Genome Biology and Evolution Advance Access published March 29, 2013 doi:10.1093/gbe/evt049

Contrasting patterns of transposable element and satellite distribution on sex chromosomes (XY_1Y_2) in the dioecious plant *Rumex acetosa*

Pavlina Steflova[1,2], Viktor Tokan[1], Ivan Vogel[1,2], Matej Lexa[2], Jiri Macas[3], Petr Novak[3], Roman Hobza[1,4], Boris Vyskot[1], Eduard Kejnovsky[1,2]

[1]*Department of Plant Developmental Genetics, Institute of Biophysics ASCR, Brno, Czech Republic, www.ibp.cz/labs/LPDG, email: kejnovsk@ibp.cz*
[2]*Laboratory of Genome Dynamics, CEITEC - Central European Institute of Technology, Masaryk University, Brno, Czech Republic*
[3] *Biology Centre ASCR, Institute of Plant Molecular Biology, Ceske Budejovice, Czech Republic*
[4] *Laboratory of Molecular Cytogenetics and Cytometry, Centre of the Region Haná for Biotechnological and Agricultural Research, Institute of Experimental Botany, Olomouc, Czech Republic*

* Corresponding author: Eduard Kejnovsky, Laboratory of Plant Developmental Genetics, Institute of Biophysics, Kralovopolska 135, CZ-612 65 Brno, Czech Republic, phone: +420 541517203, E-mail: kejnovsk@ibp.cz

Running title: Repetitive DNA in *Rumex acetosa* sex chromosomes (XY_1Y_2)

Abstract

Rumex acetosa is a dioecious plant with the XY_1Y_2 sex chromosome system. Both Y chromosomes are heterochromatic and are thought to be degenerated. We performed low-pass 454 sequencing and similarity-based clustering of male and female genomic 454 reads in order to identify and characterize major groups of *R. acetosa* repetitive DNA. We found that Copia and Gypsy retrotransposons dominated, followed by DNA transposons and non-LTR retrotransposons. CRM and Tat/Ogre retrotransposons dominated the Gypsy superfamily while Maximus/Sireviruses were most abundant among Copia retrotransposons. Only one Gypsy subfamily had accumulated on Y_1 and Y_2 chromosomes while many retrotransposons were ubiquitous on autosomes and the X chromosome, but absent on Y_1 and Y_2 chromosomes, and others were depleted from the X chromosome. One group of CRM Gypsy was specifically localized to centromeres. We also found that majority of previously described satellites (RAYSI, RAYSII, RAYSIII, RAE180) are accumulated on the Y chromosomes where we identified Y chromosome-specific variant of RAE180. We discovered two novel satellites - RA160 satellite dominating on the X chromosome, and RA690 localized mostly on the Y_1 chromosome. The expression pattern obtained from Illumina RNA sequencing showed that the expression of transposable elements is similar in leaves of both sexes and that satellites are also expressed. Contrasting patterns of TEs and satellite localization on sex chromosomes in *R. acetosa*, where not only accumulation but also depletion of repetitive DNA was observed, suggesting that a plethora of evolutionary processes can shape sex chromosomes.

Key words: sex chromosomes; sorrel (*Rumex acetosa*); transposable elements; satellites

Introduction

Sex chromosomes are the genomic regions undergoing specific evolutionary processes (Charlesworth 1991). There is also extraordinary variability in the patterns of sex chromosomes: not only the XY system dominating in mammals and the ZW system in lepidoptera and birds but also many variants with multiple X or Y chromosomes found in both animals and plants with the extreme example of five Xs and five Ys in the male platypus (McMillan 2007, Ming *et al.* 2011). The unifying feature of the Y and W chromosome is partial or complete loss of recombination with their partner X and Z chromosome, respectively, which leads to genetic degeneration of the Y or W chromosome and accumulation of repetitive DNA combined with expansion (Charlesworth *et al.* 1994, Kejnovsky *et al.* 2009a, Gvozdev *et al.* 2005). In plants, sex chromosomes are found in several dioecious species and often represent an early evolutionary stage (Ming *et al.* 2011). An incipient stage of sex chromosomes is represented by homomorphic sex chromosomes present in some plants (e.g. *Carica papaya* or *Bryonia dioica*). Other plants have evolutionarily older heteromorphic sex chromosomes with either large Y chromosome (*Silene latifolia*, *Coccinia grandis*, *Rumex acetosa*, *Cannabis sativa*) or small Y chromosome (*Cycas revoluta*, *Humulus lupulus*, *Marchantia polymorpha*, for review see Ming *et al.* 2011).

Sorrel (*Rumex acetosa*) is a dioecious plant with the XY_1Y_2 system. Dioecy in *Rumex* genus (XY system) arose about 16 mil years ago and the *acetosa* clade with multiple XY_1Y_2 system originated 12-13 mil years ago (Navajas-Pérez *et al.* 2005a). The X chromosome is the largest in male metaphase but both Y chromosomes together are bigger than the X chromosome. Five satellites have been found in *R. acetosa* - RAYSI, RAYII, RAYSII (specific for Y_1 and Y_2, Shibata *et al.* 1999, Navajas-Pérez *et al.* 2005b), RAE180 (Y_1, Y_2 and one autosome, Shibata *et al.* 2000) and RAE730 (autosomes, Shibata *et al.* 2000). RAYSI and RAE180 are the main components of the Y heterochromatin (Shibata *et al.* 1999, 2000). RAYSI is also common in other species with multiple XY_1Y_2 systems (*R. papillaris*, *R. intermedius*, *R. thyrsoides* and *R. tuberosus*) but absent in species with an XY system like *R. acetosella* and *R. suffruticosus* (Navajas-Pérez *et al.* 2005b, Cunado *et al.* 2007). RAE180 is expanded on the Y_1 chromosome in *R. acetosa*. It is also amplified on one autosome in *R. suffruticosus* and dispersed in low copy number in *R. acetosella* (Shibata *et al.* 2000, Cunado *et al.* 2007, Navajas-Pérez *et al.* 2009). RAYSI, RAYSII, RAYSIII and RAE730 satellites arose by different ancestral duplications and reshufflings from the same 120bp unit (Navajas-Pérez *et al.* 2005b, Mariotti *et al.* 2009). Intraspecific variability of Y-associated satellites like RAYSI and RAE180 is much higher than that in the autosomal RAE730 satellite which indicates a particular mode of evolution of satellites in a non-recombining genomic context (Navajas-Perez *et al.* 2005b, 2005c). To date, no TEs have been described in *Rumex* species. Only four clones originating from degenerate PCR on microdissected sex chromosomes exhibited homologies with Gypsy (DOP-47 and 61), Copia (DOP-60) and non-LTR retrotransposons (DOP-8, Mariotti *et al.* 2005). *R. acetosa* has two 45SrDNA loci on two autosomal pairs (Lengerova and Vyskot 2001).

Repetitive DNA forms a significant proportion of eukaryotic genomes. This is particularly evident in plants which have faster genome dynamics than animals (Kejnovsky *et al.* 2009b). However, the rules governing genome size and repeat composition are not fully understood. Even closely related species often significantly differ in composition of their transposable elements or satellites (Neumann *et al.* 2006). The chromosomal localization of repetitive DNA was previously thought to be only a result of selection but recent findings show that other factors such as targeting of TEs into specific chromosomal niches are also important (for review see Heslop-Harrison and Schwarzacher 2011, Kejnovsky *et al.* 2012b). The Y or W sex chromosomes often accumulate various repetitive DNA as has been proven for humans (Skaletsky *et al.* 2003), drosophila (Steinemann and Steinemann 1992), fish (Cioffi *et al.* 2011), and reptiles (Pokorna *et al.* 2011). In plants, tandem repeats (Hobza *et al.* 2007),

microsatellites (Kubat *et al.* 2008) and transposable elements (Cermak *et al.* 2008) are accumulated on the Y chromosome of *Silene latifolia*, while tandem repeats are gathered on both Y chromosomes in *Rumex acetosa* (Shibata *et al.* 1999, Mariotti *et al.* 2009), and transposable elements are accumulated on the Y chromosome in *Cannabis sativa* (Sakamoto *et al.* 2000). However, repetitive DNA can have also other patterns than simple accumulation on the Y chromosome. For example, Ogre retrotransposon is ubiquitous on all autosomes and the X chromosome but is absent on the Y chromosome in *Silene latifolia* (Cermak *et al.* 2008). Microsatellites are accumulated on the X chromosome rather than Y chromosome in fish *Hoplias malabaricus* (Cioffi *et al.* 2011) and some microsatellites are absent on the W chromosome in the lizard *Eremias velox* despite their presence on other chromosomes (Pokorna *et al.* 2011).

In this study, we analyzed the structure, genomic proportion, expression and chromosomal localization of the main classes of TEs and satellites in the dioecious plant *Rumex acetosa*. We found that and Maximus/Sireviruses (among Copia elements) and Chromoviruses (among Gypsy elements) predominate and their chromosomal localization exhibits various contrasting patterns, e.g., not only accumulation on the Y chromosomes.

Materials and methods

454 sequencing. One sequencing run of the 454 GS FLX platform (454 Life Sciences, Roche) was performed for each male and female genomic DNA isolated from healthy young leaves, resulting in 280,954 and 295,993, quality-filtered reads, respectively, with average read length 332 nucleotides for the male and 338 nucleotides for the female sample (Accession numbers SRX118072 and SRX118073). Male and female read sets were combined for the purpose of complex analysis, providing a total of 193.4 Mb of sequencing data. Given the genome size of *R. acetosa* 7.0pg in female, 7.5pg in male (2C) (Blocka-Wandas *et al.* 2007), this represents 5.7% of the genome. The sequencing reads were clustered on the basis of similarity (as described by Novak *et al.* 2010, Macas *et al.* 2011) and clusters containing at least 57 reads (representing around 0.01% of the genome) were used for further analysis.

Illumina sequencing. Pair-end sequencing was performed for two male and two female genomic DNA.These were, isolated from leaves representing parents and their single male and female progenitors (deposited under SRA062840). The leaves from the same individuals were then used for RNA-Seq experiment (deposited under SRA058606) resulting in 4 pair-end libraries of transcriptomic data. Both genomic and transcriptomic classes of reads were then analyzed using FastQC (available at http://www.bioinformatics.babraham.ac.uk) quality control tool. The reads were trimmed and filtered on the basis of quality using FASTX-toolkit (available at http://hannonlab.cshl.edu/fastx_toolkit/) and the redundant reads were removed from all datasets. Both genomic and transcriptomic libraries were then mapped to the identified clusters of genomic 454 data using BLAT (Kent 2002). The BLAT analysis was run with default parameters, except for the stepSize parameter which was reduced to nine. This was to ensure greater sensitivity of the mapping analysis (the genome is sampled with higher sampling frequency). To eliminate redundancy in the obtained alignments, the following steps were taken: only alignments with an alignment e-value of less than 10^{-20} and 10^{-15} for genomic and transcriptomic data were considered. The BLAT output was then sorted according to e-value, percent identity and alignment score. Only alignments that fitted the these criteria best were chosen for future analysis to ensure that every Illumina read was mapped only once to one of the genomic reads (locations). The numbers of mapped Illumina genomic versus 454 genomic reads (Table 1) or transcriptome reads versus 454 genomic reads (Figure 6) were counted for every cluster and subsequently for every identified TE family of *Rumex acetosa*. The weighted average (considering library sizes) of the relative expression

and genomic proportions of repetitive families was counted. In-house computational pipeline including custom-made Bash and Python scripts was used to sort and filter the alignments.

Phylogeny and classification. The reconstructed DNA sequences were analyzed for the presence of a reverse transcriptase (RT) domain by sequence similarity. Nucleotide sequences of RT cores were then used to place the clusters into a phylogenetic tree of LTR-retrotransposon RT domains. The identification of RT cores was based on a collection of consensus amino-acid sequences of known RT domains available at Gypsy Database (Llorens *et al.* 2011) and TREP database (http://wheat.pw.usda.gov/ITMI/Repeats/). This collection was used to create a BLAST+ (Camacho *et al.* 2009) database and searched using the blastx command with DNA sequences of the LTR elements in question. Regions having E-value < 10^{-3} where cut out making sure they were unique and fell into the 500-1000bp range observed for the best matches of well-known elements. The extracted RT cores were subsequently analyzed using the Geneious Pro Alignment tool (Drummond *et al.* 2011) to generate a multiple nucleotide sequence alignment. Once aligned, the Neighbor-Joining distance model of the Geneious Pro Tree Builder was used to build a phylogenetic tree.

Structural annotation of LTR elements. The reconstructed nucleotide sequences were first analyzed for the presence of structural features typical for specific classes of repetitive sequences, namely LTRs, *gag* and *pol* genes and their individual protein domains (GAG,AP,RT,RH,INT), other ORFs, PBS and/or PPT. The presence of typical protein domains was detected by sequence similarity, in the same way as the detection of RT cores in the previous paragraph, except for using the appropriate consensus sequences. The recognition of *gag* and *pol* genes relied on the combined evidence of predicted ORFs using the FrameD++ software package (Schiex *et al.* 2003) used because of its tolerance to reading frame interruptions and the presence of protein domains. However, no exact delimitation of the ORFs/genes was attempted because of the nature of the analyzed sequences (e.g. averaged from multiple loci, presence of non-functional but autonomous LTR elements). The PBS, PPT sequences were detected using the LTR finder software (Xu and Wang 2007).

Preparation of probes for FISH. Specific primers were designed, usually for reverse transcriptase or the transposase domain of individual TEs. In the first step, template DNA was amplified using PCR with a mix containing 1 x complete PCR buffer, 0.1 mM dNTPs, 0.1 µM primers, 0.5 U Taq polymerase (Top Bio) and 10-15 ng of template DNA. Reaction conditions were as follows: 94°C/4min 34x (94°C/50s + 55°C/50s + 72°C/1min) + 72°C/5min. PCR products were checked by gel electrophoresis, cleaned using the PCR purification kit (Qiagen), cloned into pDrive vector (Qiagen) and transformed to *E. coli*. Clones were sequenced to verify the presence of a specific product. Selected clones were then used for preparation on probes for FISH by PCR and labeling using Nick Translation Kit (Roche).

Fluorescent in situ hybridisation (FISH). FISH was performed on mitotic metaphase chromosomes, prepared from root tip cells. The hybridization mix contained 50% formamide, 2xSSC, 10% dextran sulphate. 1-5 ng/µl labeled DNA was denatured, added to slide and hybridized at 37°C for 18 hours. Slides were then washed 2 x 5´ in 2xSSC at 42°C, 2 x 5´ in 0.1xSSC at 42°C, 2 x 5´ in 2xSSC at 42°C, 5´ in 2xSSC at room temperature, 7´ in 4xSSC + 1% Tween and finally washed in 1xPBS. The chromosomes were counterstained with DAPI, viewed in Olympus AX70 fluorescent microscope, scanned by CCD camera and analyzed by ISIS software.

Satellite DNA sequence analysis. 454 sequencing reads were analyzed for potential repetitive sequence motifs. Known repeats were identified in clustered reads by sequence similarity to known *Rumex* satellite sequences RAYSI, RAYSII, RAYSIII, RAE180 and RAE730

(Navajas-Peréz *et al.* 2005b). The sequences were downloaded from PlantSat (Macas *et al.* 2002) and NCBI Genbank (Benson *et al.* 2012). Owing to the prevalence of plastid DNA, mitochondrial and retroelement sequences in the 454 data, clusters of reads with top BLAST hits mapping to known repetitive sequences were eliminated before further analysis. The remaining 454 reads were subjected to k-mer counting and extension by the algorithm of Macas *et al.* (2010, 2011). Identified repeat motifs were associated with clusters of origin and the cluster contigs were visually analyzed for tandemly repeated regions using the *polydot* program from the EMBOSS package at *word size*=9 (Rice *et al.* 2000). Sequences with tandem subrepeats were broken into their respective monomers at the first point of self-similarity, as determined by running BLAST (Altschul *et al.* 1997) of the sequence on itself with *word_size*=7 and a threshold of *e*=0.001. The obtained monomer sequences were used for an exhaustive search of additional matches in the 454 sequence reads. The collected sequences were aligned with CLUSTALX (Thompson *et al.* 1997) and displayed as a sequence logo using the Weblogo 3.3 program (Crooks *et al.* 2004). Each multiple alignment was used to generate a consensus monomer sequence at 80% identity threshold.

The same sequences were analysed with CLANS clustering software (CLuster ANnalysis of Sequences, Frickey and Lupas 2004) to reveal families and subgroups of all seven satellites. In each analysis the software was used to cluster sequencing reads, position the clusters for optimal visualization, as well as showing sequencing reads from male and female plants in contrasting colors.

Results

Genomic proportion and composition of repetitive DNA

We performed one 454 GS FLX platform sequencing run for each male and female genomic DNA and similarity-based clustering of the reads. The first 260 clusters (with more than 57 reads) contained 335,924 reads and represented 58.3% of genome. We obtained also 22,555 smaller clusters (with 2-57 reads) that contained 74,605 reads (12.9% of genome). The other 166,079 reads that remained as singlets represented 29% of genome (Figure 1). We found the main groups of transposable elements, satellites, rDNA loci and chloroplast DNA. The chloroplast genome was represented by 8 clusters (Figure 1). The majority of chloroplast DNA reads probably originated in contaminating cpDNA, even though a proportion might have come from nuclear cpDNA insertions (NUPTs). For this reason, we removed chloroplast DNA reads from further analysis of the nuclear genome.

We focused on TEs and satellites for which we identified individual families together with their genome proportions in male and female individuals (Table 1). Although reconstruction of elements was done using 454 data, genome proportions were estimated from Illumina data which provide more representative results (Macas *et al.* 2011). The most abundant were Maximus/Sire family of Copia retrotransposons (34.9% in male and 35.6% in female genomes) followed by Chromovirus/CRM and Tat/Ogre families of Gypsy retrotransposons (5.7% and 5.4% in male). LINE elements and two superfamilies of DNA transposons - Mutator and CACTA – were found to make up considerably smaller genome proportions (Table 1). All transposable elements represented together about 49% of the genome.

All seven types of satellites together comprised 5.15% of male and 2.54% of female genomes. The most abundant were RAE180 satellite representing 2.72% of the male genome (Table 1). The proportion of RAYSI, RAYSII and RAYSIII was much higher in males in agreement with their Y-specific localization. RAE180 was more abundant in males because of accumulation on both Y chromosomes. Other tandemly arranged sequences are rDNAs that are located on two autosomal pairs in *R. acetosa* (Lengerova and Vyskot 2001). The unexpected difference in proportion of rDNA in male (0.18%) and female (0.21%) was

probably caused by higher sensitivity of GC-rich sequences (like rDNA) to quality of sequencing as was demonstrated by Macas *et al.* (2011).

In order to classify Copia and Gypsy elements in more detail we aligned their reverse transcriptase (RT) domains in individual clusters and constructed phylogenetic trees for both superfamilies (Figure 2). Both trees contained subfamilies identified in our clusters (in red) together with representatives of known subfamilies of Copia or Gypsy from other plant species (in black). Among Copia, we identified nine subfamilies of Maximus/Sireviruses, one TAR subfamily and one Bianca subfamily (Figure 2A). Chromoviruses were dominant among Gypsy elements with seven CRM subfamilies and one Tekay subfamily. We compared *R. acetosa* CRM subfamilies with other CRM elements published by Neumann et al (2011). The phylogenetic tree based on reverse transcriptase of CRM elements showed that all seven subfamilies found in *R. acetosa* clustered together with group A (Figure S1) which is known to represent CRM elements having a CR motif and localized in the centromere. We found the CR motif in the seven *R. acetosa* CRM subfamilies, well-conserved in five subfamilies (CL25, 42, 51, 15 and 67) while only partially preserved in two subfamilies (CL28 and 48, Figure S2). In addition to CRM elements, we found one Tekay/Del subfamily (CL37) which also belonged to the chromoviruses, two Tat subfamilies (CL 11 and CL17) and one Athila subfamily (Figure 2B).

We analyzed the structure of selected TE families reconstructed from 454 sequencing data (Figure 3). We were able to discern all main features characteristic for the specific family – gag and pol genes, LTRs, PBS and PPT regions (Figure 3). In some elements (CL2, CL17), LTR regions were assembled into one LTR while in other clusters (CL5, CL25) right and left LTR were distinguished. As examples of very abundant Copia retrotransposons, we present the reconstructed Maximus/Sire subfamilies from CL2 and CL5. We found an extra ORF in the 3'UTR of Maximus/Sire subfamily corresponding to CL5. As an example of Gypsy retrotransposons, we used the Tat subfamily (CL17) which has a long 5'UTR region. We measured the coverage of all these elements with male (blue) and female (red) genomic reads (Figure 3, lines below reconstructed elements). We were unable to reconstruct the whole CRM elements with LTRs from CL42. However, coverage of elements with genomic reads was higher in male than in female which is consistent with the accumulation of this CRM subfamily on both Y chromosomes (Figure 4J).

Novel satellite sequences

K-mer frequency analysis (Macas *et al.* 2010) of the 454 sequencing reads from clusters not mapped to known sequences, helped us to identify two candidates for novel tandem repeats. The first candidate originated from cluster CL45, while the second candidate belonged to cluster CL65 (later detected also in CL38 and CL68). These candidate sequences were mapped to assembled contigs in their respective clusters and further adjusted to match existing reads in their size and composition as described in Methods, limiting the sequences to a single monomer of the repeat (consensus monomer sequence, Figure S1). This procedure had led to the discovery of two novel DNA satellites in *Rumex acetosa.* We named the new satellites RA160 (CL45) and RA690 (CL38, CL65, CL68), based on their species of origin and their approximate monomer length. Monomer consensus sequences of two novel satellites obtained from multiple alignments of individual 454 reads as well as the sequences of previously described satellites RAYSI-III, RAE180 and RAE730 assembled by our aproach were used to design PCR primers and obtain representative genomic *Rumex* sequences for each family as described in Methods (Preparation of FISH probes). These sequences are available under accession numbers KC310873-KC310879. Genomic proportions of RA160 and RA690 in males are 0.61% and 0.29%, respectively (Table 1).

Chromosomal localization of transposable elements and satellites

In order to find the chromosomal localization of all main types of transposable elements we prepared probes representing various parts of individual TE families (Table S1) from the first 60 clusters and used them for fluorescence *in situ* hybridization (FISH) on metaphase chromosomes of the male *R. acetosa*. We obtained several contrasting patterns of chromosomal distribution. The most typical patterns are shown Figure 4. The most abundant subfamilies of Maximus/Sireviruses were distributed on all chromosomes but were absent (CL2, CL7 and CL18) or depleted (CL5) on Y_1 and Y_2 chromosomes (Figure 4A, B, C, E). Respective subfamilies differed in signal intensity and the extent of subtelomeres labeling - e.g. the subfamily corresponding to CL18 was present only a short distance from centromeres (Figure 4E) while the two subfamilies (CL5 and CL7) covered the whole chromosome but the tip. The absence of Maximus/Sire on the Y chromosomes was consistent with its slightly lower genome proportion in male than female individuals (Table 1).

The patterns of hybridization of Gypsy elements were more variable. Tat elements (CL11) and Tekay/Del (CL22, CL37), like Copia elements, were absent from both Y chromosomes (Figure 4D, F). Surprisingly, Athila (CL41) was even absent on the X chromosome (Figure 4I). The CRM elements (CL42) showed accumulation on both Y chromosomes compared to a slight signal on all the other chromosomes (Figure 4J). Accumulation of elements on the Y chromosome or their absence on the X chromosome caused a higher genomic proportion in males than females as calculated from Illumina sequencing data (Table 1). Another CRM subfamily (CL25), coming from the same clade in the phylogenetic tree as the Y-accumulated CRM subfamily (CL42), gave specific centromeric signals on all chromosomes - signals on all autosomes were discrete and much stronger than on either Y chromosome with the weakest signal in the centromere of the X chromosome. There were additional signals to centromeric ones on both Y chromosomes (Figure 4G). The most abundant DNA transposon – the Mutator superfamily - was preferentially located in the subtelomeres of the majority of chromosomes (Figure 4H).

RAYSI satellite was used in most samples as the Y chromosomes marker (Figure 4A-P). RAYSI (CL30) was localized in four loci on each arm of the Y_1 chromosome and in two large loci at the p-arm and two minor loci on the q-arm of the Y_2 chromosome (Figure 4K). RAYSII (CL221) was present as two signals in the middle of the p-arm of the Y_1 chromosome but was absent on the Y_2 chromosome (Figure 4K). RAYSIII (CL109, CL126, CL158) was found in four strong loci on the Y_2 chromosome and three minor loci on the Y_1 chromosome (Figure 4L). RAE180 (CL32, CL73) was found in many loci on both Y chromosomes and on almost all autosomes and the X chromosome (Figure 4M). RAE730 (CL24) was present as a strong signal on one autosomal pair and as a minor signal on both arms of the Y_1 chromosome (Figure 4N). RA160 gave two strong and one weak signals on the p-arm and two minor signals on the q-arm of the X, weaker signal on both arms of the Y_1 chromosomes, three minor signals on the q-arm of the Y_2 chromosome and minor signals on two autosomal pairs (Figure 4O). RA690 was localized in two bands on the q-arm of the Y_1 chromosome, one minor signal was present on the q-arm of the Y_2 chromosome, in the centromere of the X chromosome and on two autosomal pairs (Figure 4P). Localization of all studied satellites on the Y_1, Y_2 and X chromosomes is summarized in a schematic map (Figure 5).

Sequence homogeneity of satellites and their putative Y-linked variants

In order to assess the homogeneity/variability of satellites and in an attempt to reveal potential male-specific (Y-linked) satellite variant(s), we clustered sequence reads corresponding to all seven satellites present in *R. acetosa* genome using CLANS software (Frickey and Lupas 2004). Reads originating from male plants are blue while reads from female plants are shown in red (Figure 6). We found that out of all analyzed satellites, RAYSI, RAYSII and RAYSIII were the most related ones. Since these satellites are localized mostly on the Y chromosomes,

blue symbols prevail in the the RAYSI, RAYSII and RAYSIII clusters. The homogeneity was highest in RAYSI (specifically localized on both Y chromosomes) and in RAE730 (localized only on one autosomal pair, Figure 4 and 6a). On the other hand, RAE180 showed the highest variability (present on all chromosomes, Figure 4 and 6a). A more detailed analysis of the RAE180 cluster revealed a male-specific variant that corresponded to RAE180 satellites present on the Y chromosomes (Figure 6b). The larger size of the subcluster formed by the blue dots indicates that putative Y-linked RAE180 satellites are more diverged compared to their X-linked and autosomal counterparts (Figure 6b). Similarly, we found two subclusters inside RAYSI and RAYSIII clusters (Figure 6c, d). It remains to be determined whether these two clusters correspond to satellite variants localized on the Y_1 and Y_2 chromosomes or represent two subgroups localized on both Y chromosomes. Sequence logos show sequence differences of putative Y-linked variants of RAE180 satellite compared to variant localized on autosomes and the X chromosome (Figure 6e).

Expression of transposable elements and satellites

We performed Illumina platform sequencing of RNA isolated from leaves of male and female *R. acetosa* plants. Reads were mapped onto the clusters corresponding to transposable elements and the relative expression of individual TE families for each sex was measured (Figure 7). The majority of expression reads corresponded to Maximus/Sire followed by CRM elements and Tat/Ogre elements. All of these are most abundant in the genome. However, when the relative expression of each TE family was compared to its genomic proportion, it was evident that CRM, TAR/Tork, Bianca, LINE and CACTA elements were relatively more transcribed at the expense of Maximus/Sire and Tekay/Del elements. The transcription of other elements (Tat/Ogre, Athila and Mutator) corresponded more or less to their genomic proportions. We found that some satellites were also expressed - expression of RAE180 corresponded to its genomic proportion, RA690 was overexpressed and RAYSI, RAYSII, RAYSIII, RAE160 and RAE730 were underexpressed (Figure 7).

Discussion

This study is the first comprehensive characterization of the repetitive fraction of the nuclear genome of *Rumex acetosa,* a model dioecious plant with a multiple sex chromosomal system (XY_1Y_2). We found that abundant repetitive DNA represents at least 49% of the genome. This estimation represents highly and middle abundant repeats found in first 260 clusters and the proportion of repetitive fraction would be higher if also other clusters with low repetitive fraction are taken into account. We showed that *R. acetosa* genome is composed of Copia LTR retrotransposons and only smaller proportion is made up of Gypsy retrotransposons, DNA transposons and satellite DNA. However, it is difficult to conclude why specific (sub)families are more abundant than others because the mechanism(s) governing the colonization of genomes by different groups of TEs are not fully understood. For example, LTR retrotransposons dominate in maize and poplar, non-LTR retrotransposons make up a significant proportion in *Brassica oleracea* and *Gossypium raimondii*, and DNA transposons are most abundant in *Lotus japonicus* and *Fragaria vesca* (for review, see Kejnovsky *et al.* 2012a).

In our work we showed that tandem repeats in *R. acetosa* are strongly gathered on Y_1, Y_2 or both Y chromosomes in contrast to the variable chromosomal patterns of TEs. Among TEs we showed that while some TEs are accumulated on both Y chromosomes (CRM - CL42), the majority of TEs are missing or underrepresented there (Maximus/Sire or Tat/Ogre). It is surprising that despite the fact that all CRM subfamilies contained the CR motif, only the CL25 subfamily was localized in centromeres, indicating that the presence of the CR motif is not a sufficient condition for centromeric localization and that other factors are also

important. Our results show that the generally accepted picture of Y chromosomes, as those where all repeats are only accumulated, should be modified. We can explain repeat distribution patterns on sex chromosomes in *R. acetosa* by high rate of colonization of Y_1 and Y_2 chromosomes by satellites that prevented transposable elements from significantly expanding there. Nonetheless, some TEs (CRM, CL42) were able to compete with satellites for Y-linked niches either by higher insertion rate or lower rate of removal.

Our data are relevant to questions on the structure, evolution and age of sex chromosomes. Known sex chromosomes in plants are mostly in the early stages of their evolution compared to much older mammalian sex chromosomes (Vyskot and Hobza, 2004). The young evolutionary age of plant sex chromosomes probably results in some satellites and retrotransposons being weakly accumulated and only slightly enriched on the Y chromosome in the most studied dioecious plant with sex chromosomes - *Silene latifolia* (Hobza *et al.* 2007, Cermak *et al.* 2008). Our findings show that the situation in *Rumex acetosa* is different: some satellites show strong accumulation or even Y chromosome-specific localization and both Y chromosomes that represent together 39% of the genome (Blocka-Wandas *et al.* 2007). Therefore, the *R. acetosa* Y chromosomes have different sequence composition than the X chromosome and autosomes. They are probably more degenerated and older than sex chromosomes in *S. latifolia*. This view is supported by the finding that both Y chromosomes in *R. acetosa* are heterochromatic while the Y chromosome in *S. latifolia* is euchromatic.

Distribution of various satellites along the whole length of the Y1 and Y2 chromosomes could indicate that there are no evolutionary strata on sex chromosomes in *R. acetosa*, similar to the strata found in human X chromosome (Lahn and Page 1999). If they were present, satellites should accumulate more intensively in a region of the Y chromosomes corresponding to a part of the X chromosome that stopped recombination earlier. However, an exact determination of an existence of the evolutionary strata would need an analyzis of dozens of genes located on the X and Y chromosomes. No genes have been identified in *R. acetosa* yet. Accumulation of several satellites at centromere of the X chromosome of *R. acetosa* (Figure 5) could indicate lowered recombination in that region.

Another question concerns the origin of two Y chromosomes in *R. acetosa*. Two alternative explanations have been proposed: the splitting of one original Y chromosome and translocation of an autosome onto the X chromosome (Vyskot and Hobza 2004). The same distribution of repetitive DNA on both Y chromosomes would indicate that their age was the same and supports the splitting hypothesis. In this study, we found that some tandem repeats had different localizations on the two Y chromosomes. However, our recent data show that CA and CAA microsatellites are strongly and evenly accumulated on both Y chromosomes (Kejnovsky *et al.* 2012b). Thus, we cannot conclude whether both Y chromosomes are of the same or a different age and hence we cannot support either of the two hypotheses. Identification of the genes localized on the Y_1, Y_2 and X chromosomes of *R. acetosa* together with detailed characterization of the genomic landscape of these sex chromosomes (sequencing of BAC clones) are necessary to shed light on their age, mechanism of origin and evolutionary trajectories.

Acknowledgements

This work was supported by the Grant Agency of the Czech Republic (grants P305/10/0930 to EK, P501/10/0102 to BV, P501/12/2220 to RH), grant AV0Z50040702 and RVO:60077344 from the Academy of Sciences of the Czech Republic, by the project "CEITEC - Central European Institute of Technology" (CZ.1.05/1.1.00/02.0068) from European Regional Development Fund and by the project OPVK (CZ.1.07/2.3.00/20.0045) and grant No. ED0007/01/01 Centre of the Region Haná for Biotechnological and Agricultural Research.

Literature Cited

Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW. 2012. GenBank. Nucl Acids Res 40: Database Issue, D32-37.

Blocka-Wandas M, Slivinska E, Grabowska-Joachimiak A, Musial K, Joachimiak AJ. 2007. Male gametophyte development and two different DNA classes of pollen grains in Rumex acetosa L., a plant with an XX/XY_1Y_2 sex chromosome system and female-biased sex ratio. Sexual Plant Reprod 20: 171-180.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10: 421.

Cermak T, Kubat Z, Hobza R, Koblizkova A, Widmer A, Macas J, Vyskot B, Kejnovsky E. 2008. Survey of repetitive sequences in Silene latifolia with respect to their distribution on sex chromosomes. Chromosome Research 16: 961-976.

Cioffi MB, Kejnovsky E, Bertollo LAC. 2011. The chromosomal distribution of microsatellite repeats in the genome of the wolf fish Hoplias malabaricus, focusing on the sex chromosomes. Cytogenetics and Genome Research 132: 289-296.

Charlesworth B. 1991. The evolution of sex chromosomes. Science 251: 1030-1033.

Charlesworth B, Sniegowski P, Stephan W. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371: 215-220.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: A sequence logo generator. Genome Research 14: 1188-1190.

Cunado N, Navajas-Pérez R, Herrán R, Ruiz Rejón C, Ruiz Rejón M, Santos JL, Garrido-Ramos MA. 2007. The evolution of sex chromosomes in the genus Rumex (Polygonaceae): Identification of a new species with heteromorphic sex chromosomes. Chromosome Res 15: 825-832.

Drummond AJ, Ashton B, Buxton S, Cheung M, Cooper A, Duran C, Field M, Heled J, Kearse M, Markowitz S, Moir R, Stones-Havas S, Sturrock S, Thierer T, Wilson A. 2011. Geneious v5.5, Available from http://www.geneious.com

Frickey T and Lupas A (2004) CLANS : a Java application for visualizing protein families based on pairwise similarity. Bioinformatics 20 : 3702-3704.

Gvozdev VA, Kogan GL, Usakin LA. 2005. The Y chromosome as a target for acquired and amplified genetic material in evolution. BioEssays 27: 1256-1262.

Heslop-Harrison JS and Schwarzacher T. 2011. Organization of the plant genome in chromosomes. Plant J 66: 18-33.

Hobza R, Lengerova M, Svoboda J, Kubekova H, Kejnovsky E and Vyskot B. 2006. An accumulation of a tandem DNA repeats on the Y chromosome in an early stages of sex chromosome evolution. Chromosoma 115: 376-382.

Kejnovsky E, Hobza R, Kubat Z, Cermak T, Vyskot B. 2009a. The role of repetitive DNA in structure and evolution of sex chromosomes in plants. Heredity 102: 533-541.

Kejnovsky E, Leitch I, Leitch A. 2009b. Contrasting evolutionary dynamics between angiosperm and mammalian genomes. Trends in Ecology and Evolution 24: 572-582.

Kejnovsky E, Hawkins JS, Feschotte C. 2012a. Plant transposable elements: biology and evolution. Plant Genome Diversity. Volume 1: Plant genomes, their residents, and their evolutionary dynamics. Ed. by JF Wendel, J Greilhuber, J Dolezel, I Leitch (Springer).

Kejnovsky E, Michalovova M, Steflova P, Kejnovska I, Manzano S, Hobza R, Kubat Z, Kovarik J, Jamilena M, Vyskot B. 2012b. Expansion of microsatellites on evolutionary young Y chromosome. PLOS ONE 8: e45519..

Kent WJ. 2002. BLAT – The BLAST – Like Alignment Tool. Genome Research 12: 656-664.

Kubat Z, Hobza R, Vyskot B, Kejnovsky E. 2008. Microsatellite accumulation on the Y chromosome in Silene latifolia. Genome 51: 350-356.

Lahn BT and Page DC (1999) Four evolutionary strata on the human X chromosome. Nature 286: 964-967.

Lengerova M and Vyskot B. 2001. Sex chromatin and nucleolar analysis in Rumex acetosa L. Protoplasma 217: 147-153.

Llorens C, Futami R, Covelli L, Dominguez-Escriba L, Viu JM, Tamarit D, Aguilar-Rodriguez J, Vicente-Ripolles M, Fuster G, Bernet GP, Maumus F, Munoz-Pomer A, Sempere JM, LaTorre A, Moya A. 2011. The Gypsy Database (GyDB) of Mobile Genetic Elements: Release 2.0 Nucl Acids Res (NARESE) 39 (suppl 1): D70-D74 doi: 10.1093/nar/gkq1061

Macas K, Meszaros T, Nouzova M. 2002. PlantSat: a specialized database for plant satellite repeats. Bioinformatics 18: 28-35.

Macas J, Neumann P, Navrátilová A. 2007a. Repetitive DNA in the pea (Pisum sativum L.) genome: Comprehensive characterization using 454 sequencing and comparison to soybean and Medicago truncatula. BMC Genomics 8: 427.

Macas J, Neumann P. 2007b. Ogre elements – A distinct group of plant Ty3/gypsy-like retrotransposons. Gene 390: 108-116

Macas J, Neumann P, Novak P, Jiang J. 2010. Global sequence characterization of rice centromeric satellite based on oligomer frequency analysis in large-scale sequencing data. Bioinformatics 26: 2101-2108.

Macas J, Kejnovsky E, Neumann P, Novak P, Koblizkova A, Vyskot B. 2011. Next generation sequencing-based analysis of repetitive DNA in the model dioceous plant Silene latifolia. PLoS ONE 6: e27335.

Mariotti B, Navajas-Pérez R, Lozano R, Parker JS, Herrán R, Ruiz Rejón C, Ruiz Rejón M, Garrido-Ramos M, Jamilena M. 2005. Cloning and characterization of dispersed repetitive DNA derived from microdissected sex chromosomes of Rumex acetosa. Genome 49: 114-121.

Mariotti B, Manzano S, Kejnovsky E, Vyskot B, Jamilena M. 2009. Accumulation of Y-specific satellite DNAs during the evolution of Rumex acetosa sex chromosomes. Molecular Genetics and Genomics 281: 249-259.

McMillan D, Miethke P, Alsop AE, Rens W, O'Brien PO et al.. 2007. Characterizing the chromosomes of the platypus (Ornithorhynchus anatinus). Chromosome Research 15: 961-974.

Ming R, Bendahmane A, Renner SS. 2011. Sex chromosomes in land plants. Annual Rev Plant Biol 62: 485-514.

Navajas-Pérez R, Herrán R, López Gonzáles GL, Jamilena M, Lozano R, Ruiz Rejón C, Ruiz Rejón M, Garrido-Ramos MA. 2005a. The evolution of reproductive systems and sex-determining mechanisms within Rumex (Polygonaceae) inferred from nuclear and chloroplastidial sequence data. Mol Biol Evol 22: 1929-1939.

Navajas-Pérez R, Herrán R, Jamilena M, Lozano R, Ruiz-Rejón C, Ruiz Rejón M, Garrido-Ramos MA. 2005b. Reduced rate of sequence evolution of Y-linked satellite DNA in Rumex (Polygonaceae). J Mol Evol 60: 391-399.

Navajas-Pérez R, Schwarzacher T, Herrán R, Ruiz Rejón C, Ruiz Rejón M, Garrido-Ramos MA. 2005c. The origin and evolution of the variability in a Y-specific satellite-DNA of Rumex acetosa and its relatives. Gene 368: 61-71.

Navajas-Pérez R, Quesada del Bosque ME, Garrido-Ramos MA. 2009. Effect of localization, organization, and repeat-copy number in satellite-DNA evolution. Mol Genet Genomics 282: 395-406.

Neumann P, Požárková D, Macas J. 2003. Highly abundant pea LTR retrotransposon Ogre is constitutively transcribed and partially spliced. Plan Mol Biol 53: 399-410.

Neumann P, Koblízková A, Navrátilová A, and Macas J. 2006. Significant expansion of Vicia pannonica genome size mediated by amplification of a single type of giant retroelement. Genetics 173: 1047-1056.

Neumann P, Navrátilová A, Koblížková A, Kejnovský E, Hřibová E, Hobza R, Widmer A, Doležel J and Macas J. 2011. Plant centromeric retrotransposons: a structural and cytogenetic perspective. Mobile DNA 2:4.

Pokorna M, Kratochvil L, Kejnovsky E. 2011. Microsatellite distribution on sex chromosomes at different stages of heteromorphism and heterochromatinization in two lizard species (Squamata: Eublepharidae: Coleonyx elegans and Laceridae: Eremias velox). BMC Genetics 12: 90, DOI: 10.1186/1471-2156-12-90.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European molecular biology open software suite. Trends Genet 16: 276-277.

Sakamoto K, Ohmido N, Fukui K, Kamada H, Satoh S. 2000. Site-specific accumulation of a LINE-like retrotransposon in a sex chromosome of the dioecious plant Cannabis sativa. Plant Mol Biol 44: 723-32.

Shibata F, Hizume M, Kuroki Y. 1999. Chromosome painting of Y chromosomes and isolation of a Y chromosome-specific repetitive sequences in the dioecious plant Rumex acetosa. Chromosoma 108: 266-270.

Shibata F, Hizume M, Kuroki Y. 2000. Differentiation and the polymorphic nature of the Y chromosomes revealed by repetitive sequences in the dioecious plant, Rumex acetosa. Chromosome Res 8: 229-236.

Schiex T, Gouzy J, Moisan A, and de Oliveira Y. 2003. FrameD: a flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences. Nucl Acids Res 31: 3738-3741.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ et al. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature 423: 825-837.

Steinemann M, Steinemann S. 1992. Degenerating Y chromosome of Drosophila miranda: A trap for retrotransposons. Proc Natl Acad Sci USA 89: 7591-7595.

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. Nucl Acids Res 25: 4876-4882.

Vyskot B, and Hobza R. 2004. Gender in plants: sex chromosomes are emerging from the fog. Trends Genet 20: 432-438.

Wicker T and Keller B. 2007. Genome-wide comparative analysis of copia retrotransposons from wheat, barley, rice and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. Genome Res 17: 1072-1081.

Xu Z, Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucl Acids Res 35(Web Server issue): W265-W268.

Figure Legends:

Table 1. Repeat composition in *Rumex acetosa* genome estimated from Illumina sequencing data.

Figure 1. Repeat composition of clusters and their genomic proportions. The height of columns represents number of reads in the each clusters, the width of column indicate genomic proportion of cluster.

Figure 2. Phylogenetic trees of *Rumex acetosa Copia* (A) and *Gypsy* (B) retrotransposons based on reverse transcriptase sequences. Retrotransposons reconstructed from 454 reads in this study are in red, representative *Copia* and *Gypsy* retrotransposons from other plant species (from GenBank) are in black. Individual families are highlighted by different colors.

Figure 3. Comparison of structure of selected retrotransposon families (A-E). Graphs of **coverage by male** (in blue) **and female** (in red) **genomic** reads are showed under the structure of each element.

Figure 4. Localization of transposable elements and satellites on metaphase chromosomes of *Rumex acetosa* using fluorescence *in situ* hybridization. The name of transposable element family together with the number of corresponding cluster is inside each figure. Bar indicates 10μm.

Figure 5. Schematic map of satellites localization on the Y1, Y2 and X chromosomes in *R. acetosa*. Each of sex chromosomes after FISH with specific satellite probe (red color) is shown next to its scheme. Green probe represents RAYSI in all FISH experiments.

Figure 6. Sequence homogeneity of satellites. Clustering of sequence reads originating from male (blue) and female (red) plants using the CLANS software (Frickey and Lupas, 2004). Each dot corresponds to a single sequencing read. Reads were mapped by CLANS onto a spherical surface to best represent pairwise sequence similarity and positioned by authors for clear visualization. All satellites together (A), detailed visualization of RAE180 (B), RAYSI (C) and RAYSIII (D) and sequence logos of RAE180 where differences between male and female consensus sequences are marked by asterisks (E). Individual clusters were rotated manually into positions showing as much internal structure as possible.

Figure 7. Proportion of various TE and satellite families on transcriptome of repetitive fraction plotted against their proportion on repetitive genomic fraction. 100% was represented by all transcipts corresponding only to repetitive DNA (y-axes) or to all genomic repetitive fraction (x-axes). Expression was measured in *R. acetosa* male (squares) and female (triangles) leaves by Illumina platform RNA sequencing and proportion on repetitive DNA fraction was measured by Illumina platform DNA sequencing (see Material and Methods). Each repeat type is shown by different color.

Table S1. Localization, length and annotation of FISH probes used in this study.

Figure S1. Phylogenetic tree of CRM families using reverse transcriptase. Groups A (yellow), B (blue) and C (green) groups identified by Neumann et al (2011) are indicated. All *R. acetosa* CRM families (in red) cluster with group A.

Figure S2. Alignment of aminoacid sequences of C-terminus of integrase from CR retrotransposon subfamilies from *Rumex acetosa* together with two sequences from other plant species. The CR motif is marked and sequence logo indicates conserved aminoacids.

Table 1

Classification			Genome proportion	
Repeat type	**Super family**	**Family**	**male (%)**	**female (%)**
Retroelements	*Gypsy*	*Chromovirus-CRM*	5.73	5.38
		Chromovirus-Tekay/Del	0.34	0.36
		Athila	0.48	0.41
		Tat/Ogre	5.36	5.69
	Copia	*Maximus/ SIRE*	34.92	35.58
		Bianca	0.19	0.20
		TAR	0.10	0.10
	LINE		0.01	0.01
DNA transposons	*Mutator*		1.30	1.33
	CACTA		0.26	0.25
TOTAL transposable elements			**48.69**	**49.31**
Satellites	RAYS I		0.79	0.0002
	RAYS II		0.061	0
	RAYS III		0.44	0.00013
	RA160		0.61	0.76
	RA690		0.29	0.27
	RAE 180		2.72	1.04
	RAE 730		0.24	0.47
TOTAL satellites			**5.15**	**2.54**
rDNA			0.18	0.21

Figure 1

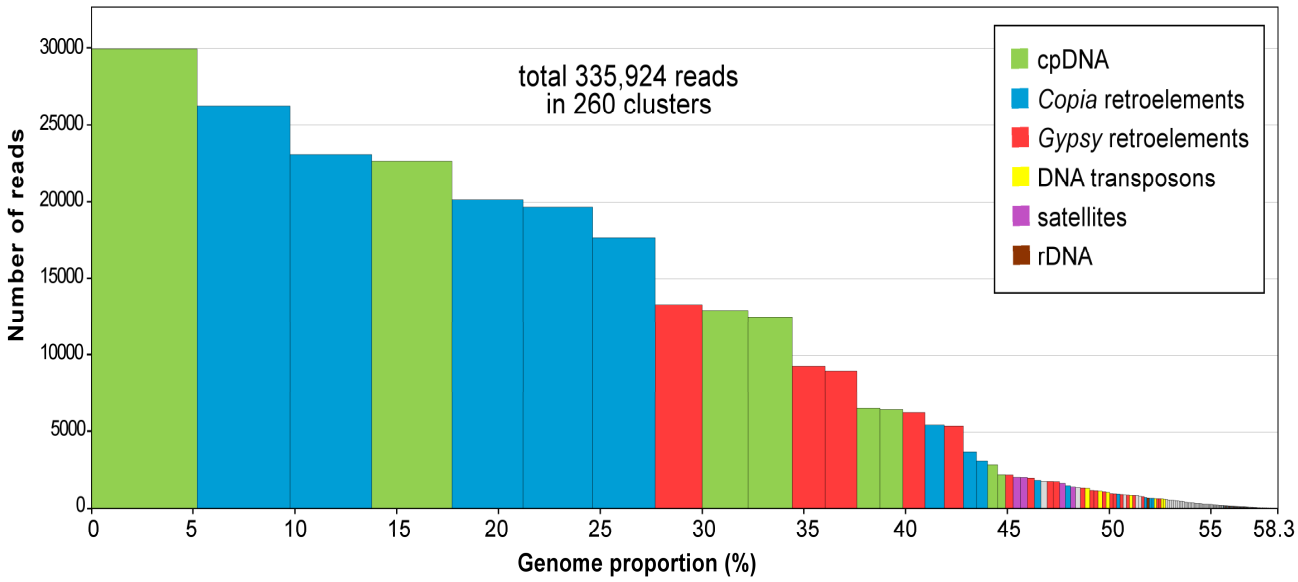
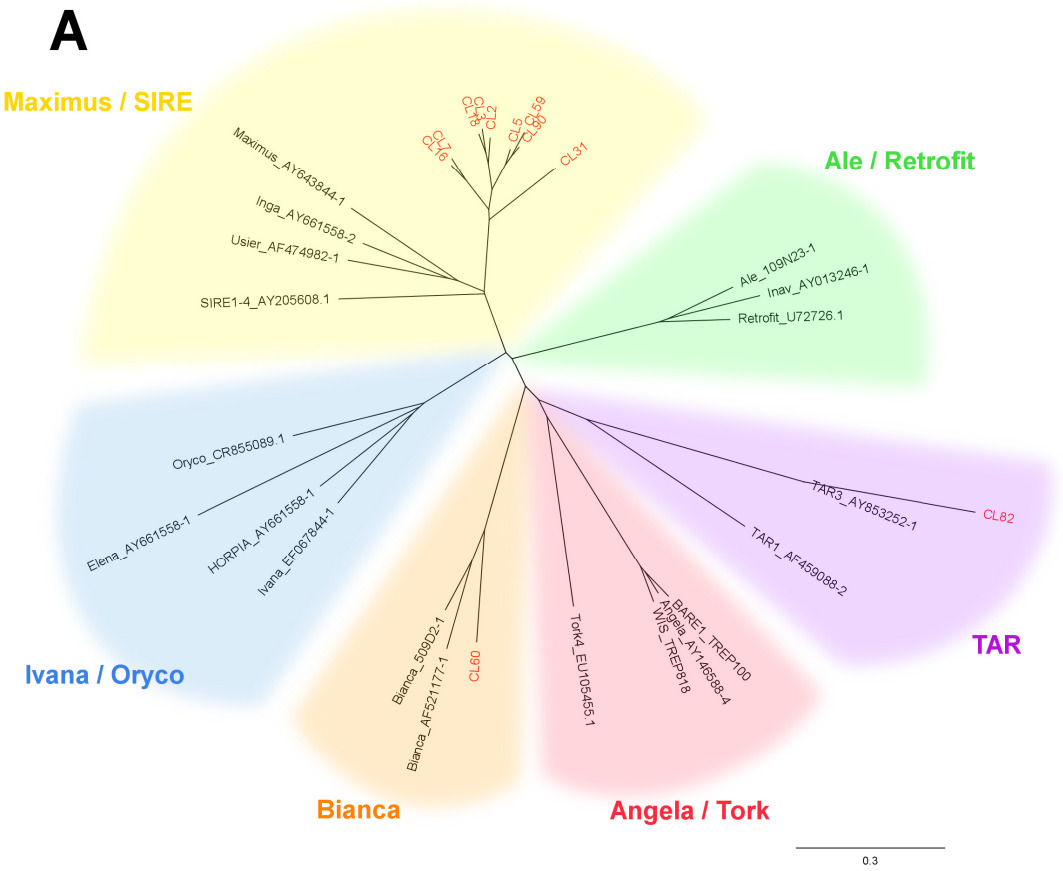


Figure 2



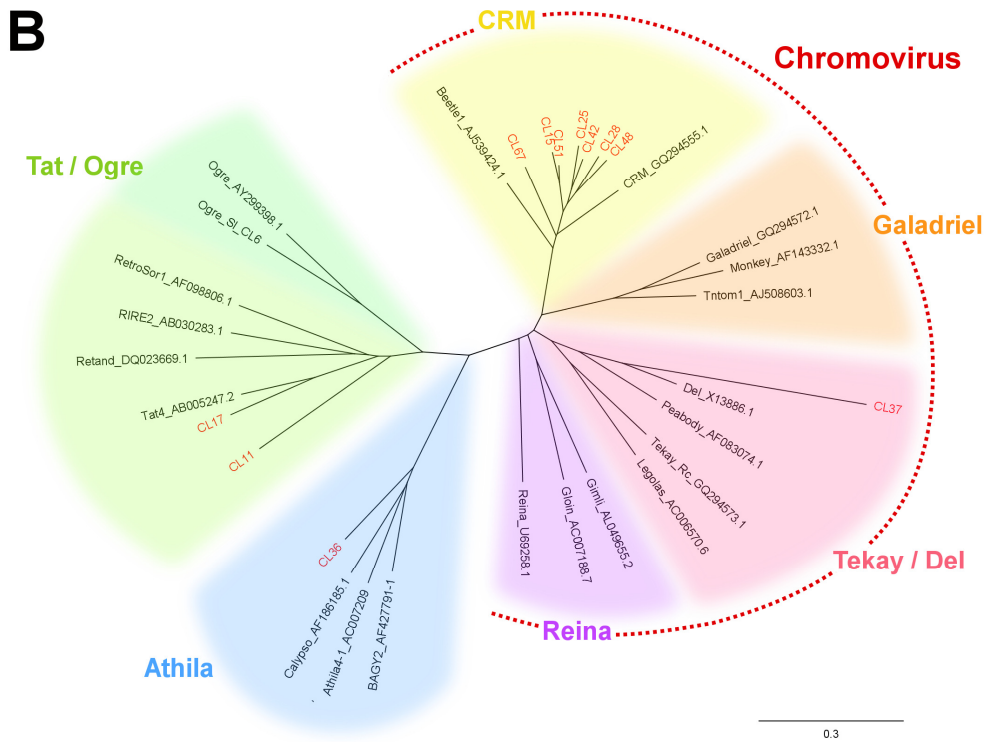


Figure 3

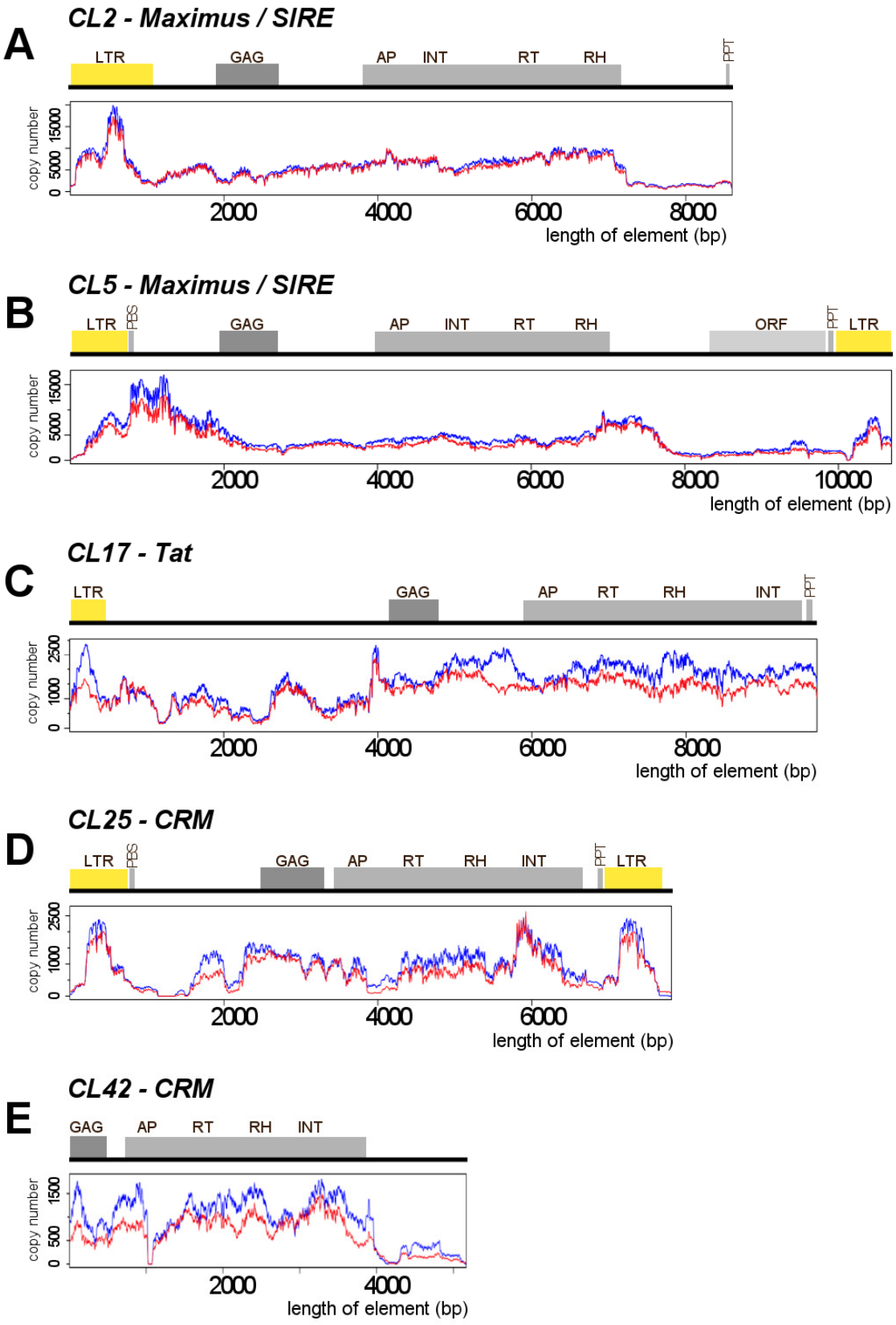


Figure 4

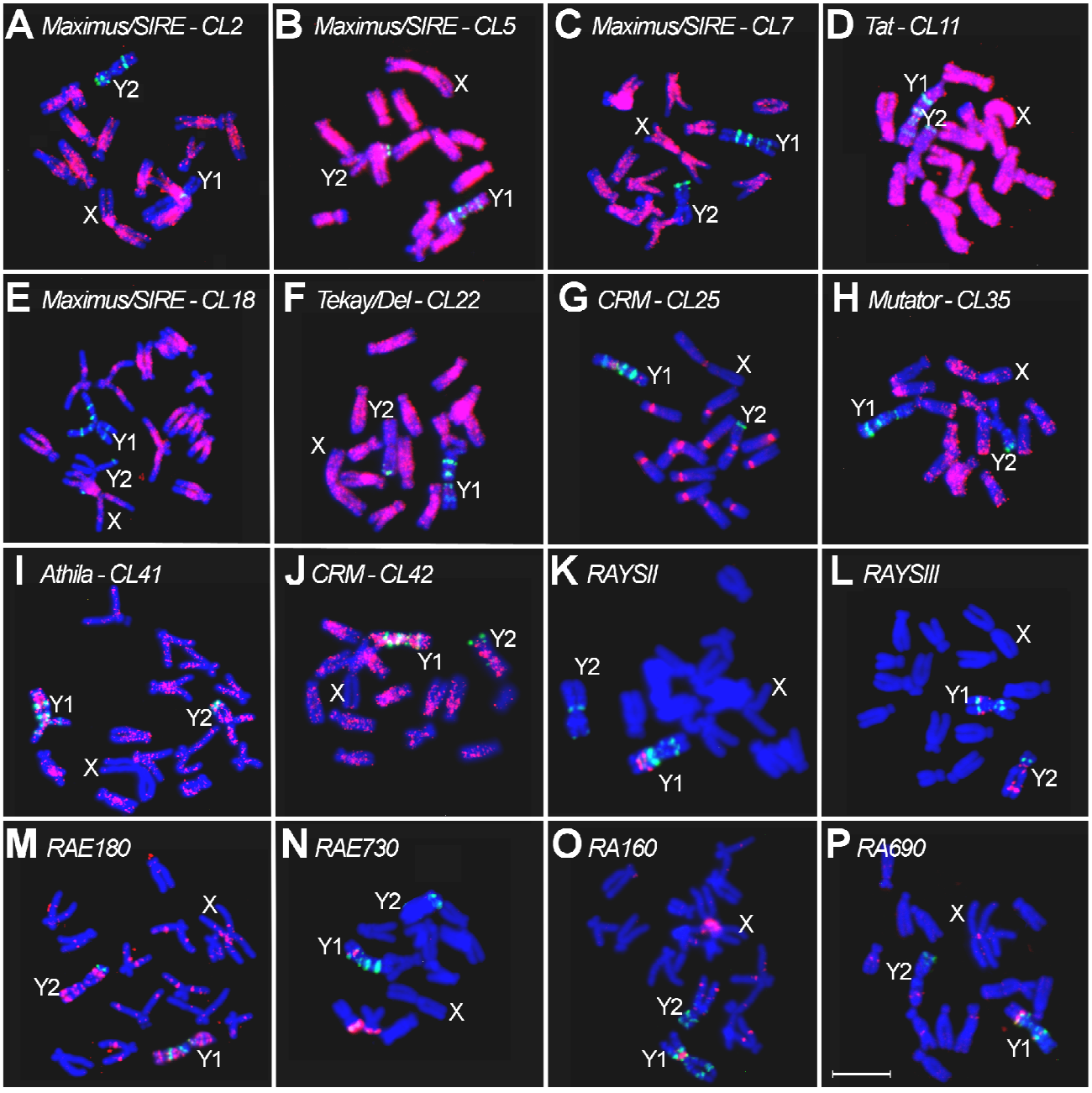


Figure 5

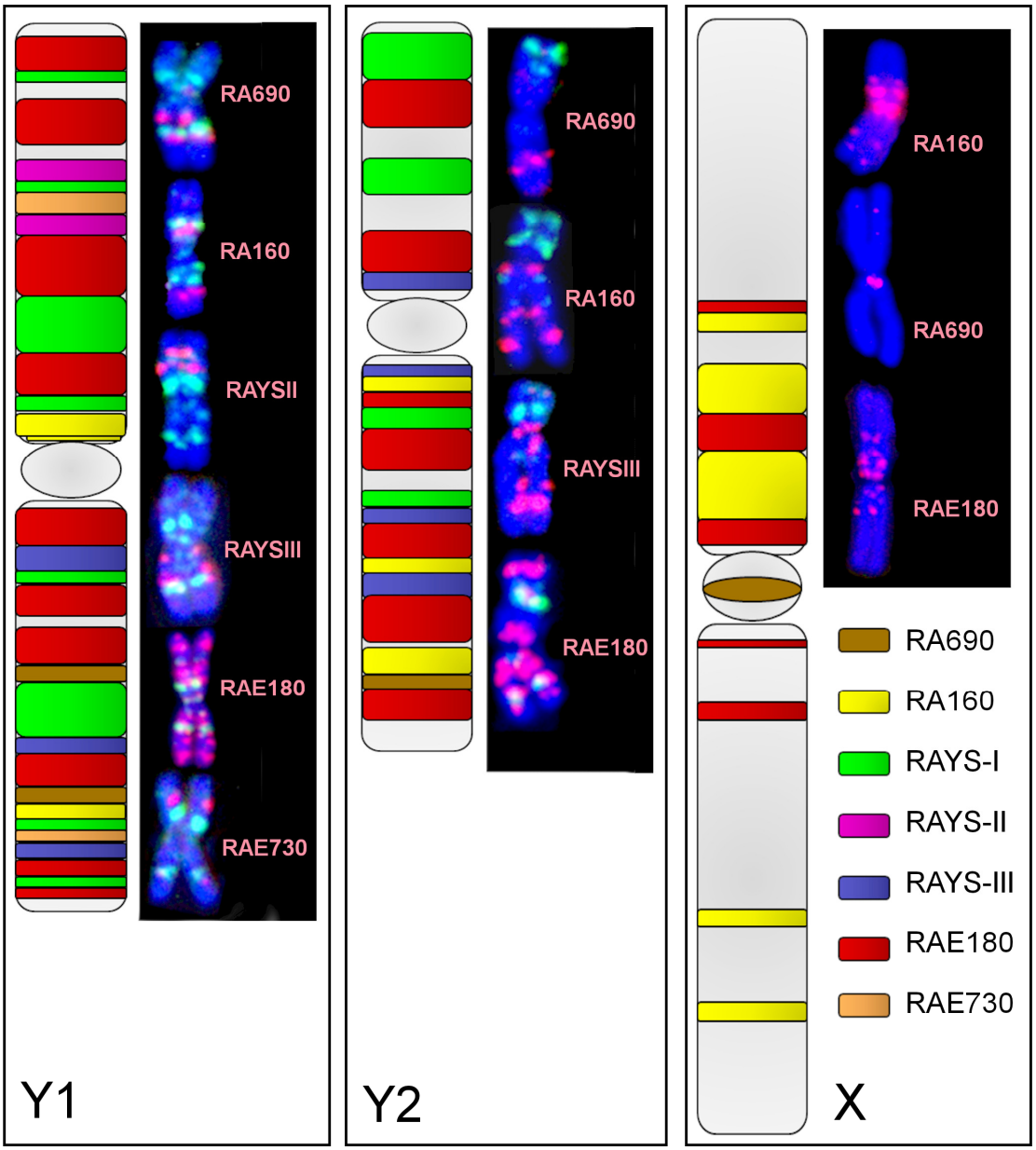


Figure 6

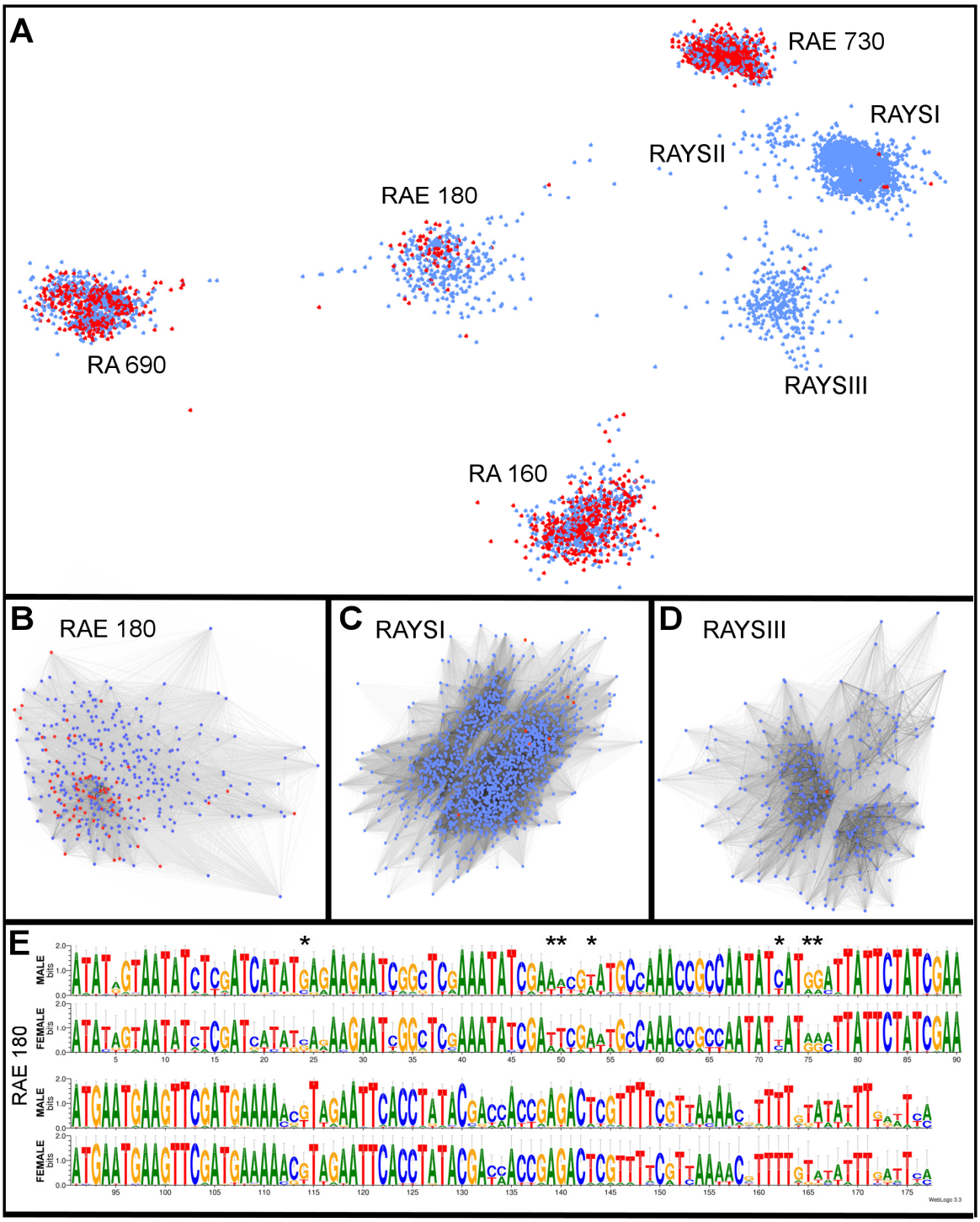


Figure 7