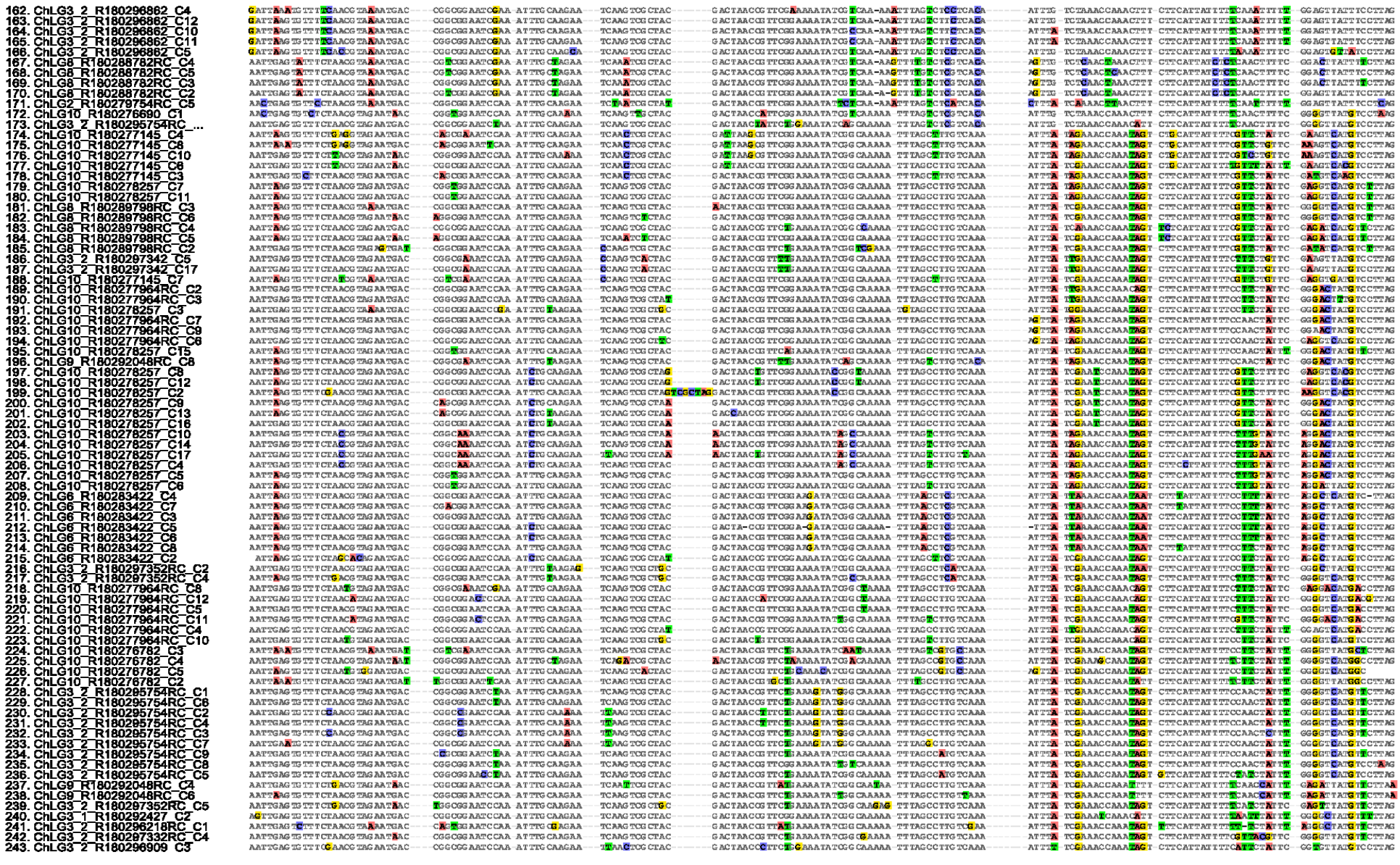


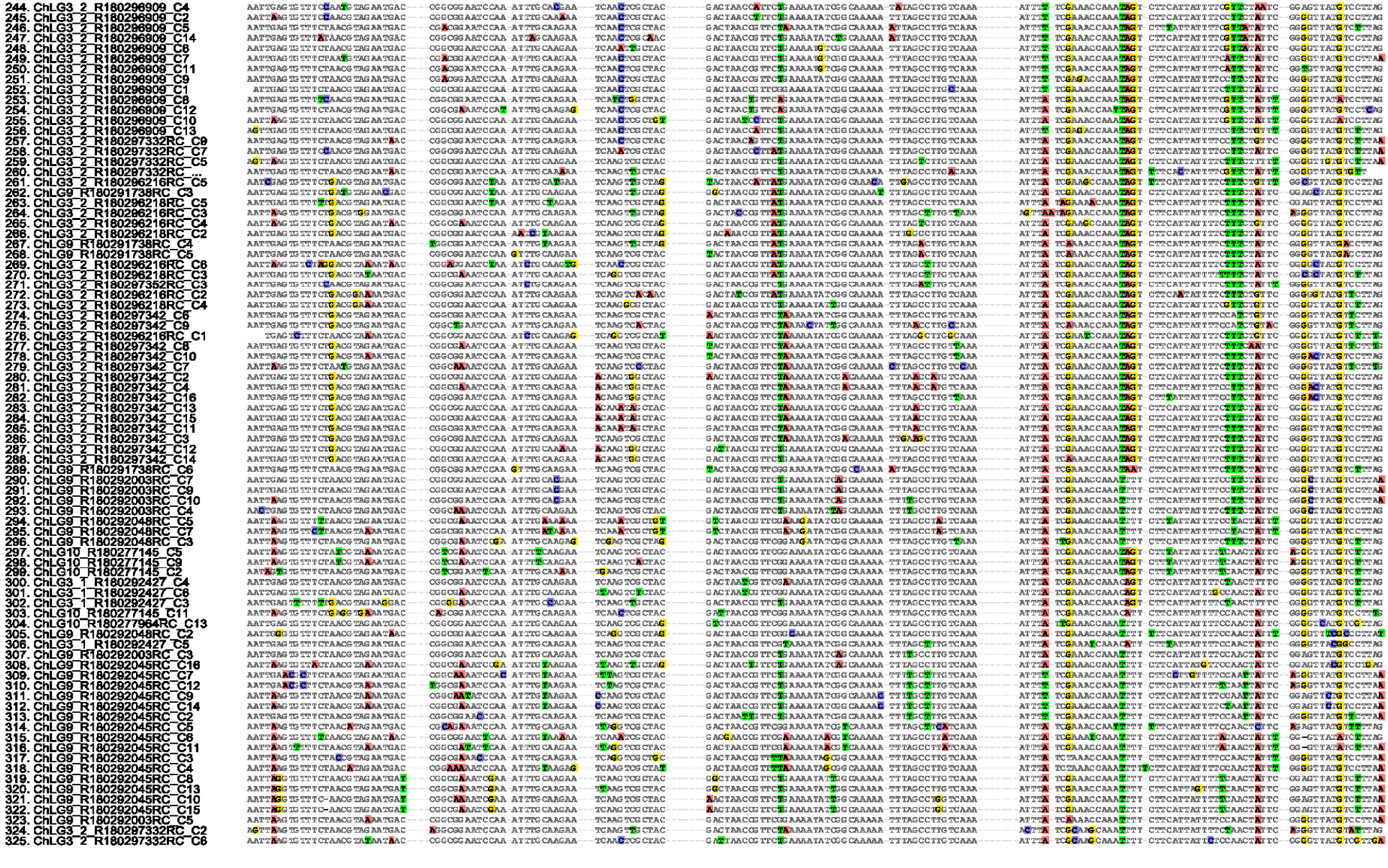
Supplementary Figure 4.2.2. Alignments of 9 clusters obtained by TRF. **a)** alignment of cluster 1; **b)** alignment of cluster 2; **c)** alignment of cluster 3; **d)** alignment of cluster 4; **e)** alignment of cluster 5; **f)** alignment of cluster 7; **g)** alignment of cluster 8; **h)** alignment of cluster 9; **i)** alignment of cluster 10

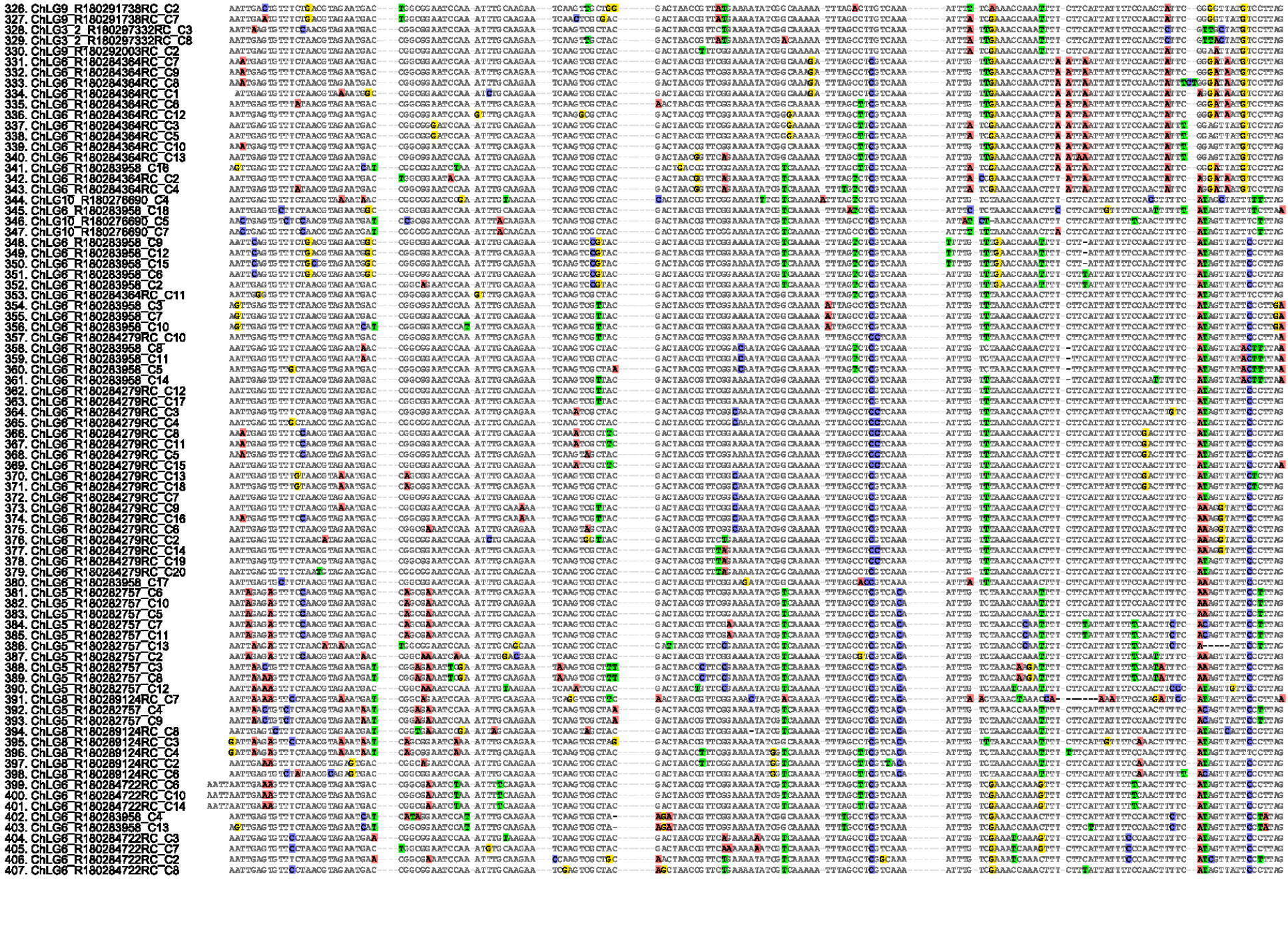
a) alignment of cluster 1

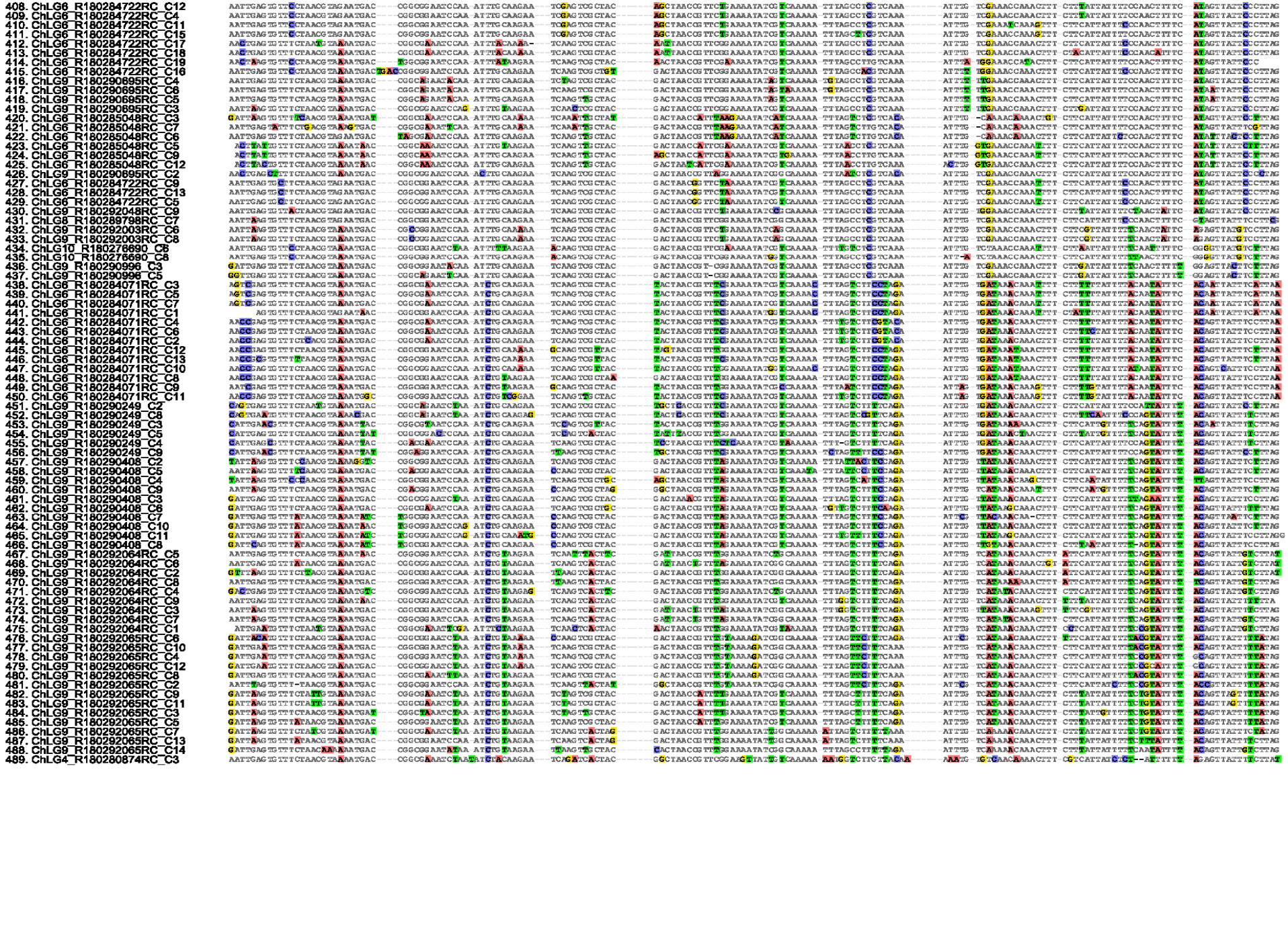



80. ChLG8_R180287672RC_C-2
81. ChLG8_R180287672RC_C10
82. ChLG7_R180286742_C4
83. ChLG7_R180286742_C6
84. ChLG7_R180285670RC_C4
85. ChLG7_R180285871_C30
86. ChLG7_R180285871_C25
87. ChLG7_R180285871_C11
88. ChLG7_R180285871_C20
89. ChLG7_R180285871_C12
90. ChLG7_R180285670RC_C6
91. ChLG7_R180285871_C7
92. ChLG7_R180285871_C15
93. ChLG7_R180285871_C4
94. ChLG7_R180285871_C13
95. ChLG7_R180285871_C22
96. ChLG7_R180285871_C26
97. ChLG7_R180285871_C24
98. ChLG7_R180285871_C6
99. ChLG7_R180285871_C21
100. ChLG7_R180285871_C29
101. ChLG7_R180285871_C18
102. ChLG7_R180285871_C32
103. ChLG7_R180285871_C14
104. ChLG7_R180285871_C31
105. ChLG7_R180285871_C17
106. ChLG7_R180285871_C28
107. ChLG7_R180285871_C3
108. ChLG7_R180285871_C23
109. ChLG7_R180285871_C8
110. ChLG7_R180285871_C9
111. ChLG7_R180285871_C5
112. ChLG7_R180285871_C27
113. ChLG7_R180285871_C2
114. ChLG7_R180285670RC_C7
115. ChLG7_R180285871_C16
116. ChLG7_R180285670RC_C5
117. ChLG7_R180285670RC_C9
118. ChLG7_R180285670RC_C8
119. ChLG7_R180285871_C19
120. ChLG7_R180285871_C10
121. ChLG7_R180285670RC_C3
122. ChLG7_R180285670RC_C2
123. ChLG6_R180283422_C1
124. ChLG6_R180284245RC_C5
125. ChLG6_R180284245RC_C7
126. ChLG6_R180284245RC_C9
127. ChLG6_R180284245RC_C6
128. ChLG6_R180284245RC_C8
129. ChLG6_R180284245RC_C2
130. ChLG6_R180284245RC_C3
131. ChLG6_R180284245RC_C4
132. ChLG10_R180276690_C2
133. ChLG10_R180276690_C3
134. ChLG6_R180285048RC_C10
135. ChLG6_R180285048RC_C13
136. ChLG6_R180285048RC_C2
137. ChLG6_R180285048RC_C14
138. ChLG6_R180285048RC_C11
139. ChLG6_R180285048RC_C4
140. ChLG6_R180285048RC_C8
141. ChLG6_R180285048RC_C15
142. ChLG2_R180279754RC_C4
143. ChLG2_R180279754RC_C6
144. ChLG2_R180279754RC_C2
145. ChLG2_R180279754RC_C3
146. ChLG9_R180290996_C7
147. ChLG9_R180290996_C9
148. ChLG9_R180290996_C2
149. ChLG9_R180290996_C4
150. ChLG9_R180290996_C6
151. ChLG9_R180290996_C8
152. ChLG3_2_R180296862_C2
153. ChLG3_2_R180296862_C8
154. ChLG3_2_R180296862_C16
155. ChLG3_2_R180296862_C3
156. ChLG3_2_R180296862_C9
157. ChLG3_2_R180296862_C6
158. ChLG3_2_R180296862_C14
159. ChLG3_2_R180296862_C7
160. ChLG3_2_R180296862_C15
161. ChLG3_2_R180296862_C13

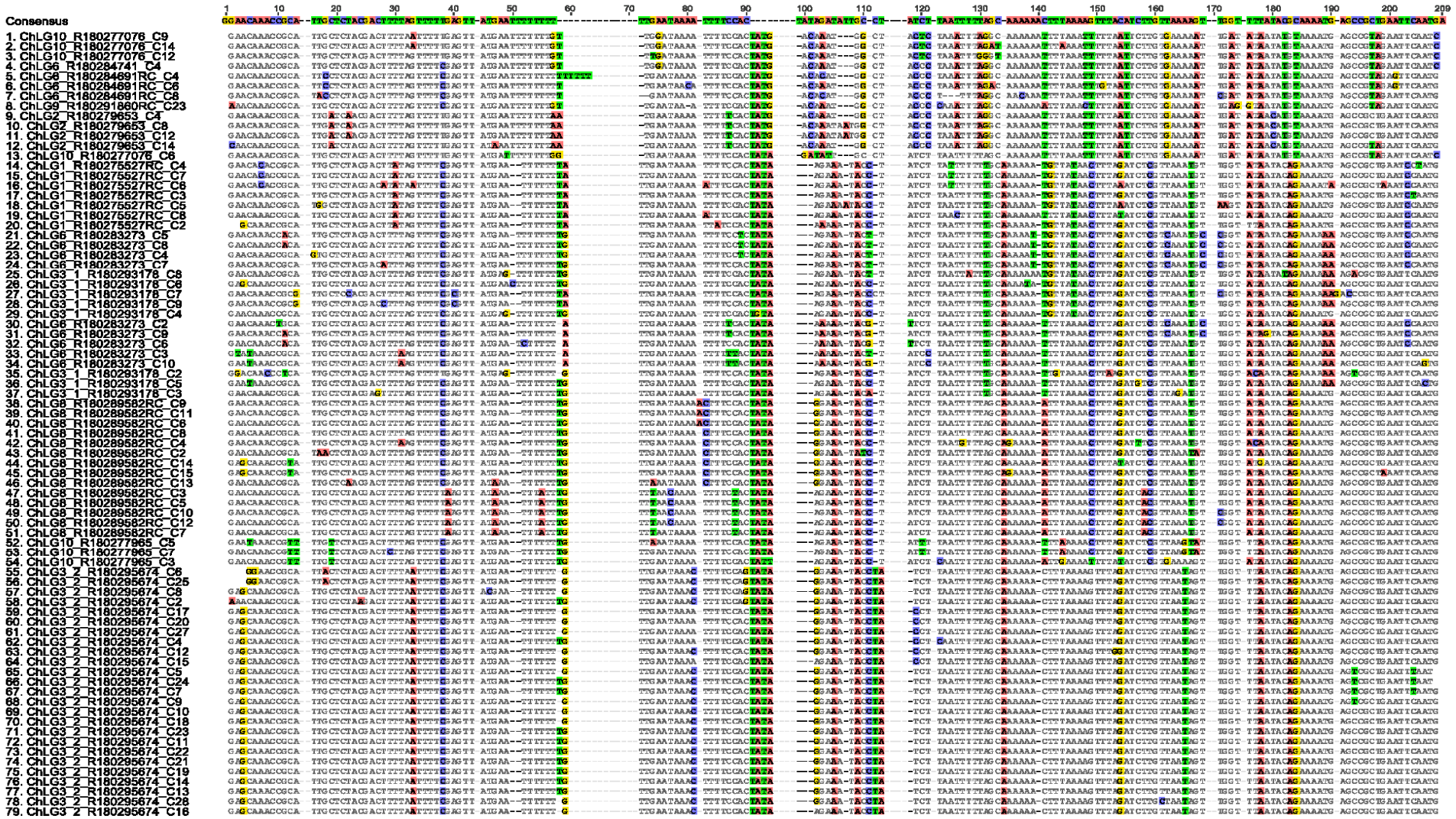

162. ChLG3_2_R180296862_C4
163. ChLG3_2_R180296862_C12
164. ChLG3_2_R180296862_C10
165. ChLG3_2_R180296862_C11
166. ChLG3_2_R180296862_C5
167. ChLG8_R180288782RC_C4
168. ChLG8_R180288782RC_C5
169. ChLG8_R180288782RC_C3
170. ChLG8_R180288782RC_C2
171. ChLG2_R180279754RC_C5
172. ChLG10_R180276690_C1
173. ChLG3_2_R180295754RC_...
174. ChLG10_R180277145_C4
175. ChLG10_R180277145_C8
176. ChLG10_R180277145_C10
177. ChLG10_R180277145_C6
178. ChLG10_R180277145_C3
179. ChLG10_R180278257_C7
180. ChLG10_R180278257_C11
181. ChLG8_R180289798RC_C3
182. ChLG8_R180289798RC_C6
183. ChLG8_R180289798RC_C4
184. ChLG8_R180289798RC_C5
185. ChLG8_R180289798RC_C2
186. ChLG3_2_R180297342_C5
187. ChLG3_2_R180297342_C17
188. ChLG10_R180277145_C7
189. ChLG10_R180277964RC_C2
190. ChLG10_R180277964RC_C3
191. ChLG10_R180278257_C3
192. ChLG10_R180277964RC_C7
193. ChLG10_R180277964RC_C9
194. ChLG10_R180277964RC_C6
195. ChLG10_R180278257_C15
196. ChLG9_R180292048RC_C8
197. ChLG10_R180278257_C8
198. ChLG10_R180278257_C12
199. ChLG10_R180278257_C2
200. ChLG10_R180278257_C9
201. ChLG10_R180278257_C13
202. ChLG10_R180278257_C16
203. ChLG10_R180278257_C10
204. ChLG10_R180278257_C14
205. ChLG10_R180278257_C17
206. ChLG10_R180278257_C4
207. ChLG10_R180278257_C5
208. ChLG10_R180278257_C6
209. ChLG6_R180283422_C4
210. ChLG6_R180283422_C7
211. ChLG6_R180283422_C3
212. ChLG6_R180283422_C5
213. ChLG6_R180283422_C6
214. ChLG6_R180283422_C8
215. ChLG6_R180283422_C2
216. ChLG3_2_R180297352RC_C2
217. ChLG3_2_R180297352RC_C4
218. ChLG10_R180277964RC_C8
219. ChLG10_R180277964RC_C12
220. ChLG10_R180277964RC_C5
221. ChLG10_R180277964RC_C11
222. ChLG10_R180277964RC_C4
223. ChLG10_R180277964RC_C10
224. ChLG10_R180276782_C3
225. ChLG10_R180276782_C4
226. ChLG10_R180276782_C5
227. ChLG10_R180276782_C2
228. ChLG3_2_R180295754RC_C1
229. ChLG3_2_R180295754RC_C6
230. ChLG3_2_R180295754RC_C2
231. ChLG3_2_R180295754RC_C4
232. ChLG3_2_R180295754RC_C3
233. ChLG3_2_R180295754RC_C7
234. ChLG3_2_R180295754RC_C9
235. ChLG3_2_R180295754RC_C8
236. ChLG3_2_R180295754RC_C5
237. ChLG9_R180292048RC_C4
238. ChLG9_R180292048RC_C6
239. ChLG3_2_R180297352RC_C5
240. ChLG3_1_R180292427_C2
241. ChLG3_2_R180296218RC_C1
242. ChLG3_2_R180297332RC_C4
243. ChLG3_2_R180296909_C3

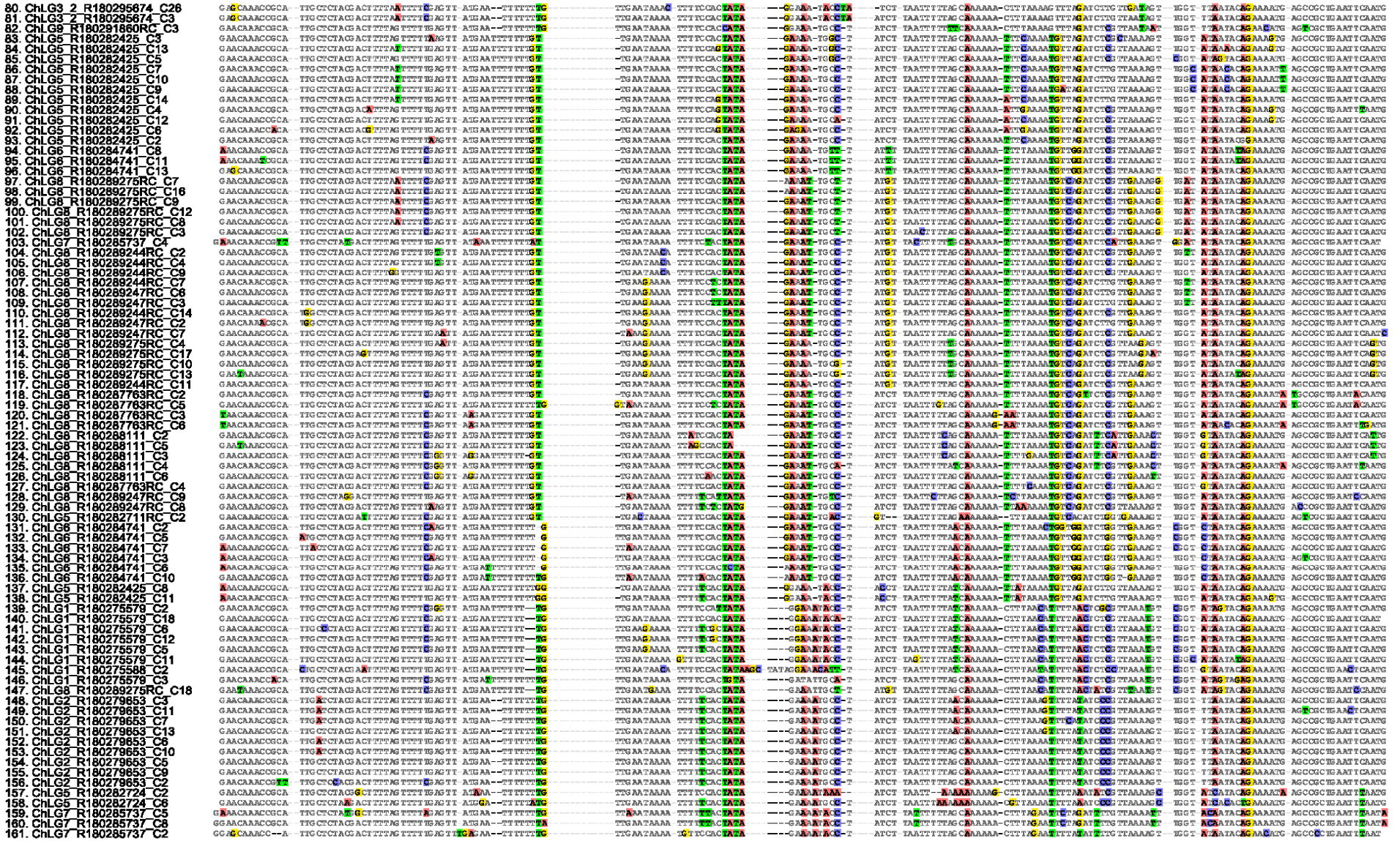

244. ChLG3_2_R180296909_C4
245. ChLG3_2_R180296909_C2
246. ChLG3_2_R180296909_C5
247. ChLG3_2_R180296909_C14
248. ChLG3_2_R180296909_C6
249. ChLG3_2_R180296909_C7
250. ChLG3_2_R180296909_C11
251. ChLG3_2_R180296909_C9
252. ChLG3_2_R180296909_C1
253. ChLG3_2_R180296909_C8
254. ChLG3_2_R180296909_C12
255. ChLG3_2_R180296909_C10
256. ChLG3_2_R180296909_C13
257. ChLG3_2_R180297332RC_C9
258. ChLG3_2_R180297332RC_C7
259. ChLG3_2_R180297332RC_C5
260. ChLG3_2_R180297332RC_...
261. ChLG3_2_R180296216RC_C5
262. ChLG9_R180291738RC_C3
263. ChLG3_2_R180296216RC_C5
264. ChLG3_2_R180296216RC_C3
265. ChLG3_2_R180296216RC_C4
266. ChLG3_2_R180296216RC_C2
267. ChLG9_R180291738RC_C4
268. ChLG9_R180291738RC_C5
269. ChLG3_2_R180296216RC_C6
270. ChLG3_2_R180296218RC_C3
271. ChLG3_2_R180297352RC_C3
272. ChLG3_2_R180296216RC_C2
273. ChLG3_2_R180296218RC_C4
274. ChLG3_2_R180297342_C6
275. ChLG3_2_R180297342_C9
276. ChLG3_2_R180296216RC_C1
277. ChLG3_2_R180297342_C8
278. ChLG3_2_R180297342_C10
279. ChLG3_2_R180297342_C7
280. ChLG3_2_R180297342_C2
281. ChLG3_2_R180297342_C4
282. ChLG3_2_R180297342_C16
283. ChLG3_2_R180297342_C13
284. ChLG3_2_R180297342_C15
285. ChLG3_2_R180297342_C11
286. ChLG3_2_R180297342_C3
287. ChLG3_2_R180297342_C12
288. ChLG3_2_R180297342_C14
289. ChLG9_R180291738RC_C6
290. ChLG9_R180292003RC_C7
291. ChLG9_R180292003RC_C9
292. ChLG9_R180292003RC_C10
293. ChLG9_R180292003RC_C4
294. ChLG9_R180292048RC_C5
295. ChLG9_R180292048RC_C7
296. ChLG9_R180292048RC_C3
297. ChLG10_R180277145_C5
298. ChLG10_R180277145_C9
299. ChLG10_R180277145_C2
300. ChLG3_1_R180292427_C4
301. ChLG3_1_R180292427_C6
302. ChLG3_1_R180292427_C3
303. ChLG10_R180277145_C11
304. ChLG10_R180277964RC_C13
305. ChLG9_R180292048RC_C2
306. ChLG3_1_R180292427_C5
307. ChLG9_R180292003RC_C3
308. ChLG9_R180292045RC_C16
309. ChLG9_R180292045RC_C7
310. ChLG9_R180292045RC_C12
311. ChLG9_R180292045RC_C9
312. ChLG9_R180292045RC_C14
313. ChLG9_R180292045RC_C2
314. ChLG9_R180292045RC_C5
315. ChLG9_R180292045RC_C6
316. ChLG9_R180292045RC_C11
317. ChLG9_R180292045RC_C3
318. ChLG9_R180292045RC_C4
319. ChLG9_R180292045RC_C8
320. ChLG9_R180292045RC_C13
321. ChLG9_R180292045RC_C10
322. ChLG9_R180292045RC_C15
323. ChLG9_R180292003RC_C5
324. ChLG3_2_R180297332RC_C2
325. ChLG3_2_R180297332RC_C6

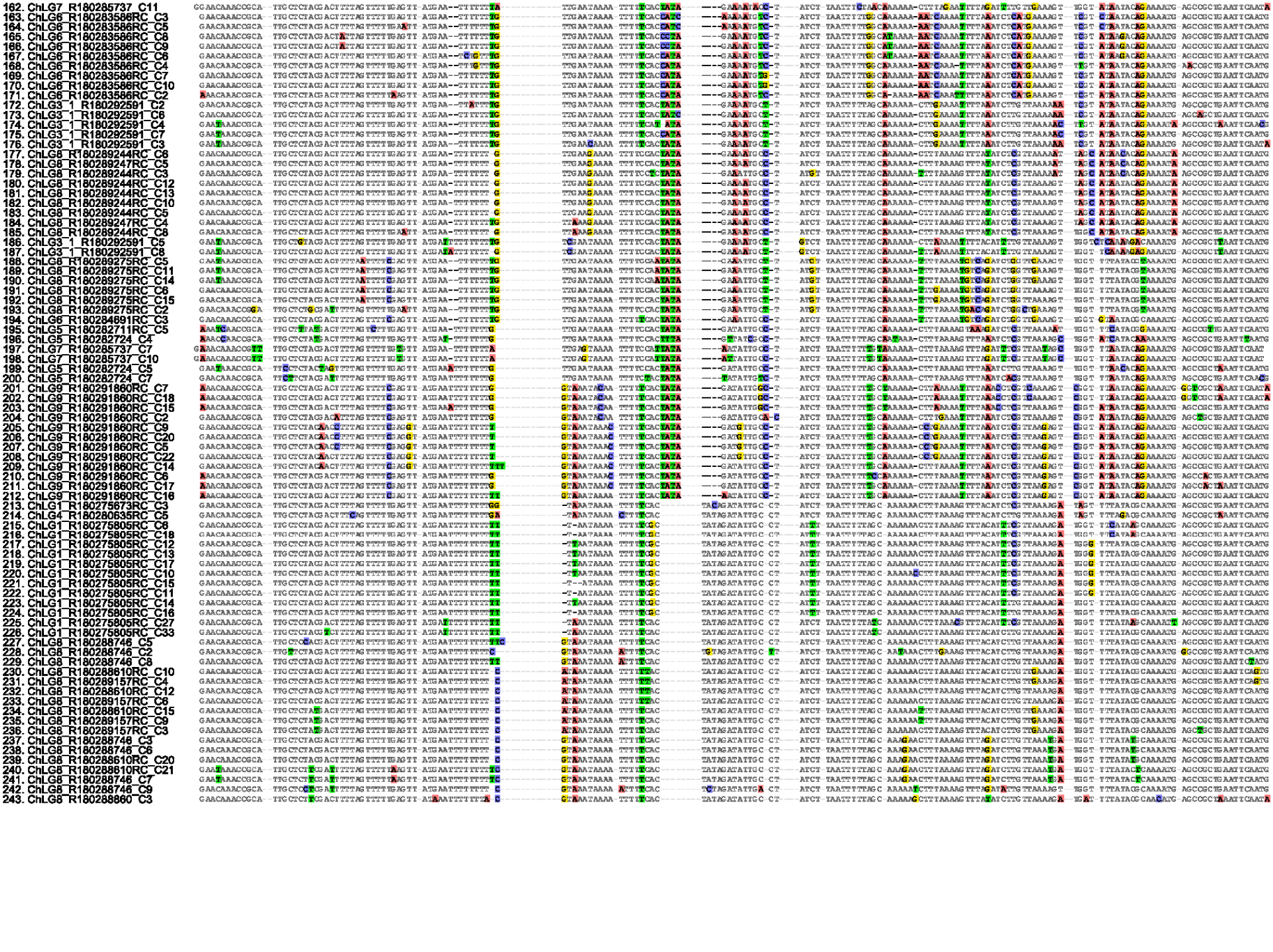

326. ChLG9_R180291738RC_C2
327. ChLG9_R180291738RC_C7
328. ChLG3_2_R180297332RC_C3
329. ChLG3_2_R180297332RC_C8
330. ChLG9_R180292003RC_C2
331. ChLG6_R180284364RC_C7
332. ChLG6_R180284364RC_C9
333. ChLG6_R180284364RC_C8
334. ChLG6_R180284364RC_C1
335. ChLG6_R180284364RC_C6
336. ChLG6_R180284364RC_C12
337. ChLG6_R180284364RC_C3
338. ChLG6_R180284364RC_C5
339. ChLG6_R180284364RC_C10
340. ChLG6_R180284364RC_C13
341. ChLG6_R180283958_C16
342. ChLG6_R180284364RC_C2
343. ChLG6_R180284364RC_C4
344. ChLG10_R180276690_C4
345. ChLG6_R180283958_C18
346. ChLG10_R180276690_C5
347. ChLG10_R180276690_C7
348. ChLG6_R180283958_C9
349. ChLG6_R180283958_C12
350. ChLG6_R180283958_C15
351. ChLG6_R180283958_C6
352. ChLG6_R180283958_C2
353. ChLG6_R180284364RC_C11
354. ChLG6_R180283958_C3
355. ChLG6_R180283958_C7
356. ChLG6_R180283958_C10
357. ChLG6_R180284279RC_C10
358. ChLG6_R180283958_C8
359. ChLG6_R180283958_C11
360. ChLG6_R180283958_C5
361. ChLG6_R180283958_C14
362. ChLG6_R180284279RC_C12
363. ChLG6_R180284279RC_C17
364. ChLG6_R180284279RC_C3
365. ChLG6_R180284279RC_C4
366. ChLG6_R180284279RC_C8
367. ChLG6_R180284279RC_C11
368. ChLG6_R180284279RC_C5
369. ChLG6_R180284279RC_C15
370. ChLG6_R180284279RC_C13
371. ChLG6_R180284279RC_C18
372. ChLG6_R180284279RC_C7
373. ChLG6_R180284279RC_C9
374. ChLG6_R180284279RC_C16
375. ChLG6_R180284279RC_C6
376. ChLG6_R180284279RC_C2
377. ChLG6_R180284279RC_C14
378. ChLG6_R180284279RC_C19
379. ChLG6_R180284279RC_C20
380. ChLG6_R180283958_C17
381. ChLG5_R180282757_C6
382. ChLG5_R180282757_C10
383. ChLG5_R180282757_C5
384. ChLG5_R180282757_C7
385. ChLG5_R180282757_C11
386. ChLG5_R180282757_C13
387. ChLG5_R180282757_C2
388. ChLG5_R180282757_C3
389. ChLG5_R180282757_C8
390. ChLG5_R180282757_C12
391. ChLG8_R180289124RC_C7
392. ChLG5_R180282757_C4
393. ChLG5_R180282757_C9
394. ChLG8_R180289124RC_C8
395. ChLG8_R180289124RC_C3
396. ChLG8_R180289124RC_C4
397. ChLG8_R180289124RC_C2
398. ChLG8_R180289124RC_C6
399. ChLG6_R180284722RC_C6
400. ChLG6_R180284722RC_C10
401. ChLG6_R180284722RC_C14
402. ChLG6_R180283958_C4
403. ChLG6_R180283958_C13
404. ChLG6_R180284722RC_C3
405. ChLG6_R180284722RC_C7
406. ChLG6_R180284722RC_C2
407. ChLG6_R180284722RC_C8

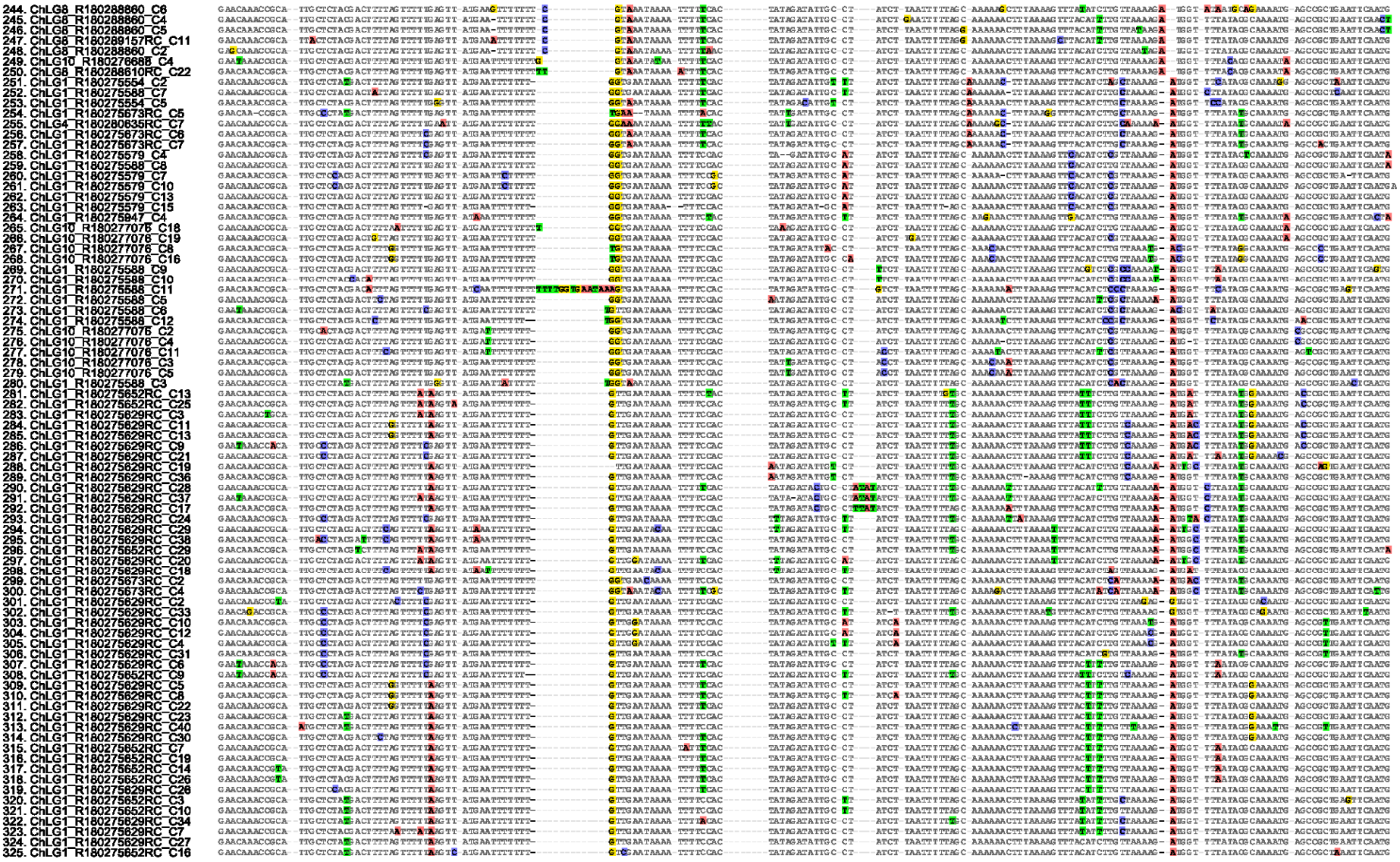

408. ChLG6_R180284722RC_C12
409. ChLG6_R180284722RC_C4
410. ChLG6_R180284722RC_C11
411. ChLG6_R180284722RC_C15
412. ChLG6_R180284722RC_C17
413. ChLG6_R180284722RC_C18
414. ChLG6_R180284722RC_C19
415. ChLG6_R180284722RC_C16
416. ChLG9_R180290695RC_C4
417. ChLG9_R180290695RC_C6
418. ChLG9_R180290695RC_C5
419. ChLG9_R180290695RC_C3
420. ChLG6_R180285048RC_C3
421. ChLG6_R180285048RC_C7
422. ChLG6_R180285048RC_C6
423. ChLG6_R180285048RC_C5
424. ChLG6_R180285048RC_C9
425. ChLG6_R180285048RC_C12
426. ChLG9_R180290695RC_C2
427. ChLG6_R180284722RC_C9
428. ChLG6_R180284722RC_C13
429. ChLG6_R180284722RC_C5
430. ChLG9_R180292048RC_C9
431. ChLG8_R180289798RC_C7
432. ChLG9_R180292003RC_C6
433. ChLG9_R180292003RC_C8
434. ChLG10_R180276690_C6
435. ChLG10_R180276690_C8
436. ChLG9_R180290996_C3
437. ChLG9_R180290996_C5
438. ChLG6_R180284071RC_C3
439. ChLG6_R180284071RC_C5
440. ChLG6_R180284071RC_C7
441. ChLG6_R180284071RC_C1
442. ChLG6_R180284071RC_C4
443. ChLG6_R180284071RC_C6
444. ChLG6_R180284071RC_C2
445. ChLG6_R180284071RC_C12
446. ChLG6_R180284071RC_C13
447. ChLG6_R180284071RC_C10
448. ChLG6_R180284071RC_C8
449. ChLG6_R180284071RC_C9
450. ChLG6_R180284071RC_C11
451. ChLG9_R180290249_C2
452. ChLG9_R180290249_C8
453. ChLG9_R180290249_C3
454. ChLG9_R180290249_C5
455. ChLG9_R180290249_C4
456. ChLG9_R180290249_C9
457. ChLG9_R180290408_C2
458. ChLG9_R180290408_C5
459. ChLG9_R180290408_C4
460. ChLG9_R180290408_C9
461. ChLG9_R180290408_C3
462. ChLG9_R180290408_C6
463. ChLG9_R180290408_C7
464. ChLG9_R180290408_C10
465. ChLG9_R180290408_C11
466. ChLG9_R180290408_C8
467. ChLG6_R180292064RC_C5
468. ChLG6_R180292064RC_C6
469. ChLG6_R180292064RC_C2
470. ChLG6_R180292064RC_C8
471. ChLG6_R180292064RC_C4
472. ChLG6_R180292064RC_C9
473. ChLG6_R180292064RC_C3
474. ChLG6_R180292064RC_C7
475. ChLG6_R180292064RC_C1
476. ChLG6_R180292065RC_C6
477. ChLG6_R180292065RC_C10
478. ChLG6_R180292065RC_C4
479. ChLG6_R180292065RC_C12
480. ChLG6_R180292065RC_C8
481. ChLG6_R180292065RC_C2
482. ChLG6_R180292065RC_C9
483. ChLG6_R180292065RC_C11
484. ChLG6_R180292065RC_C3
485. ChLG6_R180292065RC_C5
486. ChLG6_R180292065RC_C7
487. ChLG6_R180292065RC_C13
488. ChLG6_R180292065RC_C14
489. ChLG4_R180280874RC_C3

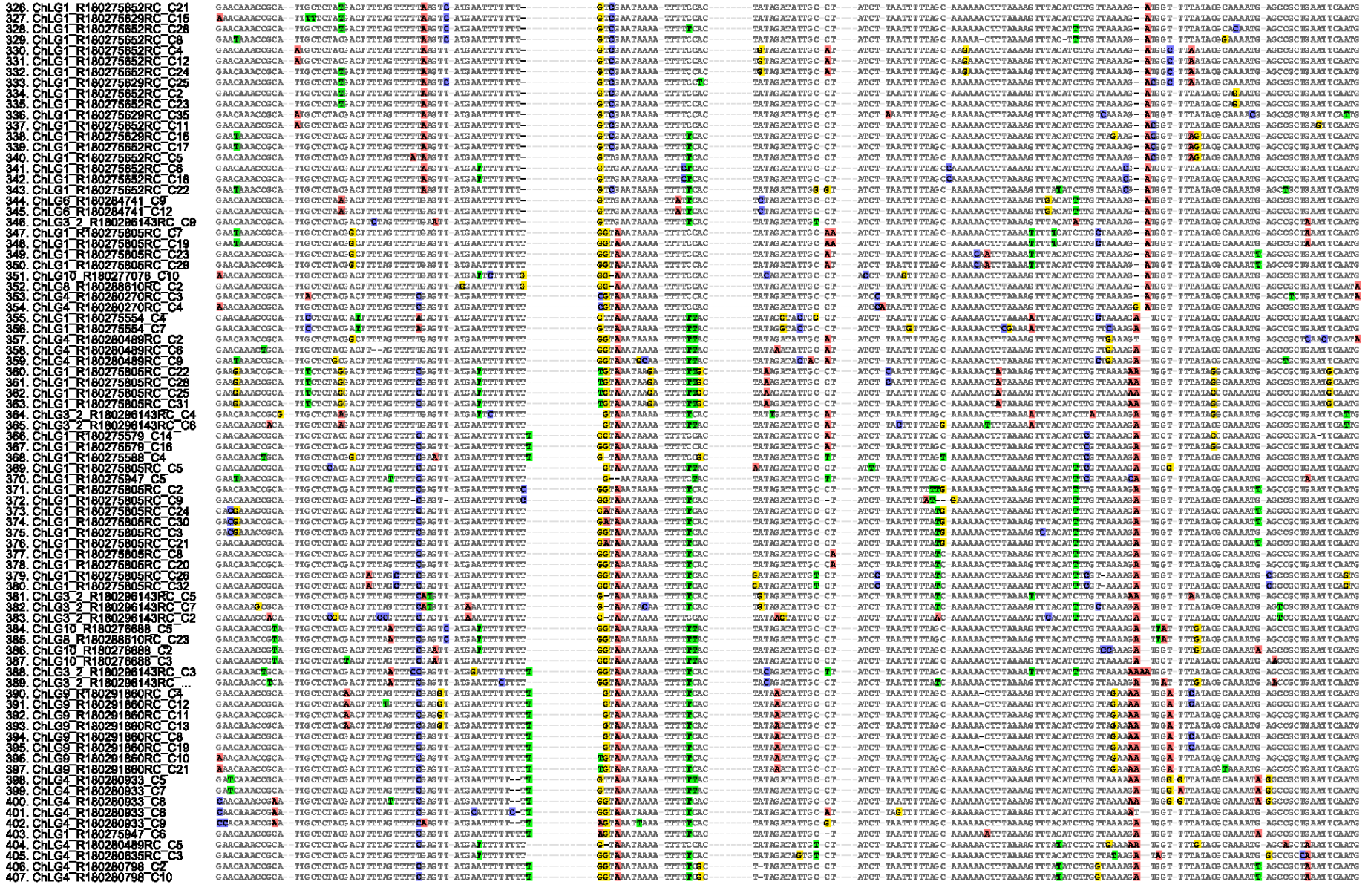
b) alignment of cluster 2

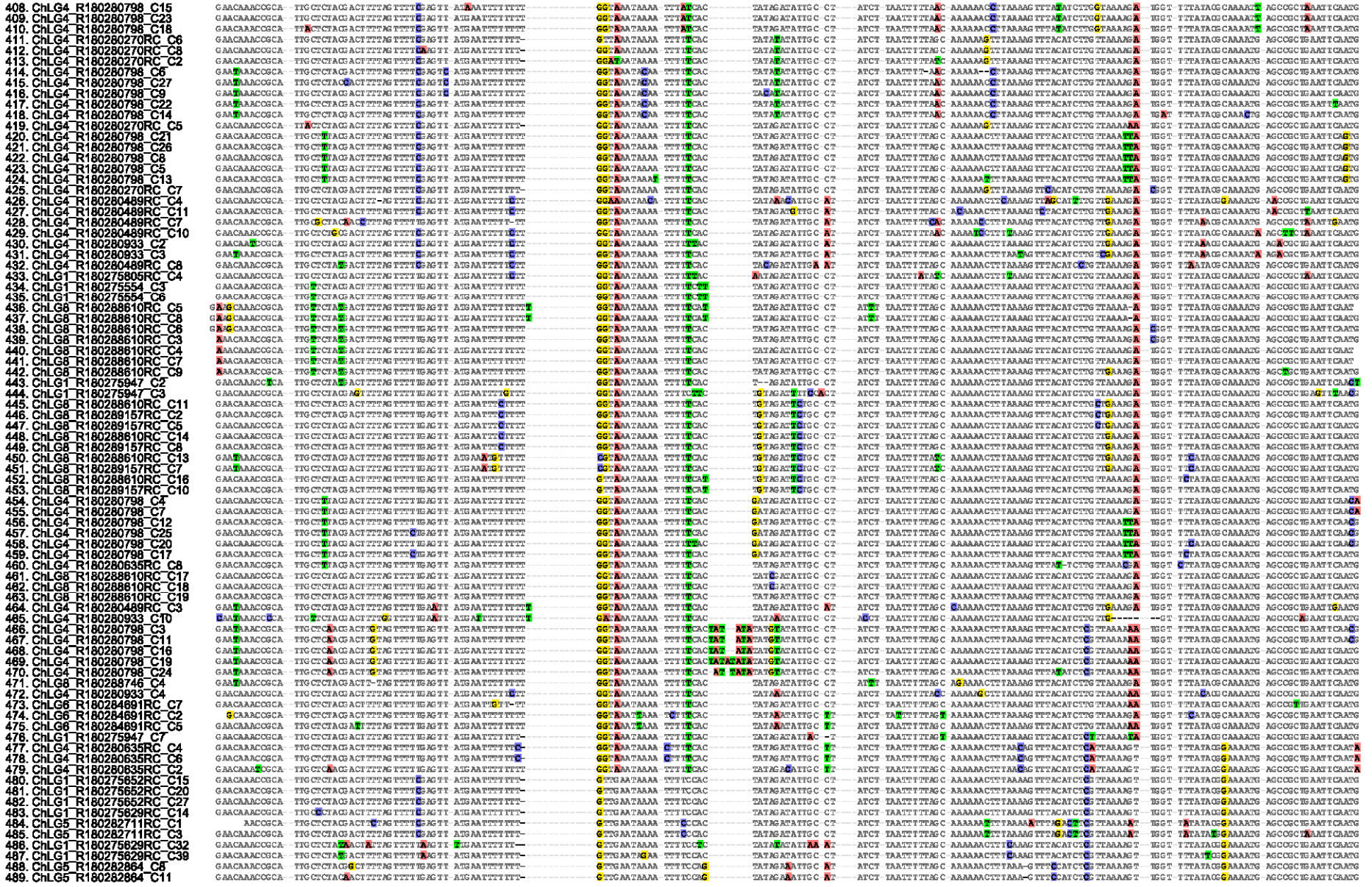


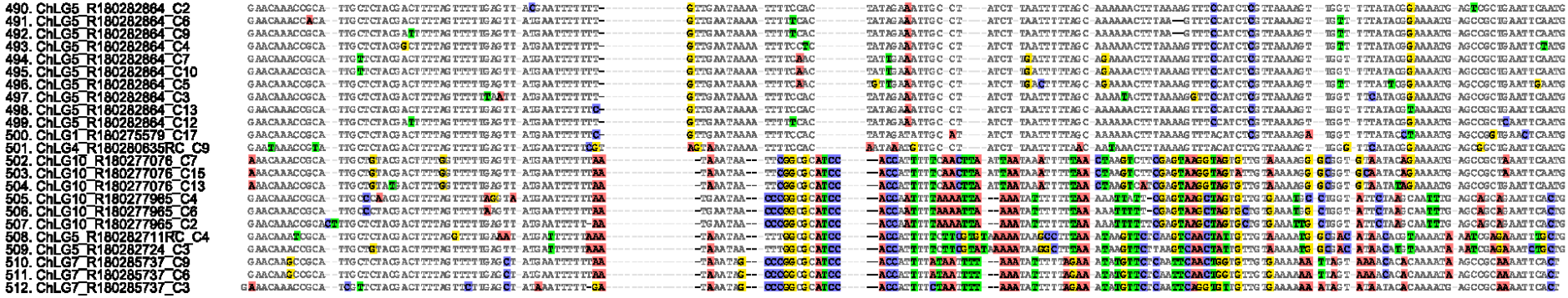

80. ChLG3_2_R180295674_C26
81. ChLG3_2_R180295674_C3
82. ChLG9_R180291860RC_C3
83. ChLG5_R180282425_C3
84. ChLG5_R180282425_C13
85. ChLG5_R180282425_C5
86. ChLG5_R180282425_C7
87. ChLG5_R180282425_C10
88. ChLG5_R180282425_C9
89. ChLG5_R180282425_C14
90. ChLG5_R180282425_C4
91. ChLG5_R180282425_C12
92. ChLG5_R180282425_C6
93. ChLG5_R180282425_C2
94. ChLG6_R180284741_C8
95. ChLG6_R180284741_C11
96. ChLG6_R180284741_C13
97. ChLG8_R180289275RC_C7
98. ChLG8_R180289275RC_C16
99. ChLG8_R180289275RC_C9
100. ChLG8_R180289275RC_C12
101. ChLG8_R180289275RC_C8
102. ChLG8_R180289275RC_C3
103. ChLG7_R180285737_C4
104. ChLG8_R180289244RC_C2
105. ChLG8_R180289244RC_C4
106. ChLG8_R180289244RC_C9
107. ChLG8_R180289244RC_C7
108. ChLG8_R180289247RC_C6
109. ChLG8_R180289247RC_C3
110. ChLG8_R180289244RC_C14
111. ChLG8_R180289247RC_C2
112. ChLG8_R180289247RC_C7
113. ChLG8_R180289275RC_C4
114. ChLG8_R180289275RC_C17
115. ChLG8_R180289275RC_C10
116. ChLG8_R180289275RC_C13
117. ChLG8_R180289244RC_C11
118. ChLG8_R180287763RC_C2
119. ChLG8_R180287763RC_C5
120. ChLG8_R180287763RC_C3
121. ChLG8_R180287763RC_C6
122. ChLG8_R180288111_C2
123. ChLG8_R180288111_C5
124. ChLG8_R180288111_C3
125. ChLG8_R180288111_C4
126. ChLG8_R180288111_C6
127. ChLG8_R180287763RC_C4
128. ChLG8_R180289247RC_C9
129. ChLG8_R180289247RC_C8
130. ChLG5_R180282711RC_C2
131. ChLG6_R180284741_C2
132. ChLG6_R180284741_C5
133. ChLG6_R180284741_C7
134. ChLG6_R180284741_C3
135. ChLG6_R180284741_C6
136. ChLG6_R180284741_C10
137. ChLG5_R180282425_C8
138. ChLG5_R180282425_C11
139. ChLG1_R180275579_C2
140. ChLG1_R180275579_C18
141. ChLG1_R180275579_C6
142. ChLG1_R180275579_C12
143. ChLG1_R180275579_C5
144. ChLG1_R180275579_C11
145. ChLG1_R180275588_C2
146. ChLG1_R180275579_C3
147. ChLG8_R180289275RC_C18
148. ChLG2_R180279653_C3
149. ChLG2_R180279653_C11
150. ChLG2_R180279653_C7
151. ChLG2_R180279653_C13
152. ChLG2_R180279653_C6
153. ChLG2_R180279653_C10
154. ChLG2_R180279653_C5
155. ChLG2_R180279653_C9
156. ChLG2_R180279653_C2
157. ChLG5_R180282724_C2
158. ChLG5_R180282724_C6
159. ChLG7_R180285737_C5
160. ChLG7_R180285737_C8
161. ChLG7_R180285737_C2


162. ChLG7_R180285737_C11
163. ChLG6_R180283586RC_C3
164. ChLG6_R180283586RC_C5
165. ChLG6_R180283586RC_C8
166. ChLG6_R180283586RC_C9
167. ChLG6_R180283586RC_C6
168. ChLG6_R180283586RC_C4
169. ChLG6_R180283586RC_C7
170. ChLG6_R180283586RC_C10
171. ChLG6_R180283586RC_C2
172. ChLG3_1_R180292591_C2
173. ChLG3_1_R180292591_C8
174. ChLG3_1_R180292591_C4
175. ChLG3_1_R180292591_C7
176. ChLG3_1_R180292591_C3
177. ChLG8_R180289244RC_C8
178. ChLG8_R180289247RC_C5
179. ChLG8_R180289244RC_C3
180. ChLG8_R180289244RC_C12
181. ChLG8_R180289244RC_C13
182. ChLG8_R180289244RC_C10
183. ChLG8_R180289244RC_C5
184. ChLG8_R180289247RC_C4
185. ChLG8_R180289244RC_C8
186. ChLG3_1_R180292591_C5
187. ChLG3_1_R180292591_C8
188. ChLG8_R180289275RC_C5
189. ChLG8_R180289275RC_C11
190. ChLG8_R180289275RC_C14
191. ChLG8_R180289275RC_C6
192. ChLG8_R180289275RC_C15
193. ChLG8_R180289275RC_C2
194. ChLG6_R180284691RC_C3
195. ChLG5_R180282711RC_C5
196. ChLG5_R180282724_C4
197. ChLG7_R180285737_C7
198. ChLG7_R180285737_C10
199. ChLG5_R180282724_C5
200. ChLG5_R180282724_C7
201. ChLG9_R180291860RC_C7
202. ChLG9_R180291860RC_C18
203. ChLG9_R180291860RC_C15
204. ChLG9_R180291860RC_C2
205. ChLG9_R180291860RC_C9
206. ChLG9_R180291860RC_C20
207. ChLG9_R180291860RC_C5
208. ChLG9_R180291860RC_C22
209. ChLG9_R180291860RC_C14
210. ChLG9_R180291860RC_C6
211. ChLG9_R180291860RC_C17
212. ChLG9_R180291860RC_C16
213. ChLG1_R180275673RC_C3
214. ChLG4_R180280635RC_C5
215. ChLG1_R180275805RC_C8
216. ChLG1_R180275805RC_C18
217. ChLG1_R180275805RC_C12
218. ChLG1_R180275805RC_C13
219. ChLG1_R180275805RC_C17
220. ChLG1_R180275805RC_C10
221. ChLG1_R180275805RC_C15
222. ChLG1_R180275805RC_C11
223. ChLG1_R180275805RC_C14
224. ChLG1_R180275805RC_C16
225. ChLG1_R180275805RC_C27
226. ChLG1_R180275805RC_C33
227. ChLG8_R180288746_C5
228. ChLG8_R180288746_C2
229. ChLG8_R180288746_C8
230. ChLG8_R180288610RC_C10
231. ChLG8_R180289157RC_C4
232. ChLG8_R180288610RC_C12
233. ChLG8_R180289157RC_C6
234. ChLG8_R180288610RC_C15
235. ChLG8_R180289157RC_C9
236. ChLG8_R180289157RC_C3
237. ChLG8_R180288746_C3
238. ChLG8_R180288746_C6
239. ChLG8_R180288610RC_C20
240. ChLG8_R180288610RC_C21
241. ChLG8_R180288746_C7
242. ChLG8_R180288746_C9
243. ChLG8_R180288860_C3

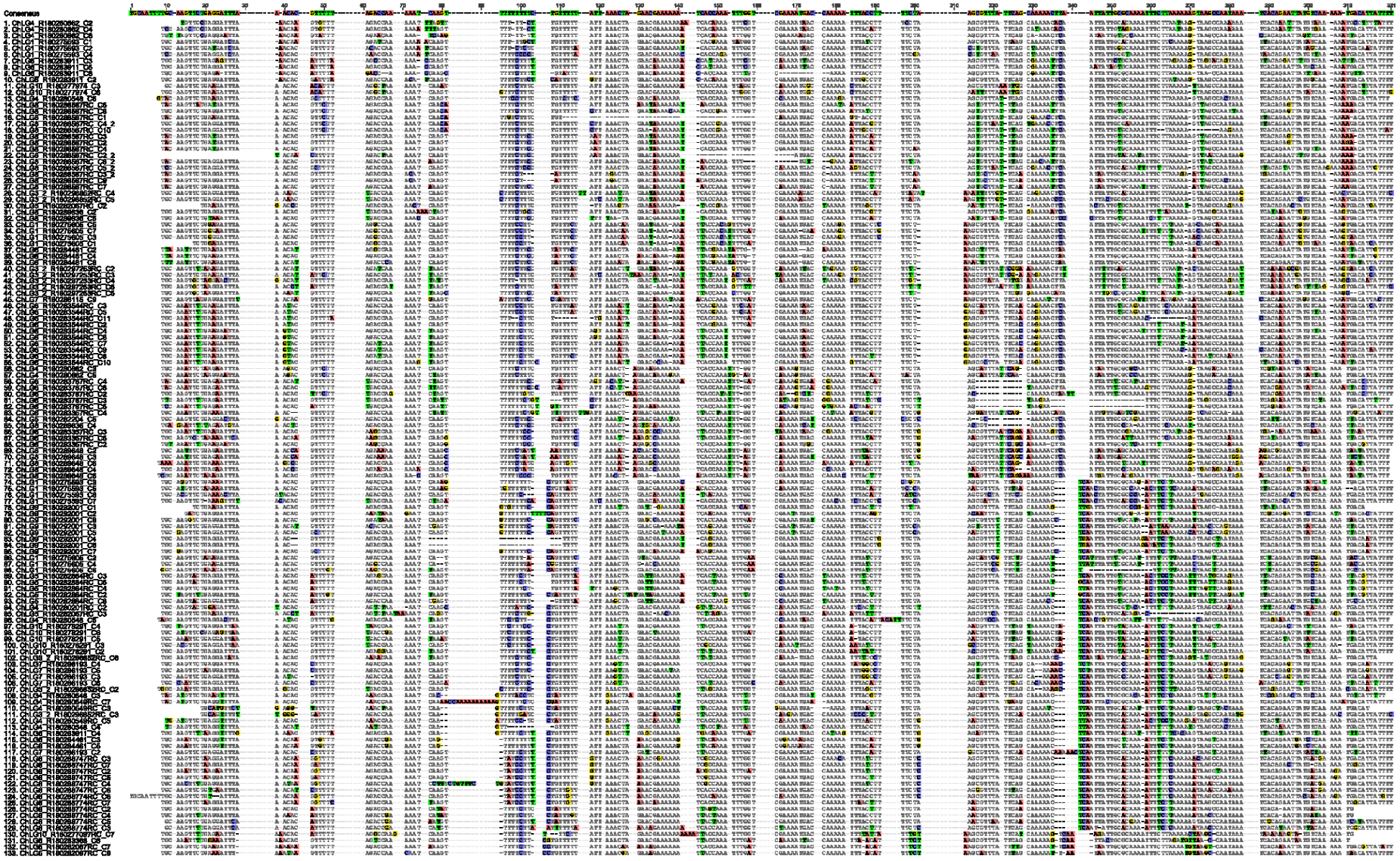

244. ChLG8_R180288860_C6
245. ChLG8_R180288860_C4
246. ChLG8_R180288860_C5
247. ChLG8_R180269157RC_C11
248. ChLG8_R180288860_C2
249. ChLG10_R180276688_C4
250. ChLG8_R180288610RC_C22
251. ChLG1_R180275554_C2
252. ChLG1_R180275588_C7
253. ChLG1_R180275554_C5
254. ChLG1_R180275673RC_C5
255. ChLG4_R180280635RC_C7
256. ChLG1_R180275673RC_C6
257. ChLG1_R180275673RC_C7
258. ChLG1_R180275579_C4
259. ChLG1_R180275588_C8
260. ChLG1_R180275579_C7
261. ChLG1_R180275579_C10
262. ChLG1_R180275579_C13
263. ChLG1_R180275579_C15
264. ChLG1_R180275847_C4
265. ChLG10_R180277076_C18
266. ChLG10_R180277076_C19
267. ChLG10_R180277076_C8
268. ChLG10_R180277076_C16
269. ChLG1_R180275588_C9
270. ChLG1_R180275588_C10
271. ChLG1_R180275588_C11
272. ChLG1_R180275588_C5
273. ChLG1_R180275588_C6
274. ChLG1_R180275588_C12
275. ChLG10_R180277076_C2
276. ChLG10_R180277076_C4
277. ChLG10_R180277076_C11
278. ChLG10_R180277076_C3
279. ChLG10_R180277076_C5
280. ChLG1_R180275588_C3
281. ChLG1_R180275652RC_C13
282. ChLG1_R180275652RC_C25
283. ChLG1_R180275629RC_C3
284. ChLG1_R180275629RC_C11
285. ChLG1_R180275629RC_C13
286. ChLG1_R180275629RC_C9
287. ChLG1_R180275629RC_C21
288. ChLG1_R180275629RC_C19
289. ChLG1_R180275629RC_C36
290. ChLG1_R180275629RC_C28
291. ChLG1_R180275629RC_C37
292. ChLG1_R180275629RC_C17
293. ChLG1_R180275629RC_C24
294. ChLG1_R180275629RC_C29
295. ChLG1_R180275629RC_C38
296. ChLG1_R180275652RC_C29
297. ChLG1_R180275629RC_C20
298. ChLG1_R180275629RC_C18
299. ChLG1_R180275673RC_C2
300. ChLG1_R180275673RC_C4
301. ChLG1_R180275629RC_C22
302. ChLG1_R180275629RC_C33
303. ChLG1_R180275629RC_C10
304. ChLG1_R180275629RC_C12
305. ChLG1_R180275629RC_C4
306. ChLG1_R180275629RC_C31
307. ChLG1_R180275629RC_C6
308. ChLG1_R180275652RC_C9
309. ChLG1_R180275629RC_C5
310. ChLG1_R180275629RC_C8
311. ChLG1_R180275629RC_C22
312. ChLG1_R180275629RC_C23
313. ChLG1_R180275629RC_C40
314. ChLG1_R180275629RC_C30
315. ChLG1_R180275652RC_C7
316. ChLG1_R180275652RC_C19
317. ChLG1_R180275652RC_C14
318. ChLG1_R180275652RC_C26
319. ChLG1_R180275629RC_C26
320. ChLG1_R180275652RC_C3
321. ChLG1_R180275652RC_C10
322. ChLG1_R180275629RC_C34
323. ChLG1_R180275629RC_C7
324. ChLG1_R180275629RC_C27
325. ChLG1_R180275652RC_C16

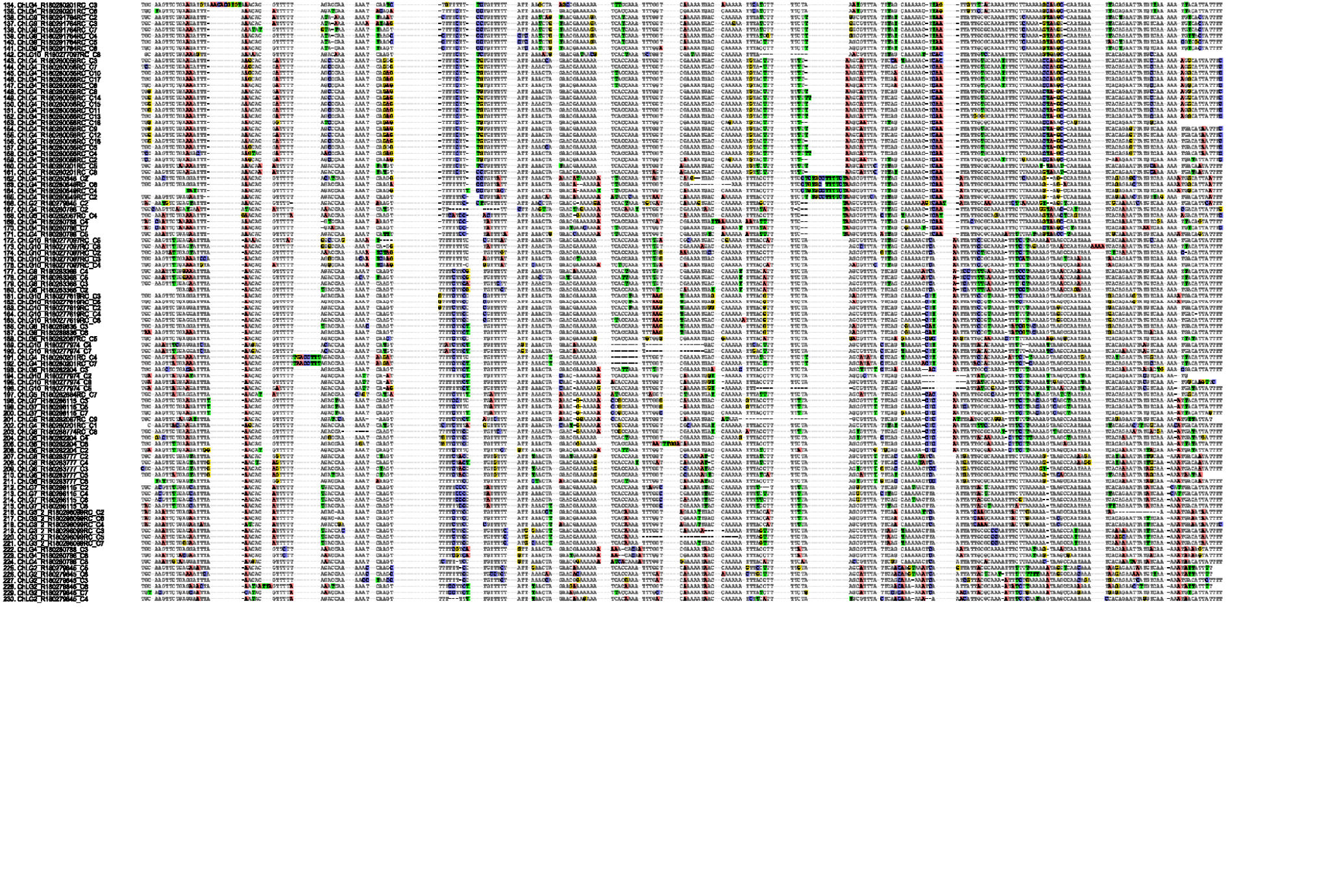

326. ChLG1_R180275652RC_C21
327. ChLG1_R180275629RC_C15
328. ChLG1_R180275652RC_C28
329. ChLG1_R180275652RC_C8
330. ChLG1_R180275652RC_C4
331. ChLG1_R180275652RC_C12
332. ChLG1_R180275652RC_C24
333. ChLG1_R180275629RC_C25
334. ChLG1_R180275652RC_C2
335. ChLG1_R180275652RC_C23
336. ChLG1_R180275629RC_C35
337. ChLG1_R180275652RC_C11
338. ChLG1_R180275629RC_C16
339. ChLG1_R180275652RC_C17
340. ChLG1_R180275652RC_C5
341. ChLG1_R180275652RC_C6
342. ChLG1_R180275652RC_C18
343. ChLG1_R180275652RC_C22
344. ChLG6_R180284741_C9
345. ChLG6_R180284741_C12
346. ChLG3_2_R180296143RC_C9
347. ChLG1_R180275805RC_C7
348. ChLG1_R180275805RC_C19
349. ChLG1_R180275805RC_C23
350. ChLG1_R180275805RC_C29
351. ChLG10_R180277076_C10
352. ChLG8_R180288610RC_C2
353. ChLG4_R180280270RC_C3
354. ChLG4_R180280270RC_C4
355. ChLG1_R180275554_C4
356. ChLG1_R180275554_C7
357. ChLG4_R180280489RC_C2
358. ChLG4_R180280489RC_C6
359. ChLG4_R180280489RC_C9
360. ChLG1_R180275805RC_C22
361. ChLG1_R180275805RC_C28
362. ChLG1_R180275805RC_C25
363. ChLG1_R180275805RC_C31
364. ChLG3_2_R180296143RC_C4
365. ChLG3_2_R180296143RC_C8
366. ChLG1_R180275579_C14
367. ChLG1_R180275579_C16
368. ChLG1_R180275688_C4
369. ChLG1_R180275805RC_C5
370. ChLG1_R180275947_C5
371. ChLG1_R180275805RC_C2
372. ChLG1_R180275805RC_C9
373. ChLG1_R180275805RC_C24
374. ChLG1_R180275805RC_C30
375. ChLG1_R180275805RC_C3
376. ChLG1_R180275805RC_C21
377. ChLG1_R180275805RC_C8
378. ChLG1_R180275805RC_C20
379. ChLG1_R180275805RC_C26
380. ChLG1_R180275805RC_C32
381. ChLG3_2_R180296143RC_C5
382. ChLG3_2_R180296143RC_C7
383. ChLG3_2_R180296143RC_C2
384. ChLG10_R180276688_C5
385. ChLG8_R180288610RC_C23
386. ChLG10_R180276688_C2
387. ChLG10_R180276688_C3
388. ChLG3_2_R180296143RC_C3
389. ChLG3_2_R180296143RC_...
390. ChLG9_R180291860RC_C4
391. ChLG9_R180291860RC_C12
392. ChLG9_R180291860RC_C11
393. ChLG9_R180291860RC_C13
394. ChLG9_R180291860RC_C8
395. ChLG9_R180291860RC_C19
396. ChLG9_R180291860RC_C10
397. ChLG9_R180291860RC_C21
398. ChLG4_R180280933_C5
399. ChLG4_R180280933_C7
400. ChLG4_R180280933_C8
401. ChLG4_R180280933_C6
402. ChLG4_R180280933_C9
403. ChLG1_R180275947_C6
404. ChLG4_R180280489RC_C5
405. ChLG4_R180280635RC_C3
406. ChLG4_R180280798_C2
407. ChLG4_R180280798_C10


408. ChLG4_R180280798_C15
409. ChLG4_R180280798_C23
410. ChLG4_R180280798_C18
411. ChLG4_R180280270RC_C6
412. ChLG4_R180280270RC_C8
413. ChLG4_R180280270RC_C2
414. ChLG4_R180280798_C6
415. ChLG4_R180280798_C27
416. ChLG4_R180280798_C9
417. ChLG4_R180280798_C22
418. ChLG4_R180280798_C14
419. ChLG4_R180280270RC_C5
420. ChLG4_R180280798_C21
421. ChLG4_R180280798_C26
422. ChLG4_R180280798_C8
423. ChLG4_R180280798_C5
424. ChLG4_R180280798_C13
425. ChLG4_R180280270RC_C7
426. ChLG4_R180280489RC_C4
427. ChLG4_R180280489RC_C11
428. ChLG4_R180280489RC_C7
429. ChLG4_R180280489RC_C10
430. ChLG4_R180280933_C2
431. ChLG4_R180280933_C3
432. ChLG4_R180280489RC_C8
433. ChLG1_R180275805RC_C4
434. ChLG1_R180275554_C3
435. ChLG1_R180275554_C6
436. ChLG8_R180288610RC_C5
437. ChLG8_R180288610RC_C8
438. ChLG8_R180288610RC_C6
439. ChLG8_R180288610RC_C3
440. ChLG8_R180288610RC_C4
441. ChLG8_R180288610RC_C7
442. ChLG8_R180288610RC_C9
443. ChLG1_R180275947_C2
444. ChLG1_R180275947_C3
445. ChLG8_R180288610RC_C11
446. ChLG8_R180289157RC_C2
447. ChLG8_R180289157RC_C5
448. ChLG8_R180288610RC_C14
449. ChLG8_R180289157RC_C8
450. ChLG8_R180288610RC_C13
451. ChLG8_R180289157RC_C7
452. ChLG8_R180288610RC_C16
453. ChLG8_R180289157RC_C10
454. ChLG4_R180280798_C4
455. ChLG4_R180280798_C7
456. ChLG4_R180280798_C12
457. ChLG4_R180280798_C25
458. ChLG4_R180280798_C20
459. ChLG4_R180280798_C17
460. ChLG4_R180280635RC_C8
461. ChLG8_R180288610RC_C17
462. ChLG8_R180288610RC_C18
463. ChLG8_R180288610RC_C19
464. ChLG4_R180280489RC_C3
465. ChLG4_R180280933_C10
466. ChLG4_R180280798_C3
467. ChLG4_R180280798_C11
468. ChLG4_R180280798_C16
469. ChLG4_R180280798_C19
470. ChLG4_R180280798_C24
471. ChLG8_R180288746_C4
472. ChLG4_R180280933_C4
473. ChLG6_R180284691RC_C7
474. ChLG6_R180284691RC_C2
475. ChLG6_R180284691RC_C5
476. ChLG1_R180275947_C7
477. ChLG4_R180280635RC_C4
478. ChLG4_R180280635RC_C6
479. ChLG4_R180280635RC_C2
480. ChLG1_R180275652RC_C15
481. ChLG1_R180275652RC_C20
482. ChLG1_R180275652RC_C27
483. ChLG1_R180275629RC_C14
484. ChLG5_R180282711RC_C1
485. ChLG5_R180282711RC_C3
486. ChLG1_R180275629RC_C32
487. ChLG1_R180275629RC_C39
488. ChLG5_R180282864_C8
489. ChLG5_R180282864_C11

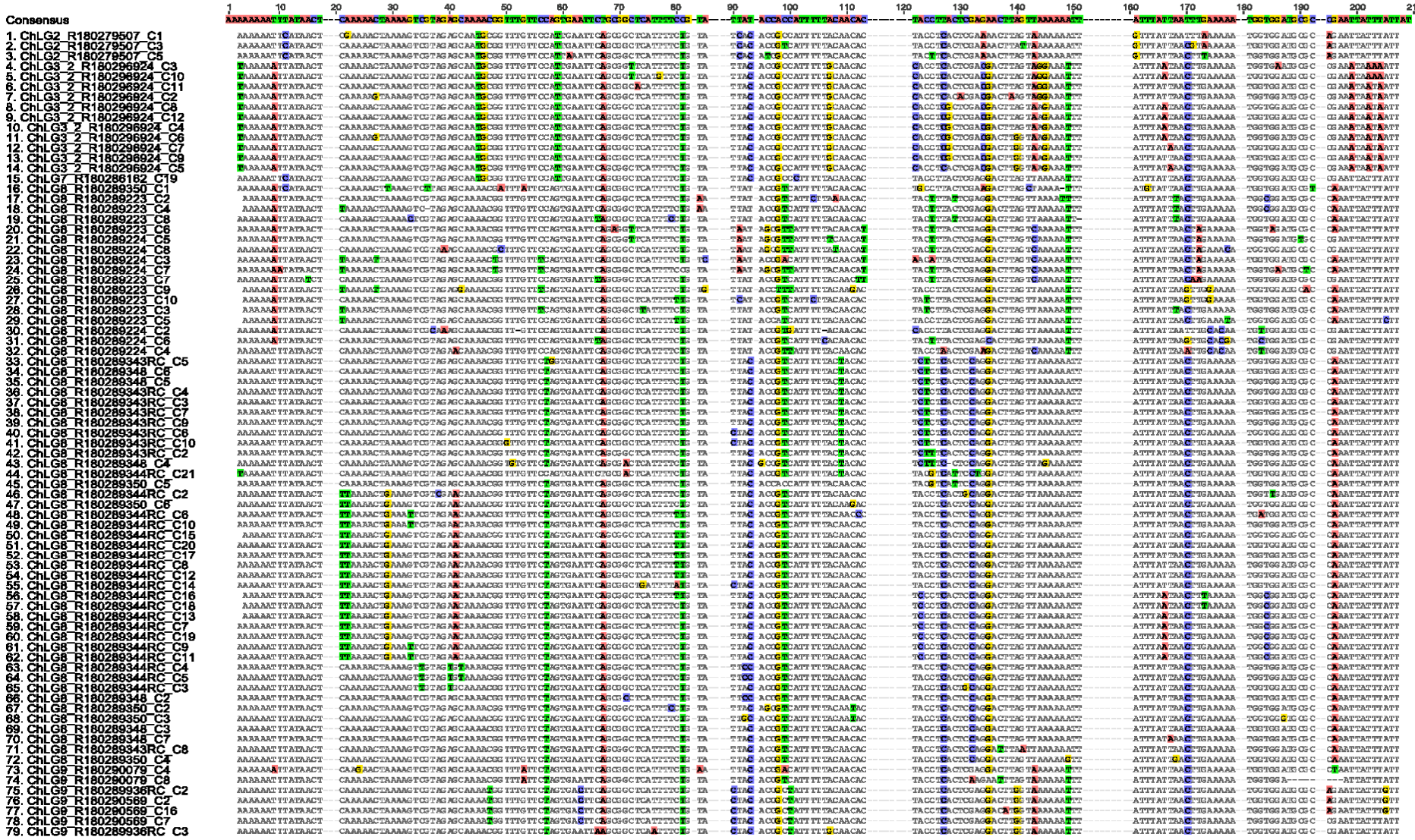

490. ChLG5_R180282864_C2
491. ChLG5_R180282864_C6
492. ChLG5_R180282864_C9
493. ChLG5_R180282864_C4
494. ChLG5_R180282864_C7
495. ChLG5_R180282864_C10
496. ChLG5_R180282864_C5
497. ChLG5_R180282864_C3
498. ChLG5_R180282864_C13
499. ChLG5_R180282864_C12
500. ChLG1_R180275579_C17
501. ChLG4_R180280635RC_C9
502. ChLG10_R180277076_C7
503. ChLG10_R180277076_C15
504. ChLG10_R180277076_C13
505. ChLG10_R180277965_C4
506. ChLG10_R180277965_C6
507. ChLG10_R180277965_C2
508. ChLG5_R180282711RC_C4
509. ChLG5_R180282724_C3
510. ChLG7_R180285737_C9
511. ChLG7_R180285737_C6
512. ChLG7_R180285737_C3

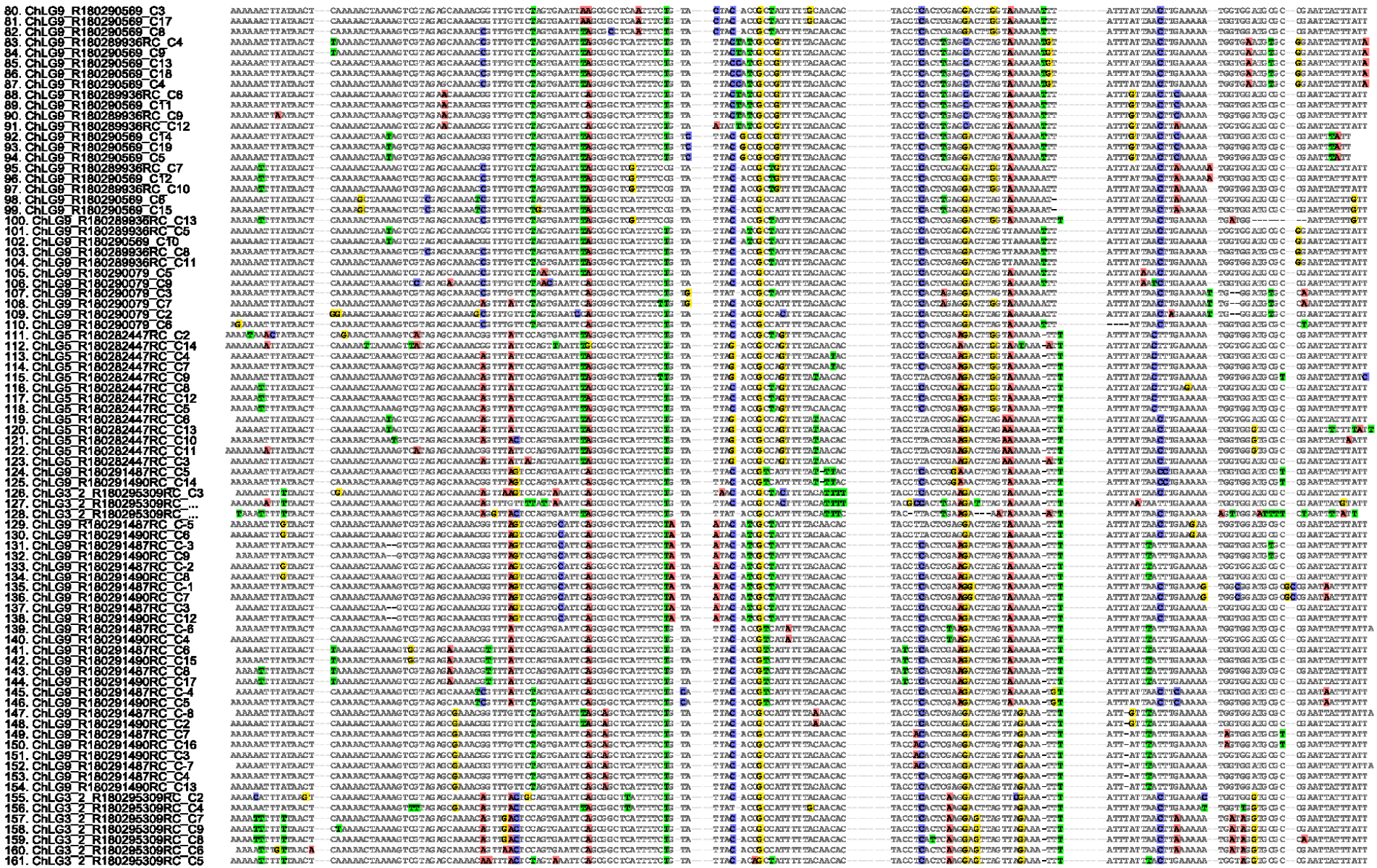
c) alignment of cluster 3

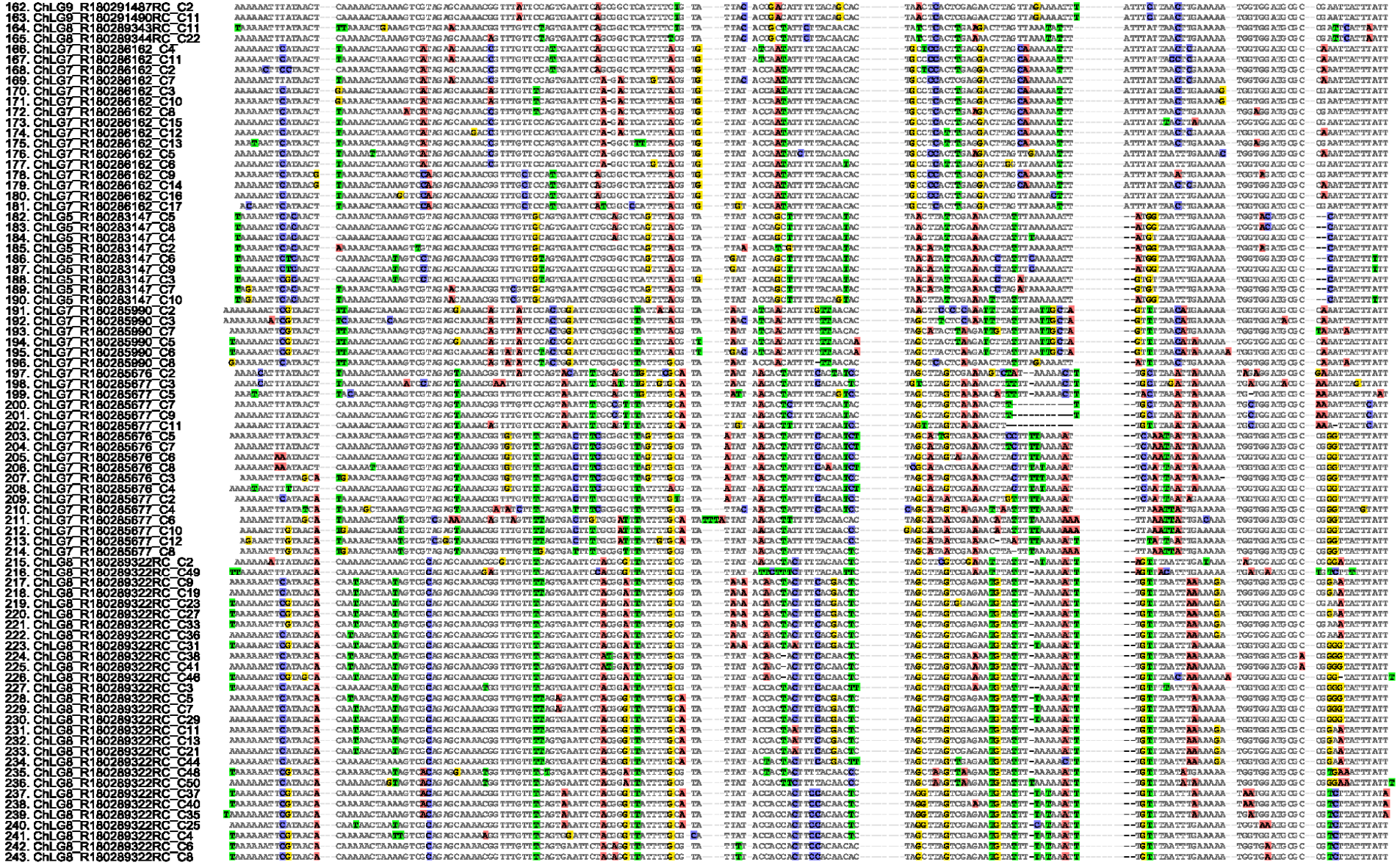


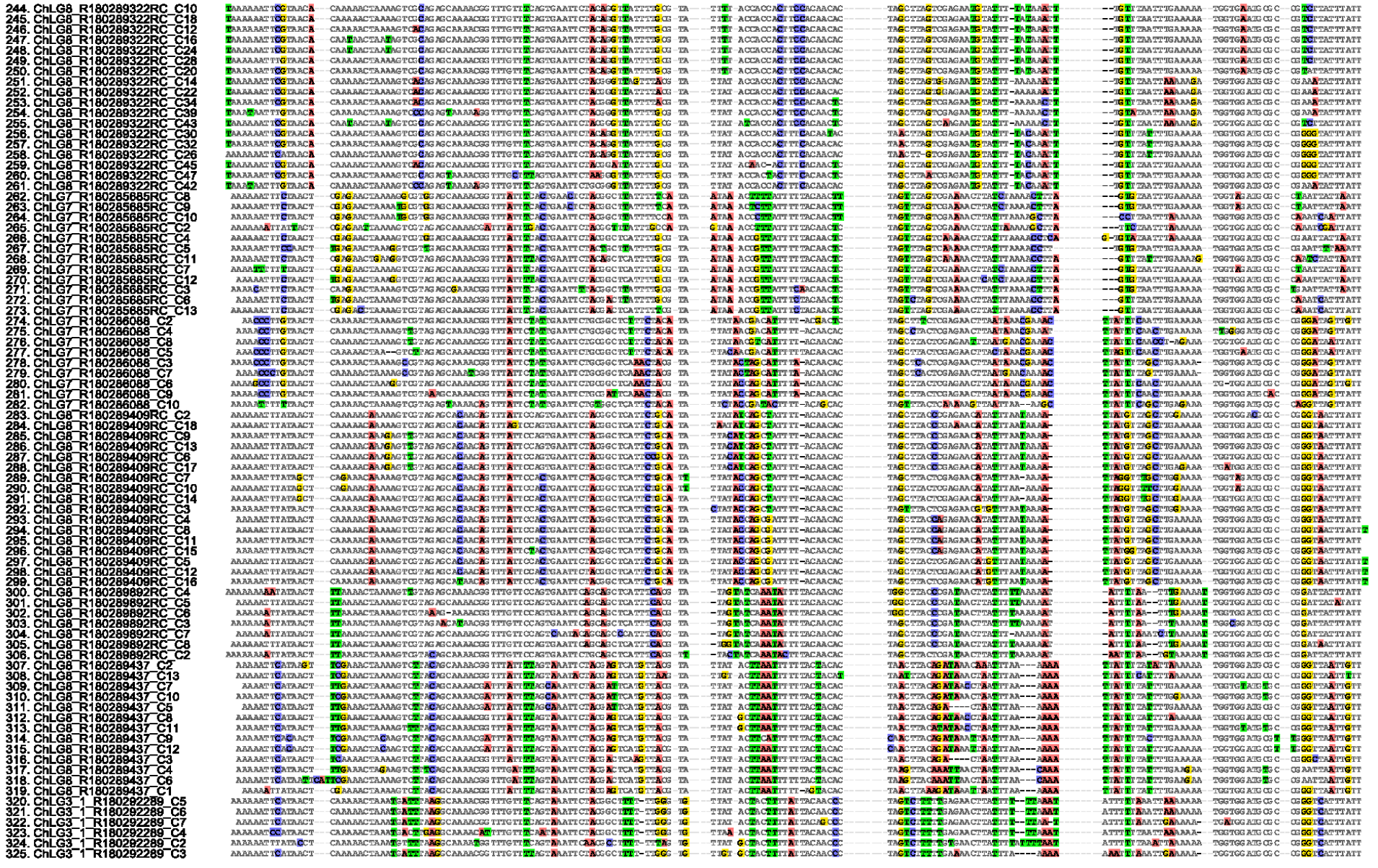


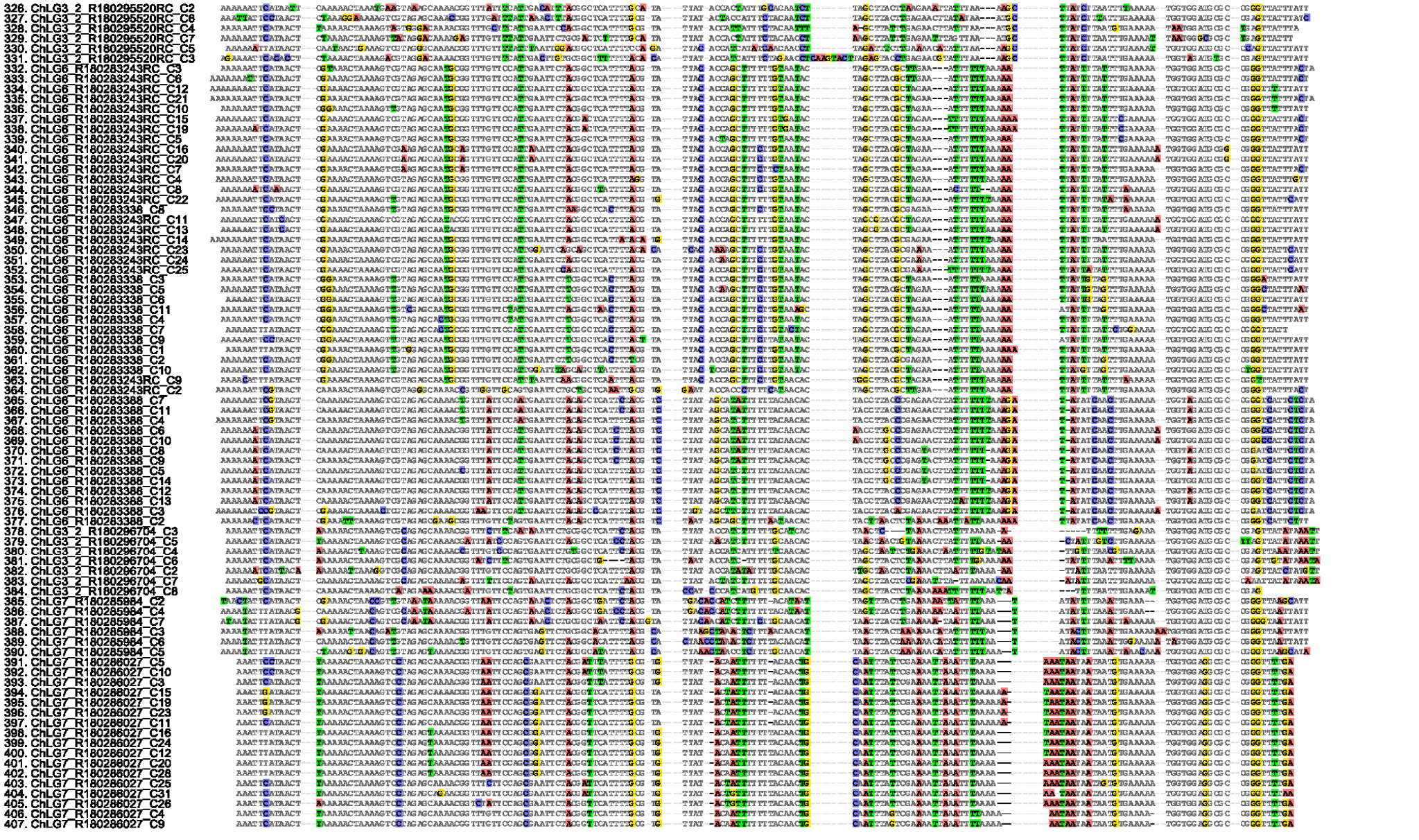
d) alignment of cluster 4

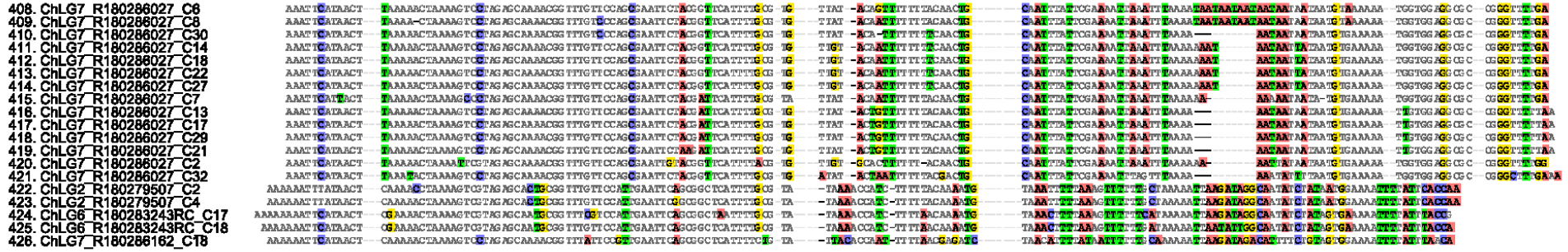



80. ChLG9_R180290569_C3
81. ChLG9_R180290569_C17
82. ChLG9_R180290569_C8
83. ChLG9_R180289936RC_C4
84. ChLG9_R180290569_C9
85. ChLG9_R180290569_C13
86. ChLG9_R180290569_C18
87. ChLG9_R180290569_C4
88. ChLG9_R180289936RC_C6
89. ChLG9_R180290569_C11
90. ChLG9_R180289936RC_C9
91. ChLG9_R180289936RC_C12
92. ChLG9_R180290569_C14
93. ChLG9_R180290569_C19
94. ChLG9_R180290569_C5
95. ChLG9_R180289936RC_C7
96. ChLG9_R180290569_C12
97. ChLG9_R180289936RC_C10
98. ChLG9_R180290569_C6
99. ChLG9_R180290569_C15
100. ChLG9_R180289936RC_C13
101. ChLG9_R180289936RC_C5
102. ChLG9_R180290569_C10
103. ChLG9_R180289936RC_C8
104. ChLG9_R180289936RC_C11
105. ChLG9_R180290079_C5
106. ChLG9_R180290079_C9
107. ChLG9_R180290079_C3
108. ChLG9_R180290079_C7
109. ChLG9_R180290079_C2
110. ChLG9_R180290079_C6
111. ChLG5_R180282447RC_C2
112. ChLG5_R180282447RC_C14
113. ChLG5_R180282447RC_C4
114. ChLG5_R180282447RC_C7
115. ChLG5_R180282447RC_C9
116. ChLG5_R180282447RC_C8
117. ChLG5_R180282447RC_C12
118. ChLG5_R180282447RC_C5
119. ChLG5_R180282447RC_C6
120. ChLG5_R180282447RC_C13
121. ChLG5_R180282447RC_C10
122. ChLG5_R180282447RC_C11
123. ChLG5_R180282447RC_C3
124. ChLG9_R180291487RC_C5
125. ChLG9_R180291490RC_C14
126. ChLG3_2_R180295309RC_C3
127. ChLG3_2_R180295309RC_...
128. ChLG3_2_R180295309RC_...
129. ChLG9_R180291487RC_C-5
130. ChLG9_R180291490RC_C6
131. ChLG9_R180291487RC_C-3
132. ChLG9_R180291490RC_C9
133. ChLG9_R180291487RC_C-2
134. ChLG9_R180291490RC_C8
135. ChLG9_R180291487RC_C-1
136. ChLG9_R180291490RC_C7
137. ChLG9_R180291487RC_C3
138. ChLG9_R180291490RC_C12
139. ChLG9_R180291487RC_C-6
140. ChLG9_R180291490RC_C4
141. ChLG9_R180291487RC_C6
142. ChLG9_R180291490RC_C15
143. ChLG9_R180291487RC_C8
144. ChLG9_R180291490RC_C17
145. ChLG9_R180291487RC_C-4
146. ChLG9_R180291490RC_C5
147. ChLG9_R180291487RC_C-8
148. ChLG9_R180291490RC_C2
149. ChLG9_R180291487RC_C7
150. ChLG9_R180291490RC_C16
151. ChLG9_R180291490RC_C3
152. ChLG9_R180291487RC_C-7
153. ChLG9_R180291487RC_C4
154. ChLG9_R180291490RC_C13
155. ChLG3_2_R180295309RC_C2
156. ChLG3_2_R180295309RC_C4
157. ChLG3_2_R180295309RC_C7
158. ChLG3_2_R180295309RC_C9
159. ChLG3_2_R180295309RC_C8
160. ChLG3_2_R180295309RC_C6
161. ChLG3_2_R180295309RC_C5

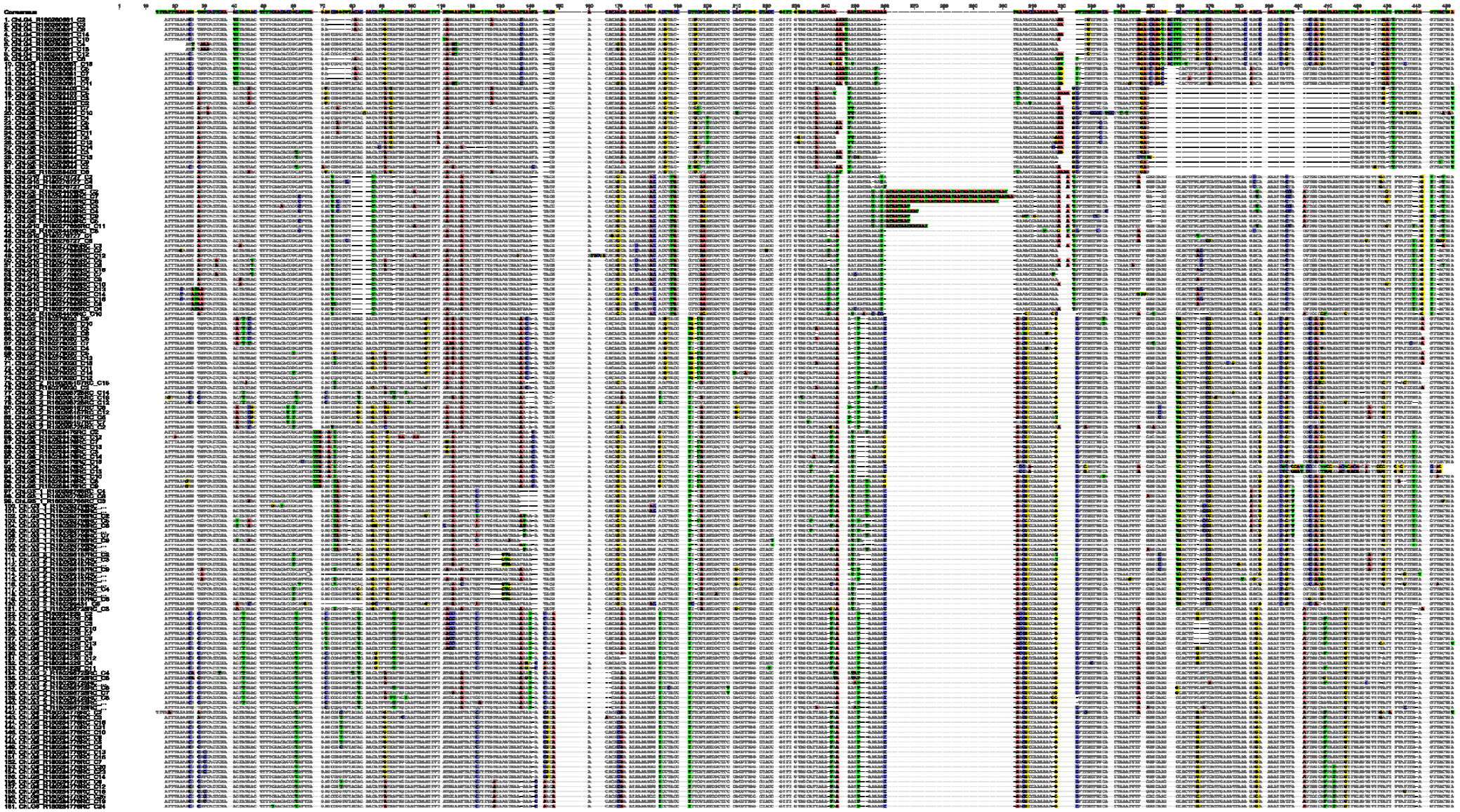

162. ChLG9_R180291487RC_C2
163. ChLG9_R180291490RC_C11
164. ChLG8_R180289343RC_C11
165. ChLG8_R180289344RC_C22
166. ChLG7_R180286162_C4
167. ChLG7_R180286162_C11
168. ChLG7_R180286162_C2
169. ChLG7_R180286162_C7
170. ChLG7_R180286162_C3
171. ChLG7_R180286162_C10
172. ChLG7_R180286162_C8
173. ChLG7_R180286162_C15
174. ChLG7_R180286162_C12
175. ChLG7_R180286162_C13
176. ChLG7_R180286162_C5
177. ChLG7_R180286162_C6
178. ChLG7_R180286162_C9
179. ChLG7_R180286162_C14
180. ChLG7_R180286162_C16
181. ChLG7_R180286162_C17
182. ChLG5_R180283147_C5
183. ChLG5_R180283147_C8
184. ChLG5_R180283147_C4
185. ChLG5_R180283147_C2
186. ChLG5_R180283147_C6
187. ChLG5_R180283147_C9
188. ChLG5_R180283147_C3
189. ChLG5_R180283147_C7
190. ChLG5_R180283147_C10
191. ChLG7_R180285990_C2
192. ChLG7_R180285990_C3
193. ChLG7_R180285990_C7
194. ChLG7_R180285990_C5
195. ChLG7_R180285990_C6
196. ChLG7_R180285990_C8
197. ChLG7_R180285676_C2
198. ChLG7_R180285677_C3
199. ChLG7_R180285677_C5
200. ChLG7_R180285677_C7
201. ChLG7_R180285677_C9
202. ChLG7_R180285677_C11
203. ChLG7_R180285676_C5
204. ChLG7_R180285676_C7
205. ChLG7_R180285676_C6
206. ChLG7_R180285676_C8
207. ChLG7_R180285676_C3
208. ChLG7_R180285676_C4
209. ChLG7_R180285677_C2
210. ChLG7_R180285677_C4
211. ChLG7_R180285677_C6
212. ChLG7_R180285677_C10
213. ChLG7_R180285677_C12
214. ChLG7_R180285677_C8
215. ChLG8_R180289322RC_C2
216. ChLG8_R180289322RC_C49
217. ChLG8_R180289322RC_C9
218. ChLG8_R180289322RC_C19
219. ChLG8_R180289322RC_C23
220. ChLG8_R180289322RC_C27
221. ChLG8_R180289322RC_C33
222. ChLG8_R180289322RC_C36
223. ChLG8_R180289322RC_C31
224. ChLG8_R180289322RC_C38
225. ChLG8_R180289322RC_C41
226. ChLG8_R180289322RC_C46
227. ChLG8_R180289322RC_C3
228. ChLG8_R180289322RC_C5
229. ChLG8_R180289322RC_C7
230. ChLG8_R180289322RC_C29
231. ChLG8_R180289322RC_C11
232. ChLG8_R180289322RC_C13
233. ChLG8_R180289322RC_C21
234. ChLG8_R180289322RC_C44
235. ChLG8_R180289322RC_C48
236. ChLG8_R180289322RC_C50
237. ChLG8_R180289322RC_C37
238. ChLG8_R180289322RC_C40
239. ChLG8_R180289322RC_C35
240. ChLG8_R180289322RC_C25
241. ChLG8_R180289322RC_C4
242. ChLG8_R180289322RC_C6
243. ChLG8_R180289322RC_C8

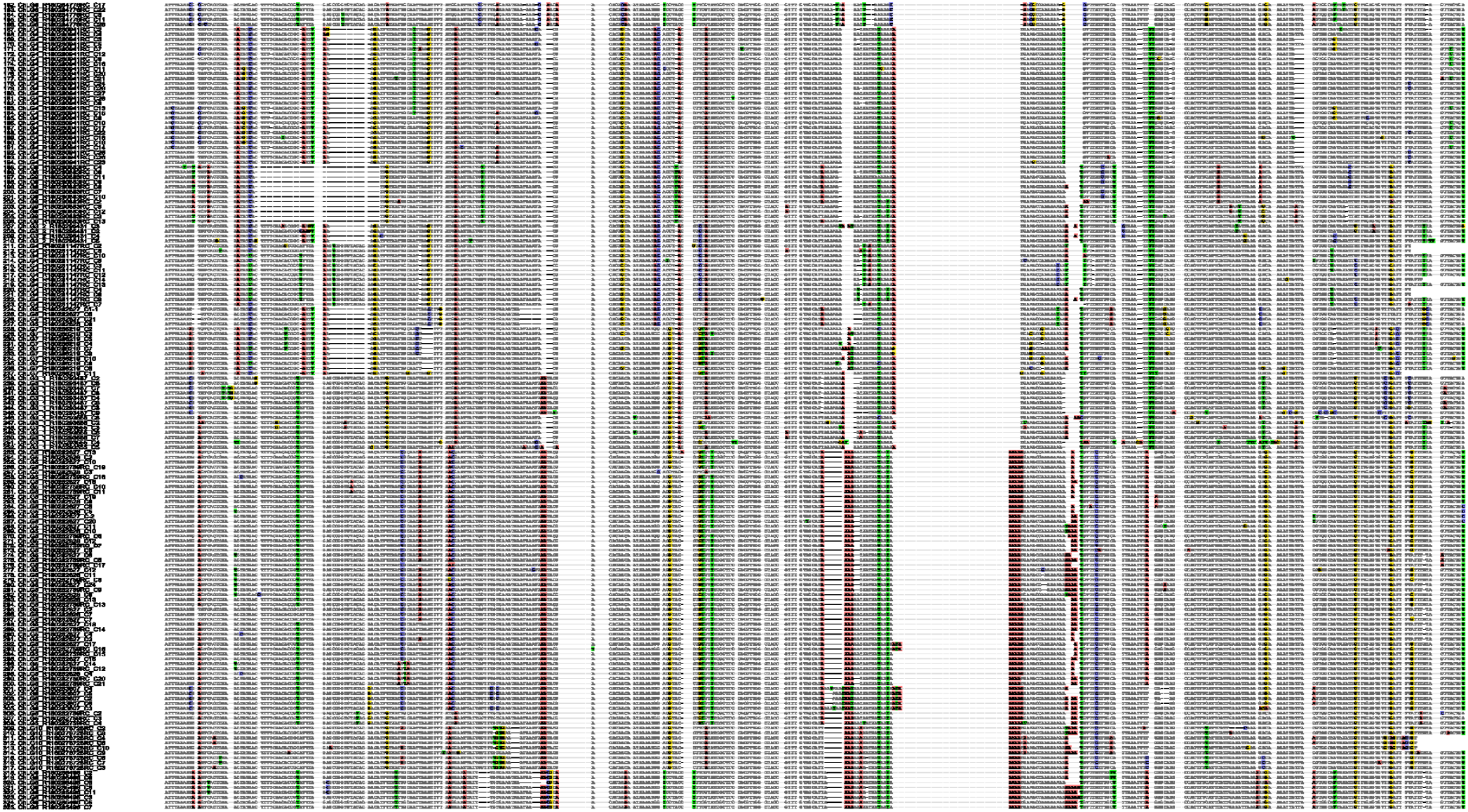

244. ChLG8_R180289322RC_C10
245. ChLG8_R180289322RC_C18
246. ChLG8_R180289322RC_C12
247. ChLG8_R180289322RC_C16
248. ChLG8_R180289322RC_C24
249. ChLG8_R180289322RC_C28
250. ChLG8_R180289322RC_C20
251. ChLG8_R180289322RC_C14
252. ChLG8_R180289322RC_C22
253. ChLG8_R180289322RC_C34
254. ChLG8_R180289322RC_C39
255. ChLG8_R180289322RC_C43
256. ChLG8_R180289322RC_C30
257. ChLG8_R180289322RC_C32
258. ChLG8_R180289322RC_C26
259. ChLG8_R180289322RC_C45
260. ChLG8_R180289322RC_C47
261. ChLG8_R180289322RC_C42
262. ChLG7_R180285685RC_C8
263. ChLG7_R180285685RC_C9
264. ChLG7_R180285685RC_C10
265. ChLG7_R180285685RC_C2
266. ChLG7_R180285685RC_C4
267. ChLG7_R180285685RC_C5
268. ChLG7_R180285685RC_C11
269. ChLG7_R180285685RC_C7
270. ChLG7_R180285685RC_C12
271. ChLG7_R180285685RC_C3
272. ChLG7_R180285685RC_C8
273. ChLG7_R180285685RC_C13
274. ChLG7_R180286088_C2
275. ChLG7_R180286088_C4
276. ChLG7_R180286088_C8
277. ChLG7_R180286088_C5
278. ChLG7_R180286088_C3
279. ChLG7_R180286088_C7
280. ChLG7_R180286088_C6
281. ChLG7_R180286088_C9
282. ChLG7_R180286088_C10
283. ChLG8_R180289409RC_C2
284. ChLG8_R180289409RC_C18
285. ChLG8_R180289409RC_C9
286. ChLG8_R180289409RC_C13
287. ChLG8_R180289409RC_C6
288. ChLG8_R180289409RC_C17
289. ChLG8_R180289409RC_C7
290. ChLG8_R180289409RC_C10
291. ChLG8_R180289409RC_C14
292. ChLG8_R180289409RC_C3
293. ChLG8_R180289409RC_C4
294. ChLG8_R180289409RC_C8
295. ChLG8_R180289409RC_C11
296. ChLG8_R180289409RC_C15
297. ChLG8_R180289409RC_C5
298. ChLG8_R180289409RC_C12
299. ChLG8_R180289409RC_C16
300. ChLG8_R180289892RC_C4
301. ChLG8_R180289892RC_C5
302. ChLG8_R180289892RC_C6
303. ChLG8_R180289892RC_C9
304. ChLG8_R180289892RC_C7
305. ChLG8_R180289892RC_C8
306. ChLG8_R180289892RC_C2
307. ChLG8_R180289437_C2
308. ChLG8_R180289437_C13
309. ChLG8_R180289437_C7
310. ChLG8_R180289437_C10
311. ChLG8_R180289437_C5
312. ChLG8_R180289437_C8
313. ChLG8_R180289437_C11
314. ChLG8_R180289437_C9
315. ChLG8_R180289437_C12
316. ChLG8_R180289437_C3
317. ChLG8_R180289437_C4
318. ChLG8_R180289437_C6
319. ChLG8_R180289437_C1
320. ChLG3_1_R180292289_C5
321. ChLG3_1_R180292289_C6
322. ChLG3_1_R180292289_C7
323. ChLG3_1_R180292289_C4
324. ChLG3_1_R180292289_C2
325. ChLG3_1_R180292289_C3

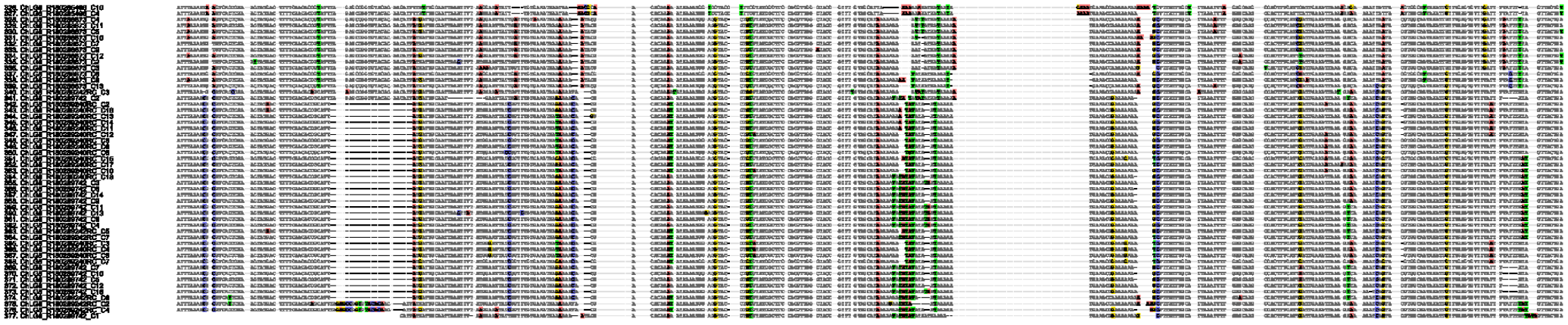

326. ChLG3_2_R180295520RC_C2
327. ChLG3_2_R180295520RC_C6
328. ChLG3_2_R180295520RC_C4
329. ChLG3_2_R180295520RC_C7
330. ChLG3_2_R180295520RC_C5
331. ChLG3_2_R180295520RC_C3
332. ChLG6_R180283243RC_C3
333. ChLG6_R180283243RC_C6
334. ChLG6_R180283243RC_C12
335. ChLG6_R180283243RC_C21
336. ChLG6_R180283243RC_C10
337. ChLG6_R180283243RC_C15
338. ChLG6_R180283243RC_C19
339. ChLG6_R180283243RC_C5
340. ChLG6_R180283243RC_C16
341. ChLG6_R180283243RC_C20
342. ChLG6_R180283243RC_C7
343. ChLG6_R180283243RC_C4
344. ChLG6_R180283243RC_C8
345. ChLG6_R180283243RC_C22
346. ChLG6_R180283338_C8
347. ChLG6_R180283243RC_C11
348. ChLG6_R180283243RC_C13
349. ChLG6_R180283243RC_C14
350. ChLG6_R180283243RC_C23
351. ChLG6_R180283243RC_C24
352. ChLG6_R180283243RC_C25
353. ChLG6_R180283338_C3
354. ChLG6_R180283338_C5
355. ChLG6_R180283338_C6
356. ChLG6_R180283338_C11
357. ChLG6_R180283338_C4
358. ChLG6_R180283338_C7
359. ChLG6_R180283338_C9
360. ChLG6_R180283338_C1
361. ChLG6_R180283338_C2
362. ChLG6_R180283338_C10
363. ChLG6_R180283243RC_C9
364. ChLG6_R180283243RC_C2
365. ChLG6_R180283388_C7
366. ChLG6_R180283388_C11
367. ChLG6_R180283388_C4
368. ChLG6_R180283388_C6
369. ChLG6_R180283388_C10
370. ChLG6_R180283388_C8
371. ChLG6_R180283388_C9
372. ChLG6_R180283388_C5
373. ChLG6_R180283388_C14
374. ChLG6_R180283388_C12
375. ChLG6_R180283388_C13
376. ChLG6_R180283388_C3
377. ChLG6_R180283388_C2
378. ChLG3_2_R180296704_C3
379. ChLG3_2_R180296704_C5
380. ChLG3_2_R180296704_C4
381. ChLG3_2_R180296704_C6
382. ChLG3_2_R180296704_C2
383. ChLG3_2_R180296704_C7
384. ChLG3_2_R180296704_C8
385. ChLG7_R180285984_C2
386. ChLG7_R180285984_C4
387. ChLG7_R180285984_C7
388. ChLG7_R180285984_C3
389. ChLG7_R180285984_C6
390. ChLG7_R180285984_C5
391. ChLG7_R180286027_C5
392. ChLG7_R180286027_C10
393. ChLG7_R180286027_C3
394. ChLG7_R180286027_C15
395. ChLG7_R180286027_C19
396. ChLG7_R180286027_C23
397. ChLG7_R180286027_C11
398. ChLG7_R180286027_C16
399. ChLG7_R180286027_C24
400. ChLG7_R180286027_C12
401. ChLG7_R180286027_C20
402. ChLG7_R180286027_C28
403. ChLG7_R180286027_C25
404. ChLG7_R180286027_C31
405. ChLG7_R180286027_C26
406. ChLG7_R180286027_C4
407. ChLG7_R180286027_C9


408. ChLG7_R180286027_C6
409. ChLG7_R180286027_C8
410. ChLG7_R180286027_C30
411. ChLG7_R180286027_C14
412. ChLG7_R180286027_C18
413. ChLG7_R180286027_C22
414. ChLG7_R180286027_C27
415. ChLG7_R180286027_C7
416. ChLG7_R180286027_C13
417. ChLG7_R180286027_C17
418. ChLG7_R180286027_C29
419. ChLG7_R180286027_C21
420. ChLG7_R180286027_C2
421. ChLG7_R180286027_C32
422. ChLG2_R180279507_C2
423. ChLG2_R180279507_C4
424. ChLG6_R180283243RC_C17
425. ChLG6_R180283243RC_C18
426. ChLG7_R180286162_C18

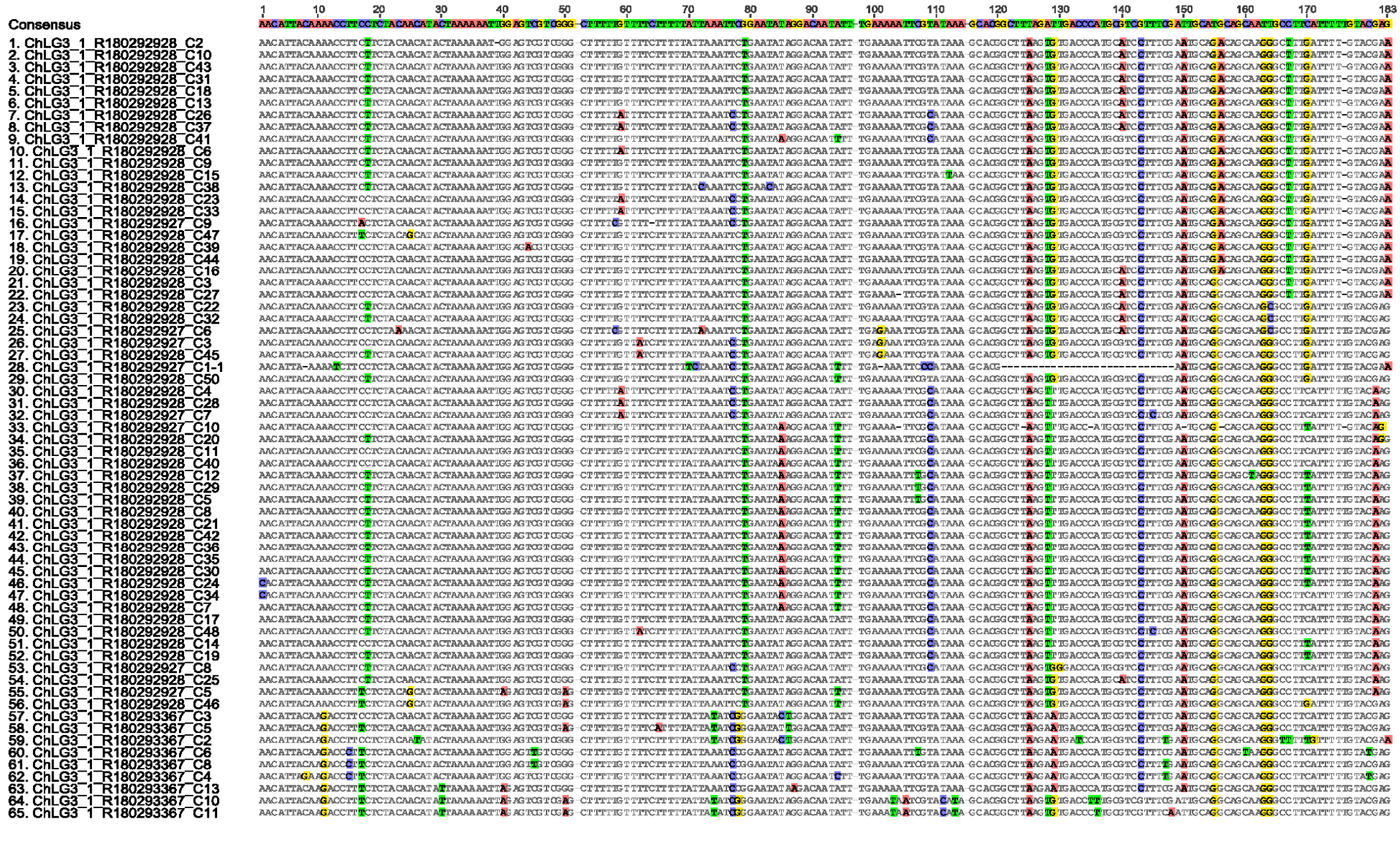
e) alignment of cluster 5

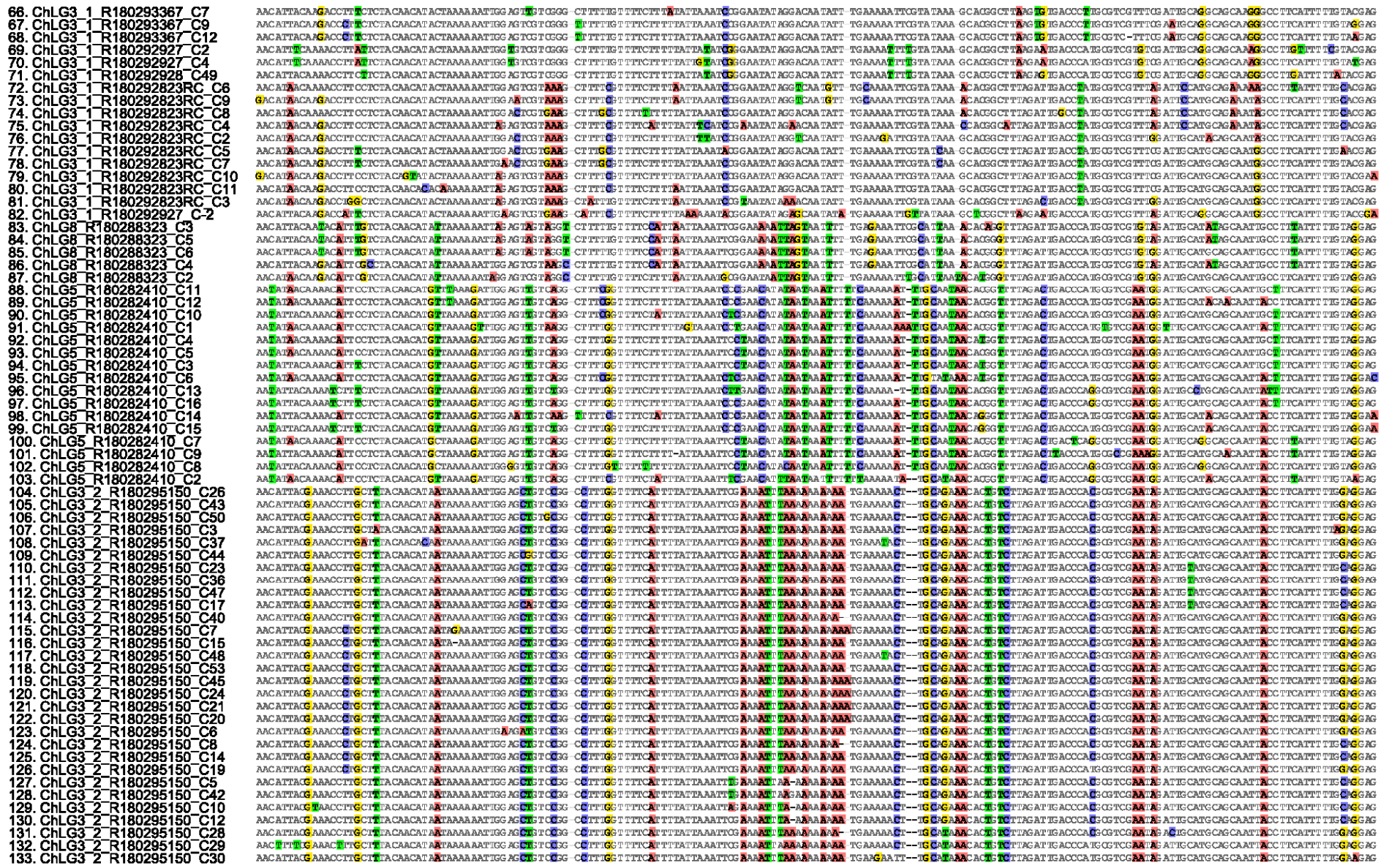




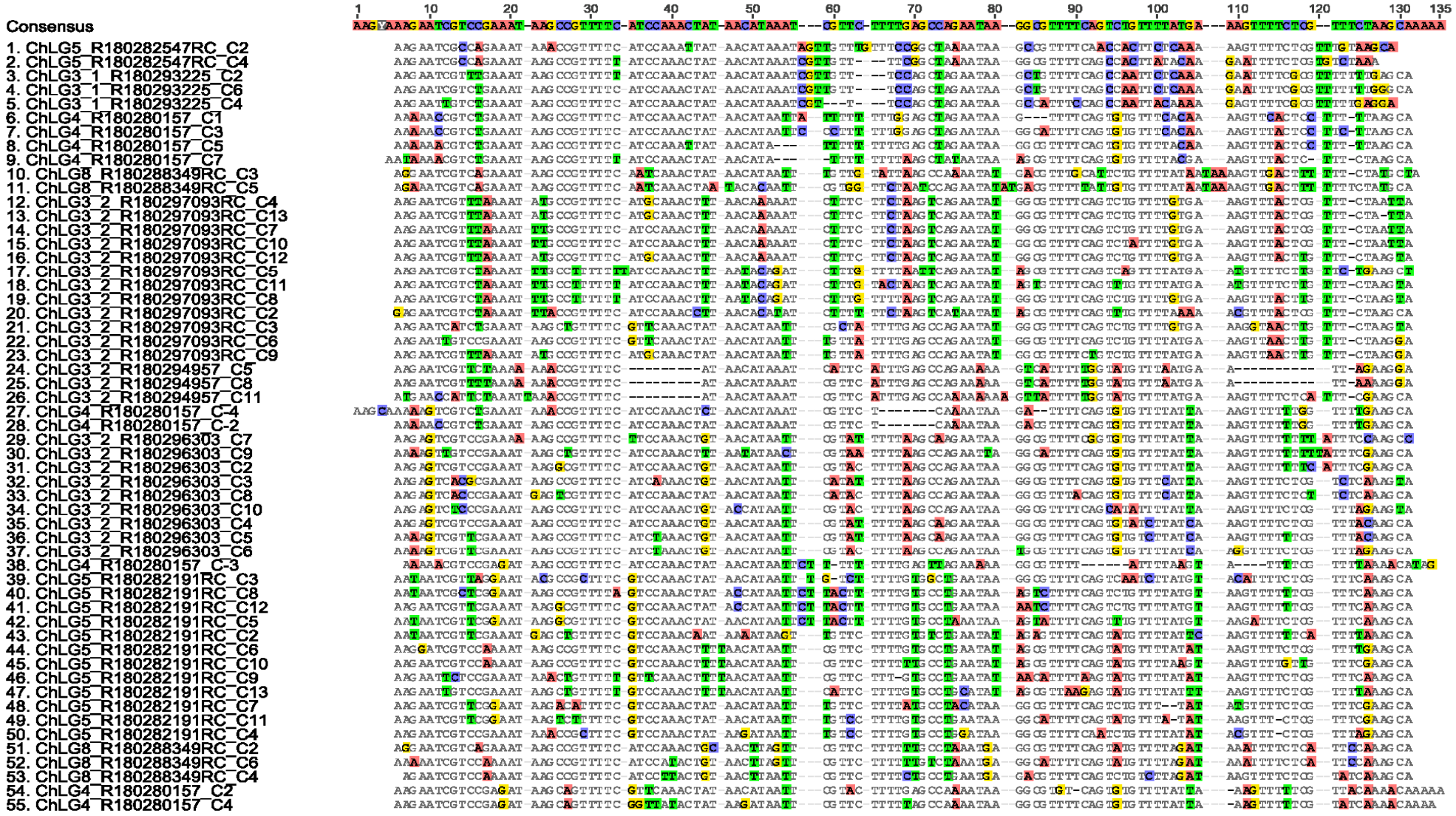


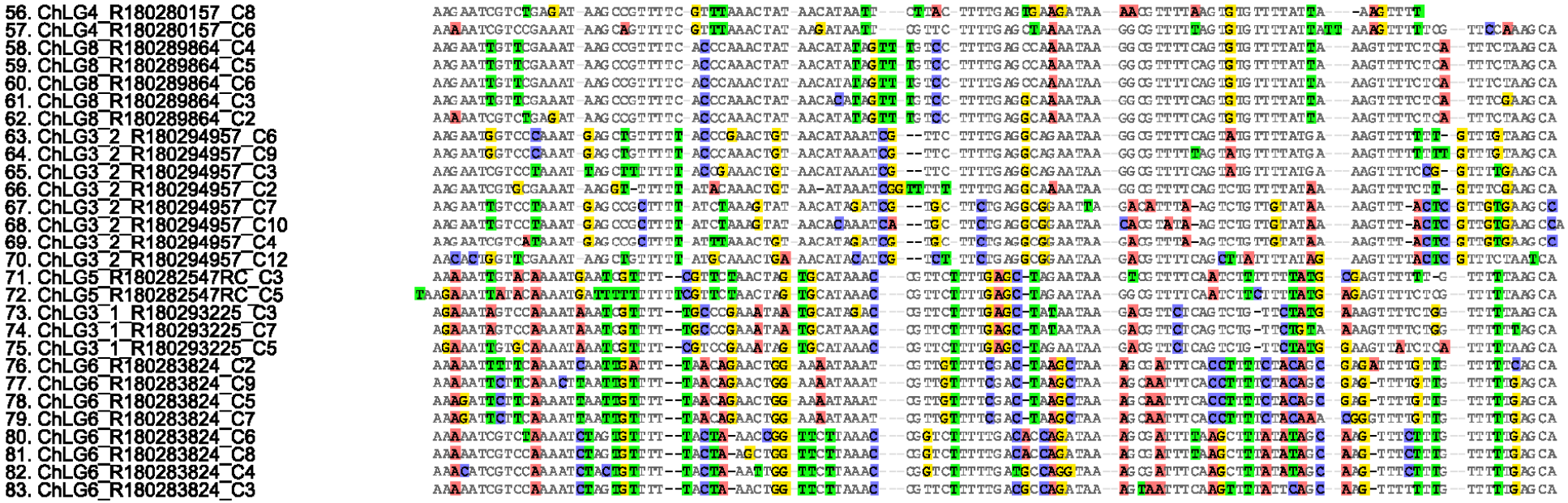
f) alignment of cluster 7



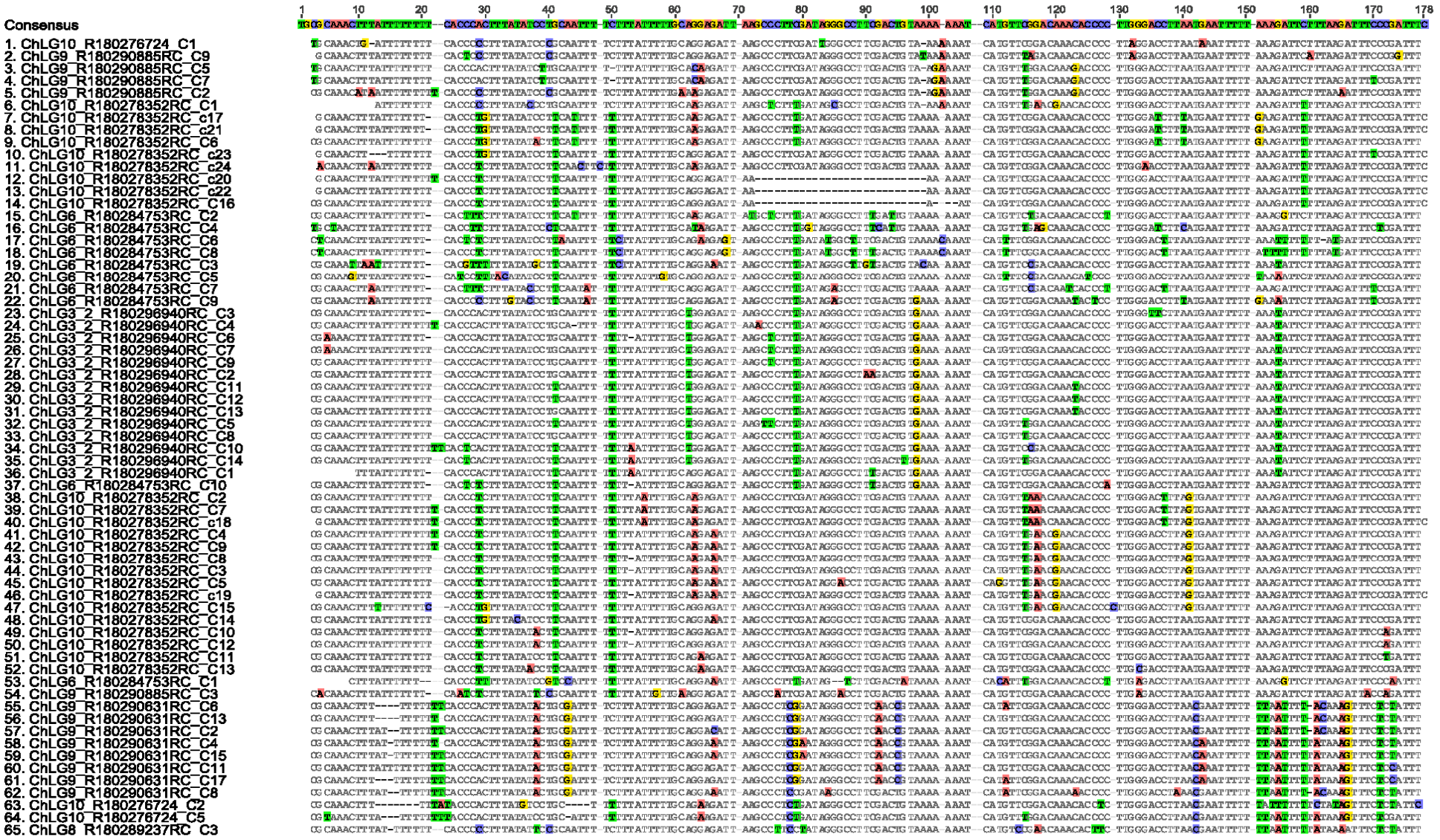


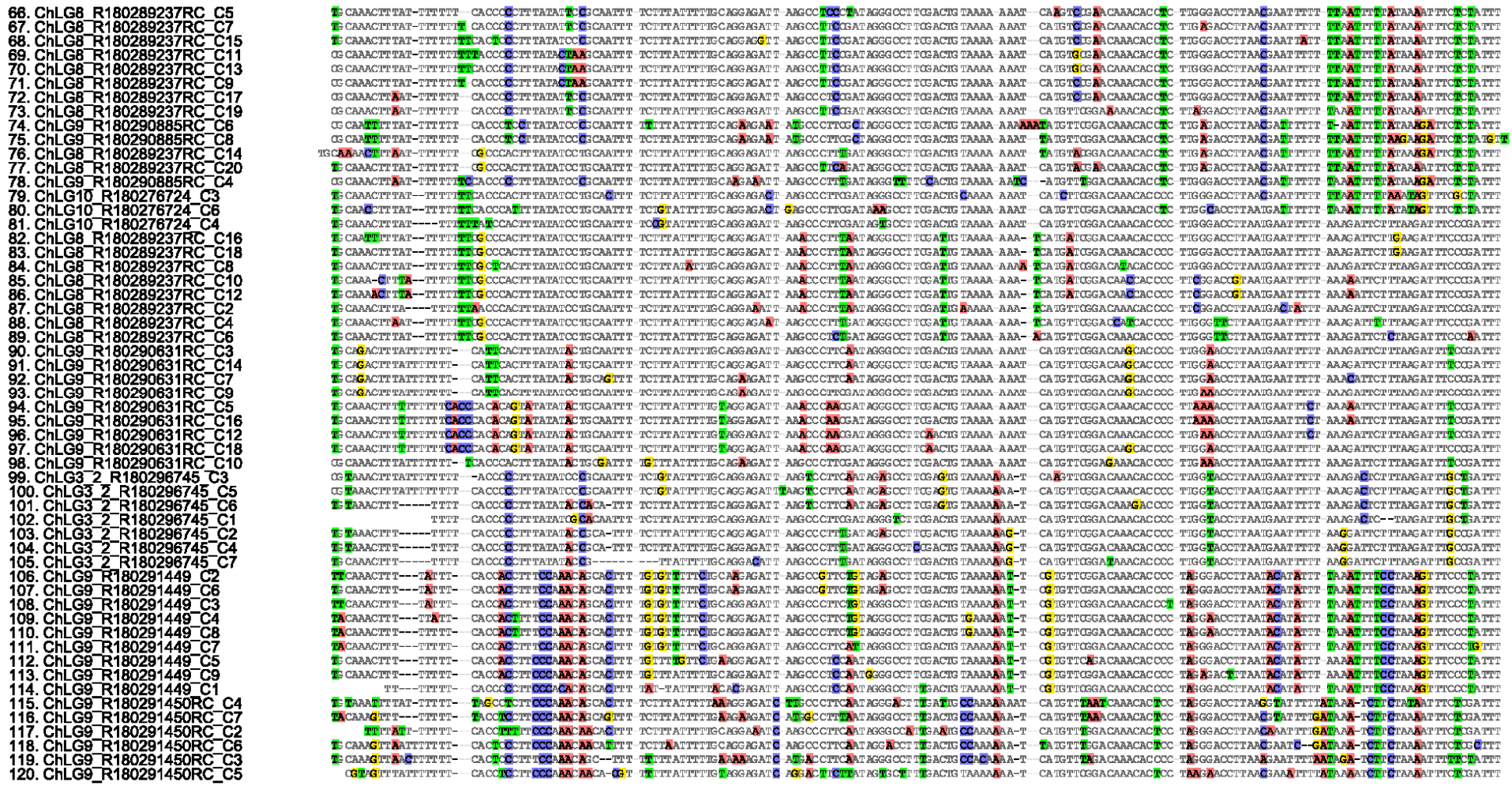
g) alignment of cluster 8


Consensus
1. ChLG5_R180282547RC_C2
2. ChLG5_R180282547RC_C4
3. ChLG3_1_R180293225_C2
4. ChLG3_1_R180293225_C6
5. ChLG3_1_R180293225_C4
6. ChLG4_R180280157_C1
7. ChLG4_R180280157_C3
8. ChLG4_R180280157_C5
9. ChLG4_R180280157_C7
10. ChLG8_R180288349RC_C3
11. ChLG8_R180288349RC_C5
12. ChLG3_2_R180297093RC_C4
13. ChLG3_2_R180297093RC_C13
14. ChLG3_2_R180297093RC_C7
15. ChLG3_2_R180297093RC_C10
16. ChLG3_2_R180297093RC_C12
17. ChLG3_2_R180297093RC_C5
18. ChLG3_2_R180297093RC_C11
19. ChLG3_2_R180297093RC_C8
20. ChLG3_2_R180297093RC_C2
21. ChLG3_2_R180297093RC_C3
22. ChLG3_2_R180297093RC_C6
23. ChLG3_2_R180297093RC_C9
24. ChLG3_2_R180294957_C5
25. ChLG3_2_R180294957_C8
26. ChLG3_2_R180294957_C11
27. ChLG4_R180280157_C-4
28. ChLG4_R180280157_C-2
29. ChLG3_2_R180296303_C7
30. ChLG3_2_R180296303_C9
31. ChLG3_2_R180296303_C2
32. ChLG3_2_R180296303_C3
33. ChLG3_2_R180296303_C8
34. ChLG3_2_R180296303_C10
35. ChLG3_2_R180296303_C4
36. ChLG3_2_R180296303_C5
37. ChLG3_2_R180296303_C6
38. ChLG4_R180280157_C-3
39. ChLG5_R180282191RC_C3
40. ChLG5_R180282191RC_C8
41. ChLG5_R180282191RC_C12
42. ChLG5_R180282191RC_C5
43. ChLG5_R180282191RC_C2
44. ChLG5_R180282191RC_C6
45. ChLG5_R180282191RC_C10
46. ChLG5_R180282191RC_C9
47. ChLG5_R180282191RC_C13
48. ChLG5_R180282191RC_C7
49. ChLG5_R180282191RC_C11
50. ChLG5_R180282191RC_C4
51. ChLG8_R180288349RC_C2
52. ChLG8_R180288349RC_C6
53. ChLG8_R180288349RC_C4
54. ChLG4_R180280157_C2
55. ChLG4_R180280157_C4

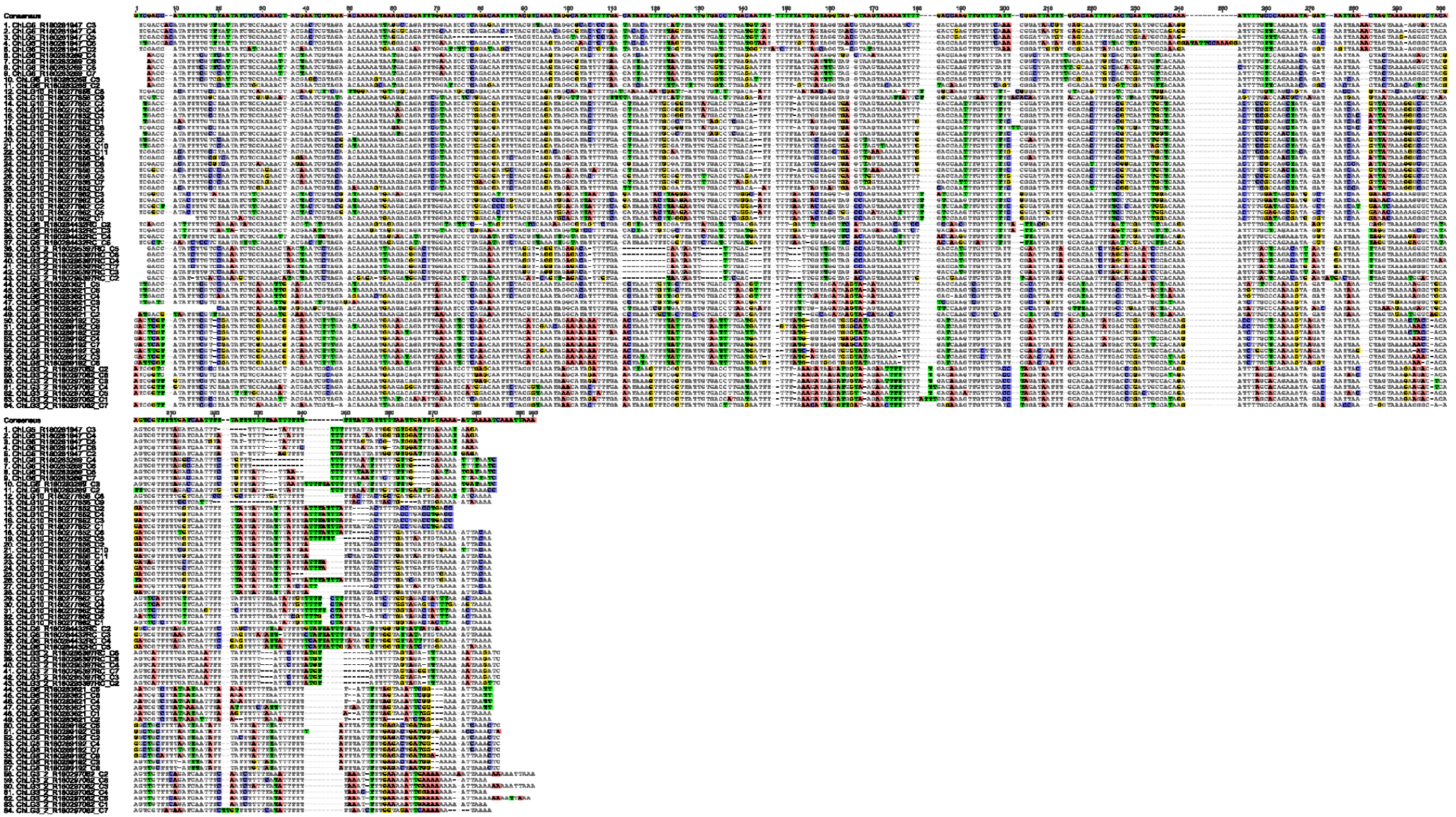

56. ChLG4_R180280157_C8
57. ChLG4_R180280157_C6
58. ChLG8_R180289864_C4
59. ChLG8_R180289864_C5
60. ChLG8_R180289864_C6
61. ChLG8_R180289864_C3
62. ChLG8_R180289864_C2
63. ChLG3_2_R180294957_C6
64. ChLG3_2_R180294957_C9
65. ChLG3_2_R180294957_C3
66. ChLG3_2_R180294957_C2
67. ChLG3_2_R180294957_C7
68. ChLG3_2_R180294957_C10
69. ChLG3_2_R180294957_C4
70. ChLG3_2_R180294957_C12
71. ChLG5_R180282547RC_C3
72. ChLG5_R180282547RC_C5
73. ChLG3_1_R180293225_C3
74. ChLG3_1_R180293225_C7
75. ChLG3_1_R180293225_C5
76. ChLG6_R180283824_C2
77. ChLG6_R180283824_C9
78. ChLG6_R180283824_C5
79. ChLG6_R180283824_C7
80. ChLG6_R180283824_C6
81. ChLG6_R180283824_C8
82. ChLG6_R180283824_C4
83. ChLG6_R180283824_C3

h) alignment of cluster 9



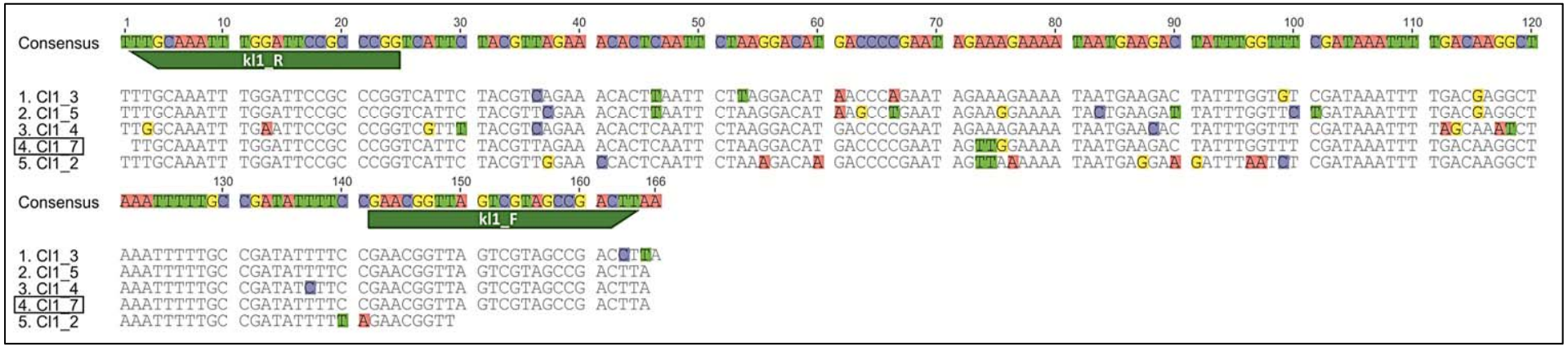

66. ChLG8_R180289237RC_C5
67. ChLG8_R180289237RC_C7
68. ChLG8_R180289237RC_C15
69. ChLG8_R180289237RC_C11
70. ChLG8_R180289237RC_C13
71. ChLG8_R180289237RC_C9
72. ChLG8_R180289237RC_C17
73. ChLG8_R180289237RC_C19
74. ChLG9_R180290885RC_C6
75. ChLG9_R180290885RC_C8
76. ChLG8_R180289237RC_C14
77. ChLG8_R180289237RC_C20
78. ChLG9_R180290885RC_C4
79. ChLG10_R180276724_C3
80. ChLG10_R180276724_C6
81. ChLG10_R180276724_C4
82. ChLG8_R180289237RC_C16
83. ChLG8_R180289237RC_C18
84. ChLG8_R180289237RC_C8
85. ChLG8_R180289237RC_C10
86. ChLG8_R180289237RC_C12
87. ChLG8_R180289237RC_C2
88. ChLG8_R180289237RC_C4
89. ChLG8_R180289237RC_C6
90. ChLG9_R180290631RC_C3
91. ChLG9_R180290631RC_C14
92. ChLG9_R180290631RC_C7
93. ChLG9_R180290631RC_C9
94. ChLG9_R180290631RC_C5
95. ChLG9_R180290631RC_C16
96. ChLG9_R180290631RC_C12
97. ChLG9_R180290631RC_C18
98. ChLG9_R180290631RC_C10
99. ChLG3_2_R180296745_C3
100. ChLG3_2_R180296745_C5
101. ChLG3_2_R180296745_C6
102. ChLG3_2_R180296745_C1
103. ChLG3_2_R180296745_C2
104. ChLG3_2_R180296745_C4
105. ChLG3_2_R180296745_C7
106. ChLG9_R180291449_C2
107. ChLG9_R180291449_C6
108. ChLG9_R180291449_C3
109. ChLG9_R180291449_C4
110. ChLG9_R180291449_C8
111. ChLG9_R180291449_C7
112. ChLG9_R180291449_C5
113. ChLG9_R180291449_C9
114. ChLG9_R180291449_C1
115. ChLG9_R180291450RC_C4
116. ChLG9_R180291450RC_C7
117. ChLG9_R180291450RC_C2
118. ChLG9_R180291450RC_C6
119. ChLG9_R180291450RC_C3
120. ChLG9_R180291450RC_C5

i) alignment of cluster 10

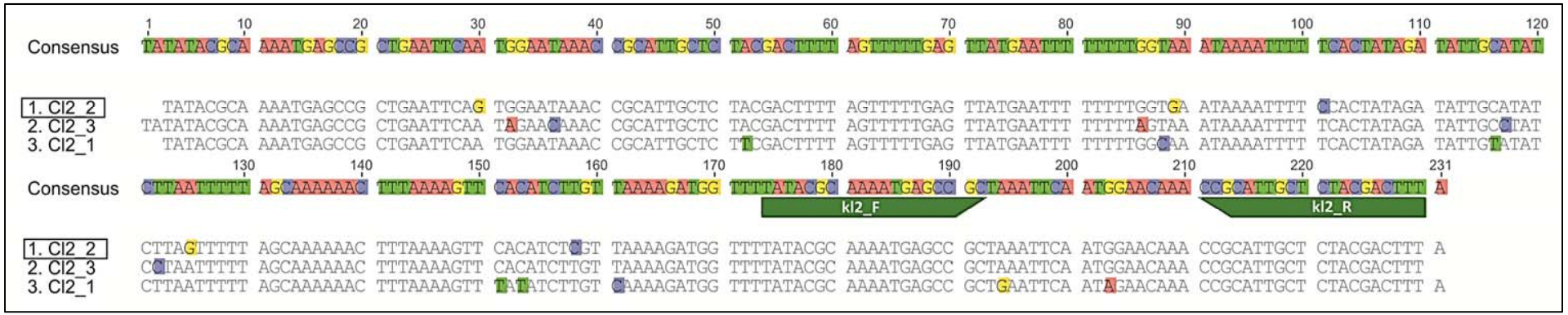


Supplementary Figure 4.2.3. Alignments of all cloned fragments obtained by PCR. **a)** Clones obtained with primers kl1_F and kl1_R; **b)** Clones obtained with primers kl2_F and kl2_R; **c)** Clones obtained with primers kl3_F and kl3_R; **d)** Clones obtained with primers kl4_F and kl4_R; **e)** Clones obtained with primers kl5_F and kl5_R; **f)** Clones obtained with primers kl7_F and kl7_R; **g)** Clones obtained with primers kl8_F and kl8_R; **h)** Clones obtained with primers kl9_F and kl9_R; **i)** Clones obtained with primers kl10_F and kl10_R

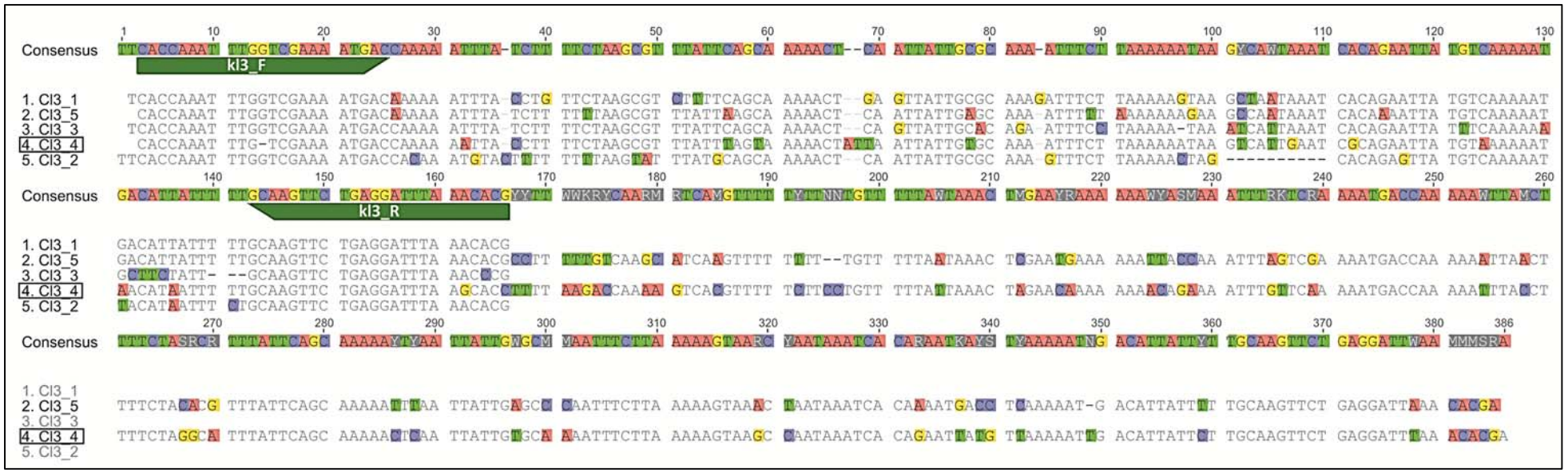
a) Clones obtained with primers kl1_F and kl1_R



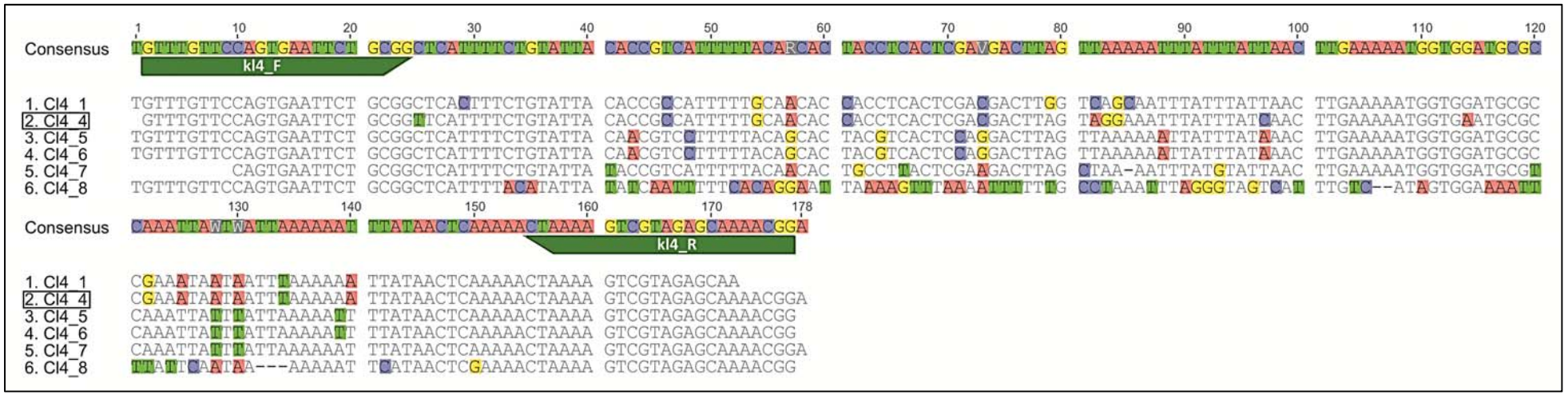
b) Clones obtained with primers kl2_F and kl2_R



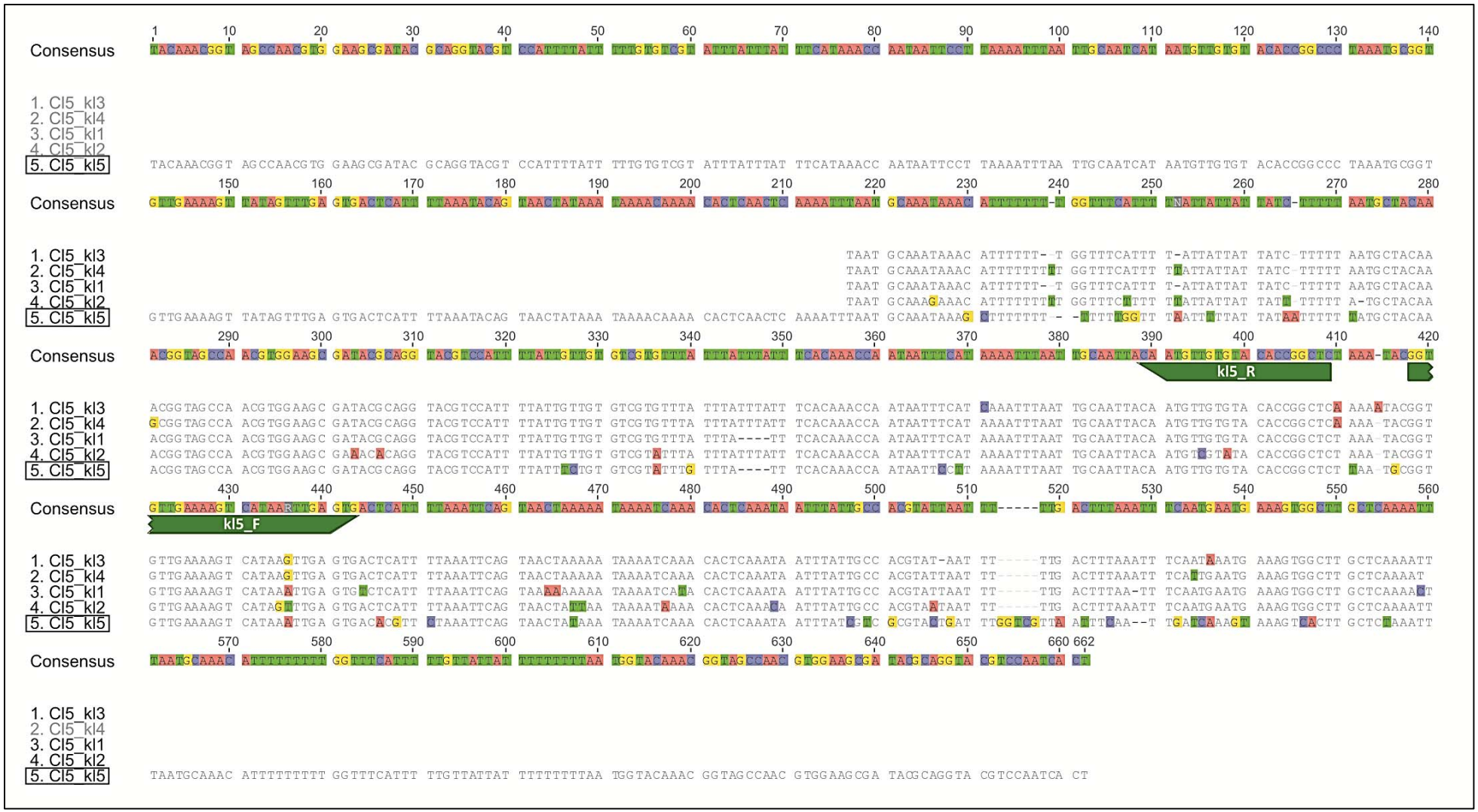
c) Clones obtained with primers kl3_F and kl3_R



d) Clones obtained with primers kl4_F and kl4_R

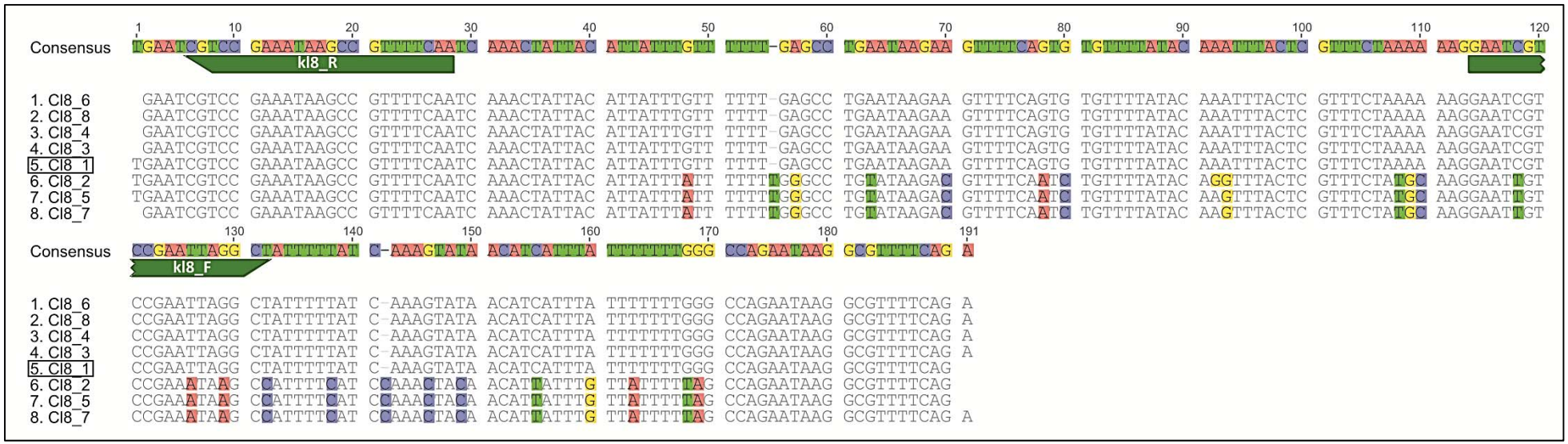


e) Clones obtained with primers kl5_F and kl5_R

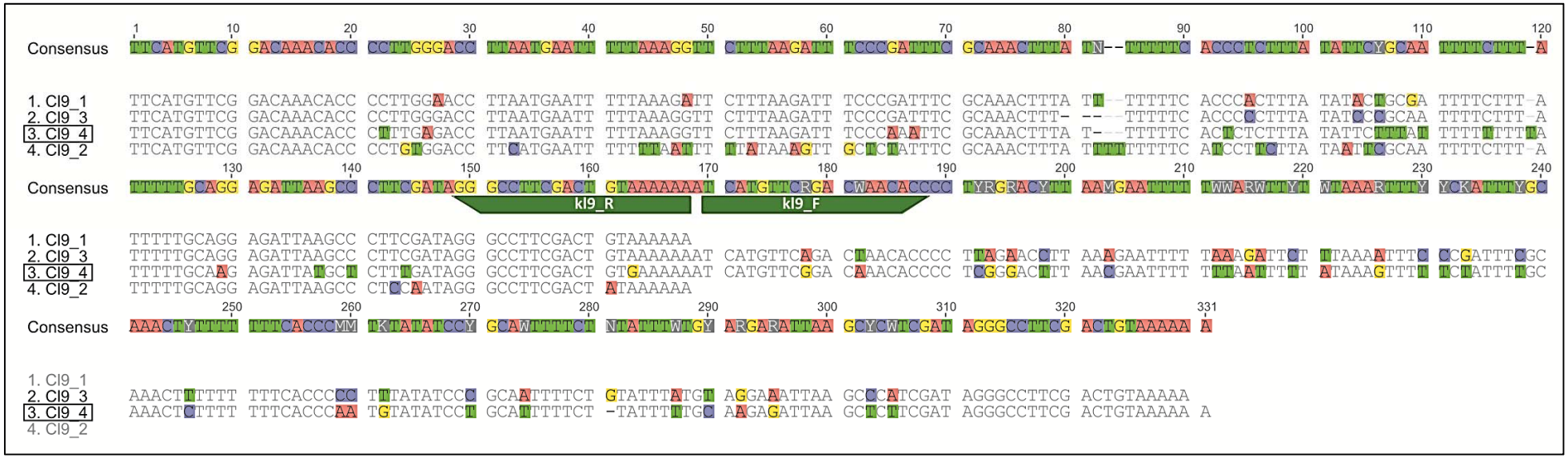


f) Clones obtained with primers kl7_F and kl7_R

g) Clones obtained with primers kl8_F and kl8_R



h) Clones obtained with primers kl9_F and kl9_R

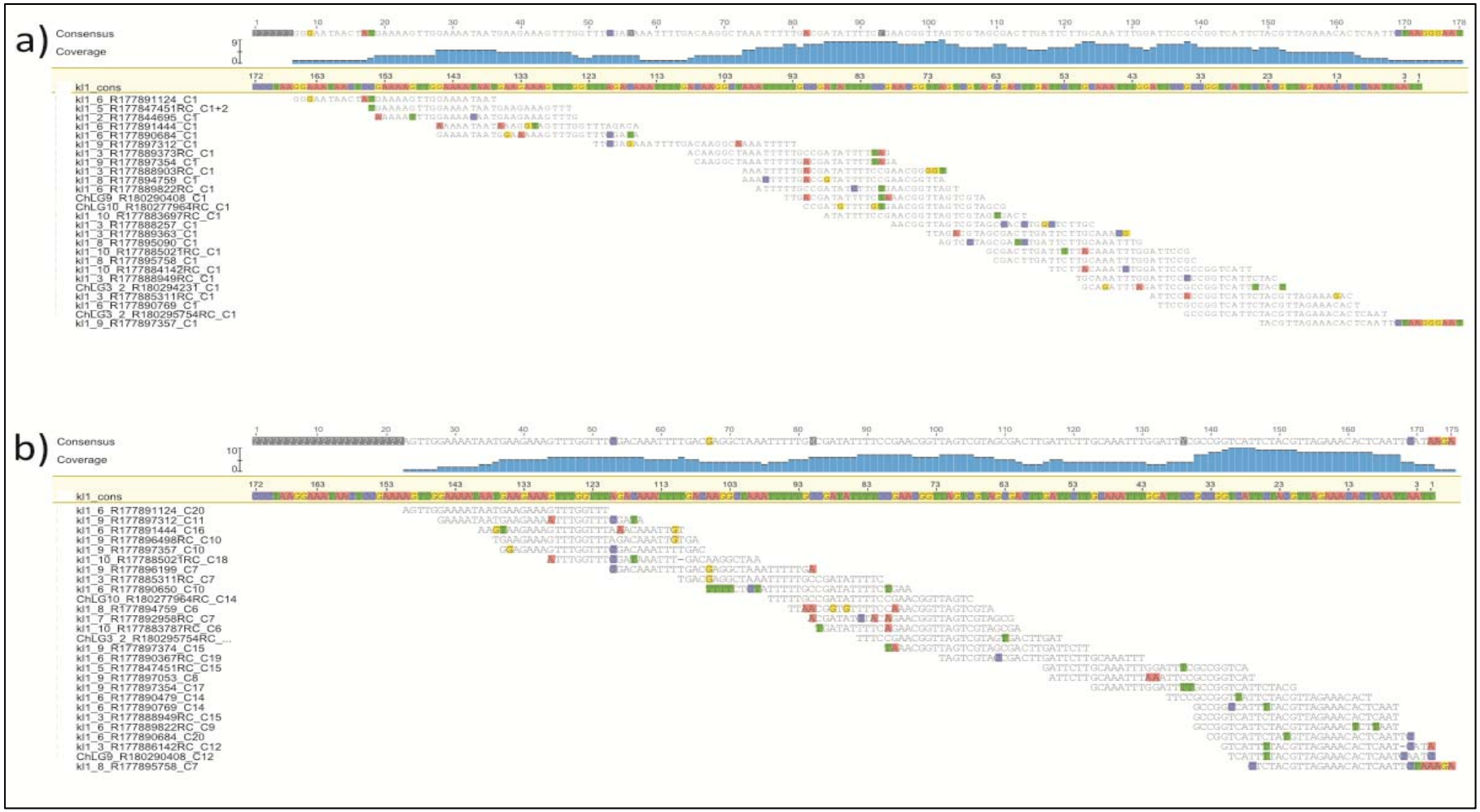


i) Clones obtained with primers kl10_F and kl10_R

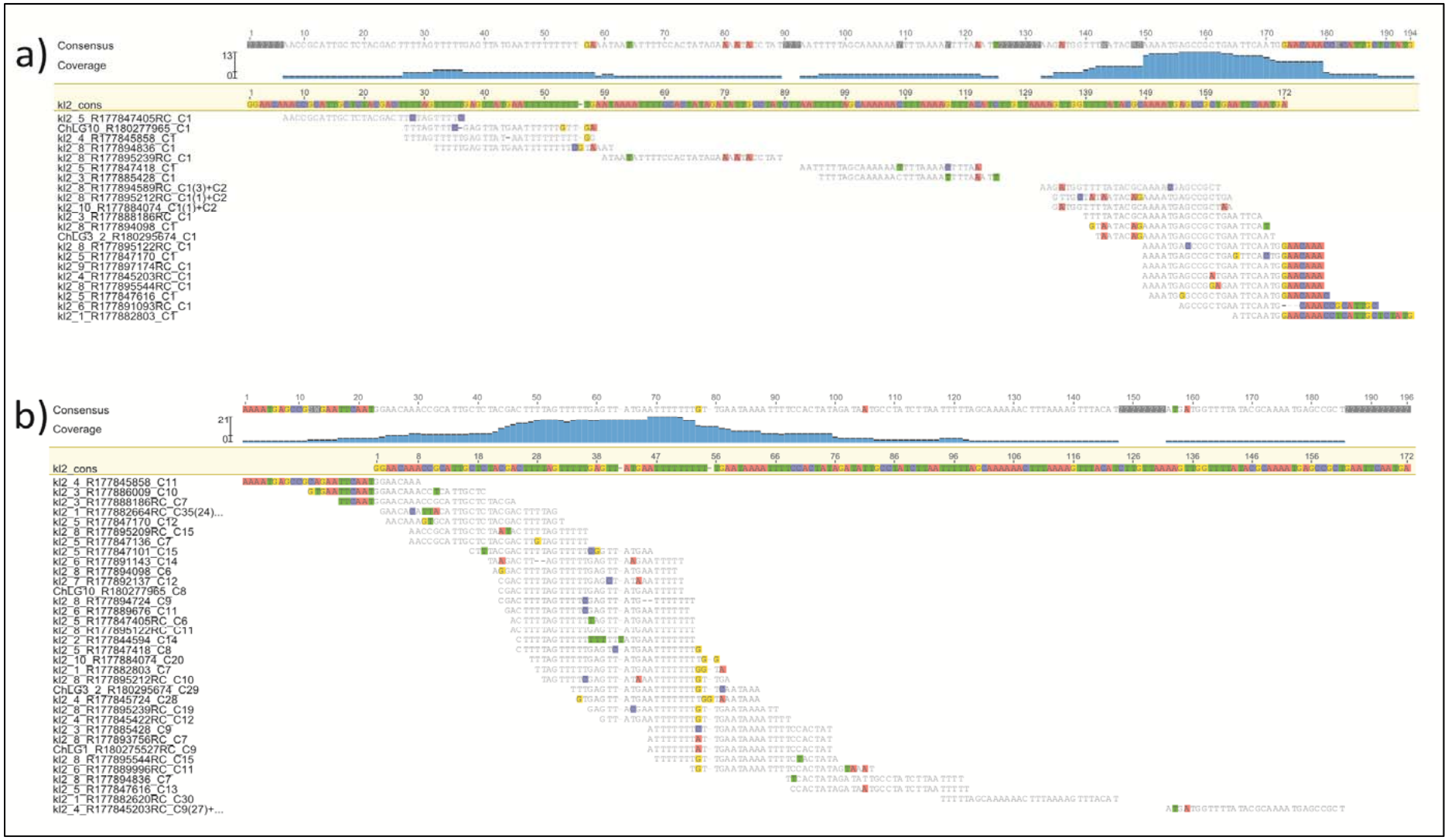
Supplementary Figure 4.2.4. Table with data for 7 clusters (marked uCl) obtained by TRF and array clustering on unassembled reads of *T. castaneum* genome. TRF arameters were the same as for the 10 chromosomes.

	Number of arrays	The average length of monomers (bp)
uCl1=Cl5	34	309-339
Cul2=Cl7	27	179-183
uCl3=Cl1	13	166-167
uCl4=Cl2	13	169-175
uCl5	11	222-224
uCl6=TCAST	9	331-337
uCl7	9	107-125

Supplementary Figure 4.2.5. Position of insertion sites of all arrays in Cluster 1 compared to consensus sequence. a) position of first 30 pb of each of the arrays in Cluster 1. b) position of last 30 pb of each of the arrays in Cluster 1.



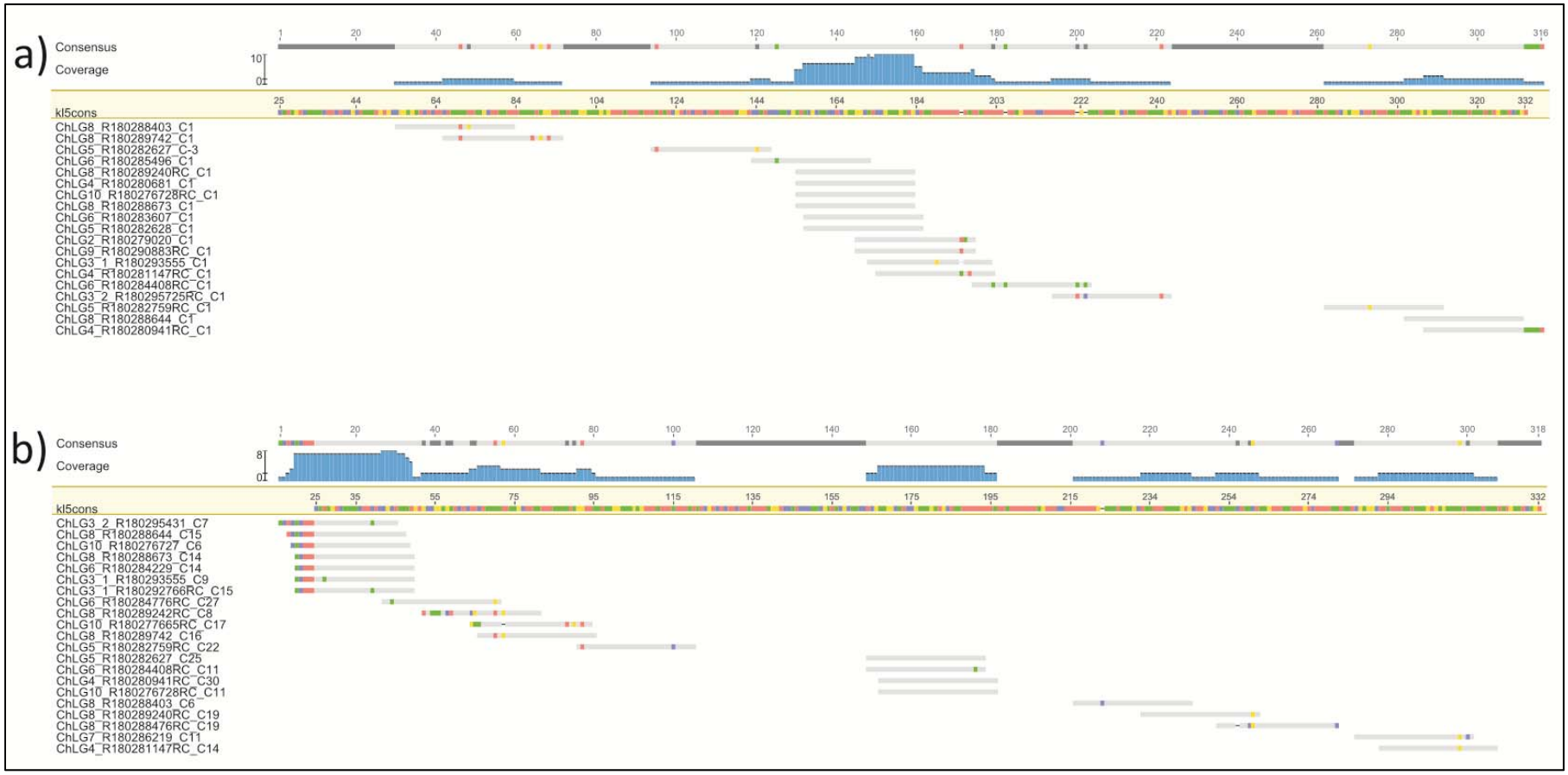
Supplementary Figure 4.2.6. Position of insertion sites of all arrays in Cluster 2 compared to consensus sequence. a) position of first 30 pb of each of the arrays in Cluster 2. b) position of last 30 pb of each of the arrays in Cluster 2.



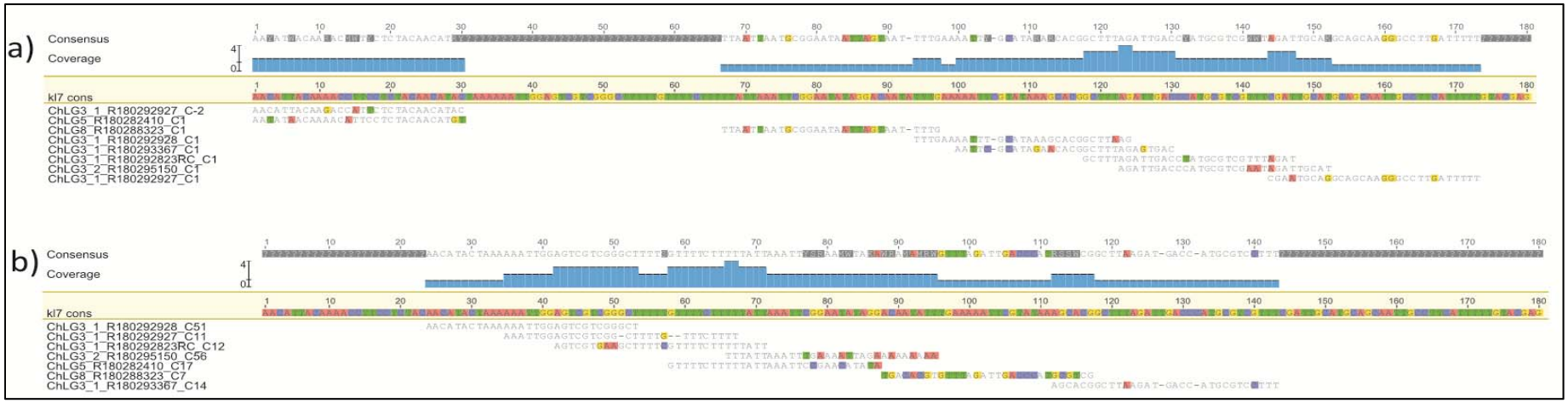
Supplementary Figure 4.2.7. Position of insertion sites of all arrays in Cluster 3 compared to consensus sequence. a) position of first 30 pb of each of the arrays in Cluster 3. b) position of last 30 pb of each of the arrays in Cluster 3.

Supplementary Figure 4.2.8. Position of insertion sites of all arrays in Cluster 4 compared to consensus sequence. a) position of first 30 pb of each of the arrays in Cluster 4. b) position of last 30 pb of each of the arrays in Cluster 4.

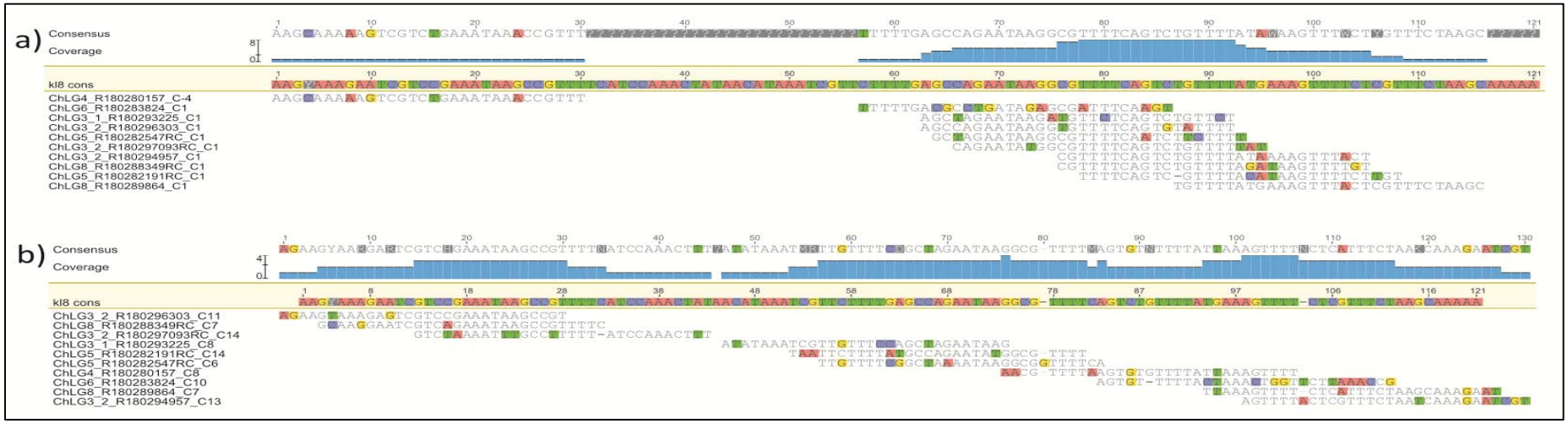
Supplementary Figure 4.2.9. Position of insertion sites of all arrays in Cluster 5 compared to consensus sequence. a) position of first 30 pb of each of the arrays in Cluster 5. b) position of last 30 pb of each of the arrays in Cluster 5.



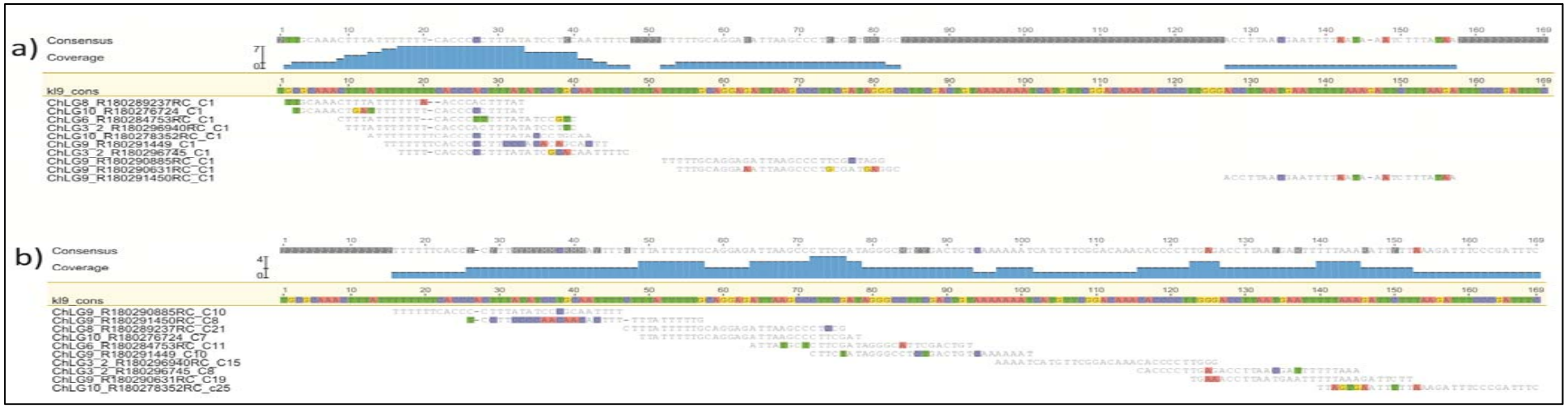
Supplementary Figure 4.2.10. Position of insertion sites of all arrays in Cluster 7 compared to consensus sequence. a) position of first 30 pb of each of the arrays in Cluster 7. b) position of last 30 pb of each of the arrays in Cluster 7.



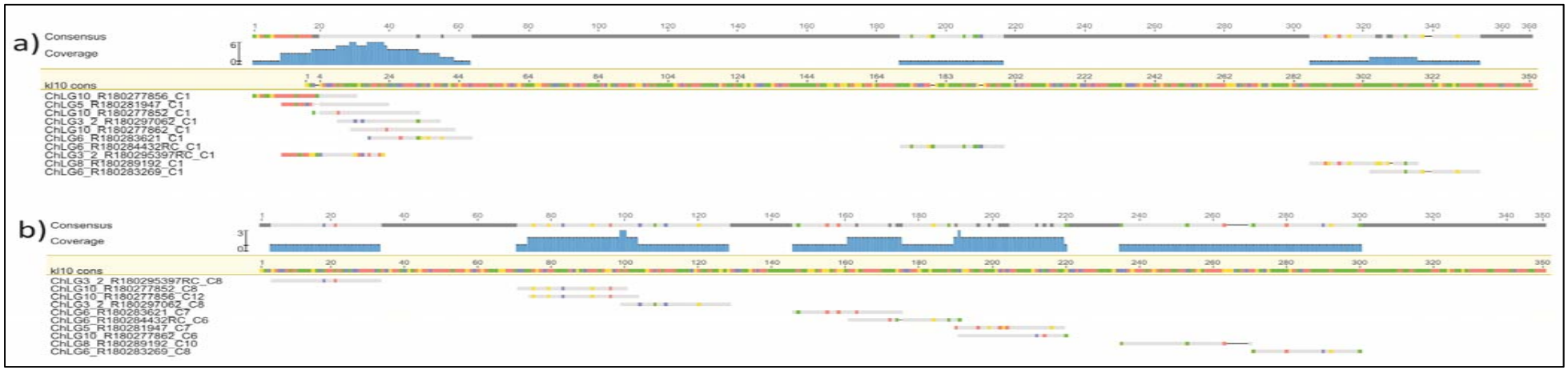
Supplementary Figure 4.2.11. Position of insertion sites of all arrays in Cluster 8 compared to consensus sequence. a) position of first 30 pb of each of the arrays in Cluster 8. b) position of last 30 pb of each of the arrays in Cluster 8.



Supplementary Figure 4.2.12. Position of insertion sites of all arrays in Cluster 9 compared to consensus sequence. a) position of first 30 pb of each of the arrays in Cluster 9. b) position of last 30 pb of each of the arrays in Cluster 9.



Supplementary Figure 4.2.13. Position of insertion sites of all arrays in Cluster 10 compared to consensus sequence. a) position of first 30 pb of each of the arrays in Cluster 10. b) position of last 30 pb of each of the arrays in Cluster 10.



REFERENCES

Rosandić M, Paar V, Basar I. 2003. Key-string segmentation algorithm and higher-order repeat 16mer (54 copies) in human alpha satellite DNA in chromosome 7. J Theor Biol 221:29–37.