

satDNA Analyzer 1.2 as a Valuable Computing Tool for Evolutionary Analysis of Satellite-DNA Families: Revisiting Y-Linked Satellite-DNA Sequences of *Rumex* (Polygonaceae)

Rafael Navajas-Pérez¹, Manuel Ruiz Rejón¹, Manuel Garrido-Ramos¹, José Luis Aznarte², and Cristina Rubio-Escudero²

¹ Departamento de Genética, Facultad de Ciencias, Universidad de Granada
Campus de Fuentenueva s/n 18071, Granada, Spain

² Departamento de Ciencias de la Computación e Inteligencia Artificial,
Universidad de Granada
C/ Periodista Daniel Saucedo Aranda s/n 18071, Granada, Spain

Abstract. In a previous paper [1] we showed that Y-linked satellite-DNA sequences of *Rumex* (Polygonaceae) present reduced rates of evolution in relation to other autosomal satellite-DNA sequences. In the present paper, we re-analyze the same set of sequences by using the satDNA Analyzer 1.2 software, specifically developed by us for analysis of satellite DNA evolution. We do not only confirm our previous findings but also prove that the satDNA Analyzer 1.2 package constitutes a powerful tool for users interested in evolutionary analysis on satellite-DNA sequences. In fact, we are able to gather more accurate calculations regarding location of Strachan positions and evolutionary rates calculations, among others useful statistics. All results are displayed in a very comprehensive multicoloured graphic representation easy to use as an html file. Furthermore, satDNA Analyzer 1.2 is a time saving feature since every utility is automatized and collected in a single software package, so the user does not need to use different programs. Additionally, it significantly reduces the rate of data miscalculations due to human errors, very prone to occur specially in large files.

1 Introduction

Despite of sex chromosomes having evolved independently in several different groups of organisms (such as fishes-[2]-, reptiles -[3]- birds - [4]- , mammals -[5]-, insects - [6] - or plants - [7]), they seem to share some common evolutionary features [8]. The commonality is the presence of a pair of heteromorphic sex chromosomes in males (XY) consequence of differentiation and degeneration of Y chromosome. In fact, sex chromosomes undergo a process of gradual suppression of recombination that converts the Y chromosome in a relict chromosome with no counterpart to recombine with. Thus, this process leads to progressive divergence and to the erosion of the Y chromosome [9]. The final outcome of this process is the accumulation of mutations in dispensable regions of Y architecture (high rates of mutation have been

described in Y-linked genes- [10]; [11]) and the subsequent loss of function of many genes within the Y chromosome [12]. Y-chromosome degeneration is also accompanied by the accumulation of a set of diverse repetitive sequences such as mobile elements and satellite DNAs [13]; [14];[15]; [16]).

In the present work, we want to emphasize the role of satellite-DNA sequences in the Y degeneration process. Models of evolutionary dynamics for satellite DNA predict its accumulation in chromosomal regions where recombination rates are low [17]. Good examples of this are the non-recombining Y chromosomes ([16]; [18]). However, little is known about how this occurs or about how the absence of recombination affects the subsequent evolutionary fate of the repetitive sequences in the Y chromosome. In the present study, we focus on satellite-DNA sequences accumulation and evolution using as models the dioecious species of Genus *Rumex*, *R. acetosa* and *R. papillaris*, and by means of new computing utilities gathered together in satDNA Analyzer 1.2 package (<http://satdna.sourceforge.net>).

2 Antecedents and Motivation

Males of *R. acetosa* and *R. papillaris* have a karyotype with 15 chromosomes including a complex XX/XY₁Y₂ sex-chromosome system, while females have 14 chromosomes, being 2n= 12 + XX. During meiosis, the two Y chromosomes pair only with the ends of each X arm (own observations). All the data indicate that the Ys and the X chromosomes are highly differentiated and that the Y chromosomes are degenerated, as they are heterochromatic and rich in satellite-DNA sequences [19]. In fact, two satellite-DNA families have been found in both species to be massively accumulated in the Y chromosomes, the RAYSI family [20]; [21] and the RAE180 family [22]. Additionally, other satellite-DNA family, RAE730, has been described in heterochromatic segments of some autosome pairs [23].

To elucidate evolutionary rates of Y-linked sequences in relation to autosomal ones, we performed a comparative analysis between *R. acetosa* and *R. papillaris* sequences belonging to three different satellite-DNA families separately. Basically, we performed distance calculations according to the Jukes-Cantor method [24] and from these, we estimated evolutionary rates for both Y and autosomal-linked families. We found that Y-linked satellite sequences evolve two-fold to five-fold slower than autosomal-linked ones. Additionally, we proposed that shared polymorphisms should be removed when analyzing closely related species for more accurate calculations, since they might indicate ancestral variation before splitting of both species but not true divergence. In contrast, non-shared polymorphisms would be automorphies and represent different transitional stages in the process of intraspecific homogenization and interspecific divergence (for full details see [1]).

This study was reinforced by analyzing concerted evolution status (see [25]) of every three satellite-DNA sequences. We followed the method described in [26] which allows to analyze the pattern of variation at each nucleotide site between a pair of species for every of three marker studied (see Materials and Methods for further details).

Our aim is to confirm our previous findings by using the software satDNA Analyzer 1.2 and analyzing the same set of sequences described before. This supposes

a non-time demanding method since every step is automatized (location of different Strachan positions, removing shared polymorphisms from alignment and all statistics such as average consensus sequences, the average base pair contents, the distribution of variant sites, the transition to transversion rate and different estimates of intra and inter-specific variation) and collected in an unique package, so the user does not need to resort to different softwares. Additionally, the use of the software prevents from data miscalculations due to human errors, very prone to occur specially in large files.

3 Materials and Methods

3.1 Biological Material and Laboratory Procedures

Sequences analyzed in the present work were taken from the EMBL database (<http://www.ebi.ac.uk/embl/>) with accession numbers AJ580328 to AJ580343, AJ580382 to AJ580398, AJ580457 to AJ580463, AJ580468 to AJ580485, AJ580494 to AJ580496, AJ634478 to AJ634526, AJ634533 to AJ63456 and AJ639709 to AJ639741. These sequences belong to three different satellite-DNA families (RAE180, RAE730 and RAYSI) that we previously isolated in *R. acetosa* and *R. papillaris*. Biological material procedence and laboratory methodologies are fully described in [1].

3.2 Sequence Analysis

For the present work, we have revisited the sequences described above by using a new computing tool , satDNA Analyzer 1.2, a software package for the analysis of satellite-DNA sequences from aligned DNA sequence data implemented in C++. It allows fast and easy analysis of patterns of variation at each nucleotide position considered independently amongst all units of a given satellite-DNA family when comparing sets of sequences belonging to two different species. The program classifies each site as monomorphic or polymorphic, discriminates shared from non-shared polymorphisms and classifies each non-shared polymorphism according to the model proposed by [26] in six different stages of transition during the spread of a variant repeat unit toward its fixation (for a detailed explanation of this method, see also [27]). Briefly described, the class 1 site represents complete homogeneity between two species, whereas classes 2 to 4 represent intermediate stages in which one of the species shows polymorphism. The frequency of the new nucleotide variant at the site considered is low in stage 2 and intermediate in stage 3, while class 4 comprises sites in which a mutation has replaced the progenitor base in most members of repetitive family in the other species (almost fully homogenized site). Class 5 represents diagnostic sites in which a new variant is fully homogenized and fixed in all members of one of the species while the other species retain the progenitor nucleotide. Class 6 represents an additional step over stage 5 (new variants appear in some of the members of the repetitive family at a site fully divergent between two species). Furthermore, this program implements several other utilities for satellite-DNA analysis evolution such as the design of the average consensus sequences, the average base pair contents, the distribution of variant sites, the transition to transversion rate, and different estimates of intra and inter-specific variation.

Aprioristic hypotheses on factors influencing the molecular drive process and the rates and biases of concerted evolution can be tested with this program. Additionally, satDNA Analyzer generates an output file containing an alignment to be used for further evolutionary analysis by using different phylogenetic softwares. The novelty of this feature is that it allows to discard the shared polymorphisms for the analysis, which as we have demonstrated in [1], can interfere with the results when analyzing closely related species.

satDNA Analyzer 1.2 is freely available at <http://satdna.sourceforge.net> where supplementary documentation can be also found. satDNA Analyzer 1.2 has been designed to operate under Windows, Linux and MAC operating systems.

4 Results and Discussion

4.1 Subfamilies Detection

One of the main problems researchers have to face up is the recurrent formation of subfamilies in satellite-DNA sets of sequences, due to differential regions within the repeats or by the presence of diagnostic positions specifically fixed in one or another species ([28], [29]). The non-detection of these types of sequences before carrying out further evolutionary analysis can lead to the comparison of non-orthologous sequences and then to subsequent miscalculations. In this work we test the ability of satDNA Analyzer 1.2 to detect such cases. We previously described two paralogous RAYSI subfamilies in *R. acetosa*, called RAYSI-S and RAYSI-J [1]; [21]. We have used as input for our software a set of sequences of RAYSI isolated from *R. acetosa* genome. The study of Strachan stages included as a feature of our software reveals the existence of 72 diagnostic (fixed or almost fixed) positions, what shows the capacity of satDNA Analyzer 1.2 to discriminate both subfamilies (see Figure 1A). This approximately corresponds with our previous estimation of 83 of such as sites. Additionally, 20 sites are in transition stage 6, indicating the beginning of a new cycle of mutation-homogenization. This is supported by the fact that the mean inter-family divergence between both types of sequences is around 18% (see Figure 1B) while the mean intra-family percentage of differences is 4.2% and 5.1% for RAYSI-S and RAYSI-J respectively. Both subfamilies additionally have diagnostic deletions found at different positions in the RAYSI monomers also recognized by satDNA Analyzer as irrelevant positions due to indels.

We then have divided the RAYSI sequences of *R. acetosa* in two different files for RAYSI-J and RAYSI-S respectively. For further analysis, we combined both sets of sequences with RAYSI sequences belonging to *R. papillaris*. The study of diagnostic sites shows that RAYSI sequences of *R. papillaris* belong to the RAYSI-J subfamily since they share more diagnostic positions with this subfamily than with RAYSI-S (see also [21]).

4.2 Evolutionary Analysis and Concerted Evolution

In the present work, we have analyzed the rate of concerted evolution of the three satellite-DNA families studied in *R. acetosa* and *R. papillaris*: the Y-linked RAYSI

We have also studied the transitional stages in the process of concerted evolution according to the Strachan model [26]. SatDNA Analyzer 1.2 distinguishes Strachan stages from 1 to 6. For practical purposes we grouped together stages 2 and 3 in *Initial Stages Class (ISC)* and stages 4 and 5 in *Fully or almost Fully Homogenized Class (FHC)*. RAE730 sequences show higher percentage of **FHC** sites (47 sites) in relation to Y-linked RAYSI-J and RAE180 (4 and 3 respectively). In fact, most positions in Y-linked sequences seem to be in **ISC** yet (312 for RAYSI-J and 33 for RAE180). As shown in Table 1, these calculations differ slightly to our previous results gathered manually. It is probably due to the fact that we considered some indels as positions in the manual calculations. However, this reveals more accuracy in results obtained by satDNA Analyzer 1.2, especially significant in long satellite-DNAs (see Table 1). The mean length of satellite-DNA families has been suggested to be 165 bp in plants [31], but significantly longer sequences have been described in both plants and animals (as the case of RAYSI in *Rumex* with 930-bp repeats- [20]- or some mammals- [32] described a satellite-DNA family with repeats of 2600 bp in bovids) for which satDNA Analyzer 1.2 would be especially suitable.

We have gathered data that correlate significantly with those in [1]. Particularly, we found that within the RAE180 repeat units approximately 59% of the sites represent shared polymorphisms between *R. acetosa* and *R. papillaris*. However, we detected only one nearly fixed difference (0.5% of the sites) between these two species and 17% of polymorphic transitional stages. These data contrast with those found for the RAE730 sequences. In this case, 5.5% of nucleotide sites represent

Table 1. Comparative between data from Navajas-Perez et al., 2005a (**stated as previous data**) and data gathered by satDNA Analyzer in this paper. (**Top**) Mean intraspecific variability and interspecific divergence of three satellite-DNA families considering shared polymorphisms (**SP**), (**Down**) Analysis after excluding shared polymorphic sites (see text for details). Notes: (**ISC**) Initial Stages Class and (**FHC**) Fully or Almost Fully Homogenized Class.

<i>With SP</i>	<i>Mean distance</i>		<i>Evolutionary Rate</i>	<i>Differences between species</i>	
	<i>Intraspecific (R.acetosa/R.papillaris)</i>	<i>Intersp.</i>		<i>FHC (stages 4+5)</i>	<i>ISC (stages 2+3)</i>
RAE730					
previous data	0.055/0.036	0.099			
satDNA Analyzer	0.055/0.036	0.099			
RAYSI-J					
previous data	0.048/0.054	0.063			
satDNA Analyzer	0.051/0.056	0.065			
RAE180					
previous data	0.195/0.203	0.228			
satDNA Analyzer	0.199/0.211	0.235			
<i>Without SP</i>					
RAE730					
previous data	0.046/0.029	0.088	22x10 ⁻⁹	47	281
satDNA Analyzer	0.046/0.028	0.087	21.65x10⁻⁹	28	222
RAYSI-J					
previous data	0.037/0.042	0.047	11.74x10 ⁻⁹	3	407
satDNA Analyzer	0.036/0.043	0.047	11.63x10⁻⁹	4	312
RAE180					
previous data	0.036/0.037	0.045	11.25x10 ⁻⁹	3	74
satDNA Analyzer	0.028/0.029	0.033	8.25x10⁻⁹	1	33

shared polymorphic sites, while 4% are fixed differences between *R. acetosa* and *R. papillaris* and 30% transitional stages. Clearly, the data support the contention that the rate of concerted evolution is lower for the RAE180 satellite DNA at the Y chromosomes than for the RAE730 autosomic satellite DNA, as it is for RAYSI sequences, since *R. acetosa* and *R. papillaris* differ by only 0.4% of the sites and show 33% transitional stages. However, as opposed to RAE180, RAYSI sequences of the two species share only 6% of polymorphisms. This difference in the number of shared polymorphisms could be explained by the fact that RAE180 sequences are older than RAYSI, and therefore have accumulated a higher number of ancestral polymorphisms. Recent data gathered using Southern-blot hybridization may indicate that RAE180 sequences indeed have an older origin than do RAYSI sequences (own observations).

Additionally, we have tested this software with different sets of sequences gathering same and satisfactory results. However, these results were out of the purposes of this paper and are not shown. In the present work, we do demonstrate the utility of satDNA Analyzer in sets of sequences with main problems when carrying out evolutionary analysis on satellite DNAs, which are: low rates of concerted evolution and subfamilies formation. To summarize, satDNA Analyzer 1.2 constitutes a unique tool for evolutionary analysis of satellite-DNA. In this work we have proved that aprioristic hypotheses on factors influencing the molecular drive process and the rates and biases of concerted evolution can be tested with this program, as comparative analysis of rate between Y-linked and autosomal sequences or subfamily detection. Furthermore, satDNA Analyzer 1.2 supposes a non-time demanding method since every utility is automatized and collected in a unique package, so the user does not need to resort to different softwares. The results are displayed in a very comprehensive multicoloured graphic representation easy to use as an html file (see Figures 1A and 1B). Additionally, the use of the software prevents from data miscalculations due to human errors, very prone to occur specially in large files. Other utilities not shown in this work (as for example design of the average consensus sequences, the average base pair contents, the transition to transversion rate) are included in the software, constituting a complete package for satellite-DNA researchers.

Acknowledgments

This work was supported by Grant CGL2006-00444/BOS awarded by the Ministerio de Educación y Ciencia, Spain. R. N-P was granted by Plan Propio of University of Granada and is currently a Fulbright Postdoctoral Scholar of the Ministerio de Educación y Ciencia, Spain, at Plant Genome Mapping Laboratory in the Genetic Applied Genetic Technologies Center at the University of Georgia, Athens, GA.

References

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

2. Devlin, RH. & Nagahama, Y. (2002). Sex determination and sex differentiation in fish: an overview of genetic, physiological and environmental influences. *Aquaculture* 208:191-364.
3. Bull, JJ. (1980). Sex determination in Reptiles. *The Quarterly Review of Biology*, 55(1): 3-21.
4. Fridolfsson, AK., Cheng, H., Copeland, NG., Jenkins, NA., Liu, HC., Raudsepp, T., Woodage, T., Chowdhary, B., Halverson, J. & Ellegren, H. (1998). Evolution of the avian sex chromosomes from an ancestral pair of autosomes. *PNAS*, 95(14):8147-8152.
5. Marshall Graves, JA. & Watson, JM. (1991). Mammalian sex chromosomes: Evolution of organization and function. *Chromosoma*, 101(2):63-68.
6. Traut, W. & Marec, F. (1996). Sex chromatin in Lepidoptera. *The Quarterly Review of Biology*, 71(2):239-256.
7. Ruiz Rejón, M. (2004). Sex chromosomes in plants. Encyclopedia of Plant and Crop Sciences Vol. IV of Dekker *Agropedia* (6 Vols.). Marcel Dekker Inc., New York, pp. 1148-1151.
8. Charlesworth, B. (1996). The evolution of chromosomal sex determination and dosage compensation. *Curr. Biol.*, 6(2):149-162.
9. Filatov, DA., Moneger, F., Negrutiu, I. & Charlesworth, D. (2000). Low variability in a Y-linked plant gene and its implications for Y-chromosome evolution. *Nature*, 404, 388-390.
10. Shimmin, LC., Chang, BH. & Li, CH. (1994). Contrasting rates of nucleotide substitution in the X-linked and Y-linked zinc finger genes. *J Mol Evol.*, 39(6):569-578.
11. Filatov, DA. (2005). Substitution rates in a new *Silene latifolia* sex-linked gene, SlsX/Y. *Mol Biol Evol.* 22(3):402-408. Erratum in: *Mol Biol Evol.*, 22(4):1157.
12. Charlesworth, D. (2002). Plant sex determination and sex chromosomes. *Heredity*, 88: 94-101.
13. Steinemann, M. & Steinemann, S. (1997). The enigma of the Y chromosome degeneration: TRAM, a novel retrotransposon is preferentially located on the neo-Y chromosome of *Drosophila miranda*. *Genetics*, 145:261-266.
14. Bachtrog, D. (2003a). Adaptation shapes patterns of genome evolution on sexual and asexual chromosomes in *Drosophila*. *Nat. Genet.* 34, 215-219.
15. Bachtrog, D. (2003b). Accumulation of *Spock* and *Worf*, two novel non-LTR retrotransposons, on the neo-Y chromosome of *Drosophila miranda*. *Mol. Biol. Evol.*, 20:173-181.
16. Skaletsky, H. et al. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, 423:825-837.
17. Charlesworth, B., Sniegowski, P. & Stephan, W. (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* 371:215-220.
18. Jobling, MA. & Tyler-Smith, C. (2003) The human Y chromosome: an evolutionary marker comes of age. *Nat. Rev. Genet.*, 4(8):598-612.
19. Ruiz Rejón, C., Jamilena, M., Garrido-Ramos, M.A., Parker, J.S., Ruiz Rejón, M. (1994). Cytogenetic and molecular analysis of the multiple sex chromosome system of *Rumex acetosa*. *Heredity*, 72:209-215.
20. Shibata, F., Hizume, M. & Kuroki, Y. (1999). Chromosome painting of Y chromosomes and isolation of a Y chromosome-specific repetitive sequence in the dioecious plant *Rumex acetosa*. *Chromosoma*, 108:266-270.
21. Navajas-Pérez, R., Schwarzacher, T., de la Herrán, R., Ruiz Rejón, C., Ruiz Rejón, M. & Garrido-Ramos, MA. (2006). The origin and evolution of the variability in a Y-specific satellite-DNA of *Rumex acetosa* and its relatives. *Gene*, 368:61-71.
22. Shibata, F., Hizume, M. & Kuroki, Y. (2000a). Molecular cytogenetic analysis of supernumerary heterochromatic segments in *Rumex acetosa*. *Genome*, 43:391-397.

23. Shibata, F., Hizume, M. & Kuroki, Y. (2000b). Differentiation and the polymorphic nature of the Y chromosomes revealed by repetitive sequences in the dioecious plant, *Rumex acetosa*. *Chromosome Res.*, 8: 229–236.
24. Jukes, T.H. & Cantor, C.R. (1969). Evolution of protein molecules. In: Munro, H.D. (ed), *Mammalian protein metabolism*. Academic press, New York, 21-132.
25. Dover, G. (1982). Molecular drive: a cohesive mode of species evolution. *Nature*, 299, 111–117.
26. Strachan, T., Webb, D. & Dover, G. (1985). Transition stages of molecular drive in multiple-copy DNA families in *Drosophila*. *Chromosoma*, 108:266–270.
27. Pons, J., Petitpierre, E. & Juan, C. (2002) Evolutionary dynamics of satellite DNA family PIM357 in species of the genus *Pimelia* (Tenebrionidae, Coleoptera). *Mol. Biol. Evol.*, 19:1329–1340.
28. Willar, HF. (1991). Evolution of alpha satellite. *Current Opinion in Genetics and Development*, 1:509-514.
29. Garrido-Ramos, MA., Jamilena, M., Lozano, R., Ruiz-Rejón, C. & Ruiz-Rejón, M. (1994). Cloning and characterization of a fish centromeric satellite DNA. *Cytogenetic and Cell Genetics*, 65:233-237.
30. Navajas-Pérez, R., de la Herrán, R., López González, G., Jamilena, M., Lozano, R., Ruiz Rejón, C., Ruiz Rejón, M. & Garrido-Ramos, MA. (2005b). The evolution of reproductive systems and sex-determining mechanisms within *Rumex* (Polygonaceae) inferred from nuclear and chloroplastidial sequence data. *Mol. Biol. Evol.*, 22(9):1929-1939.
31. Macas, J., Meszaros, T. & Nouzova, M. (2002). PlantSat: a specialized database for plant satellite repeats. *Bioinformatics* 18(1):28-35.
32. Singer, MF. (1982). Highly repeated sequences in mammalian genomes. *Int. Rev. Cytol.* 76:67-112.