HeteroGenome

Database of Genome Periodicity

User Manual

What kind of information can a user find in the HeteroGenome Database?

HeteroGenome is a non-redundant catalog of DNA periodicity regions in a number of complete genomes of various organisms. It shows the results of an automated spectral-statistical approach to the genome-wide heterogeneity (latent periodicity) search. The significance level of revealed heterogeneities is 10^{-6} and below. Heterogeneity regions listed in the database are potential regions of a latent periodicity indicated mainly by approximate tandem repeats.

A specified Period Length, in which statistically significant heterogeneity has been determined, presents an estimation of periodicity pattern size. The Exponent shows the number of copies for this pattern in the revealed periodicity region. The average invariance for copies of the estimated periodicity pattern is characterized by a Preservation Level (ranging from 0.4 to 1.0) for the characters across period positions.

Latent periodicity and heterogeneity

The spectral-statistical approach [1-3], used for the analysis of genomes in HeteroGenome, is based on χ^2 – statistics for testing homogeneity in DNA sequences at significance level characteristic for approximate tandem repeats which sequences are obviously heterogeneous. However, significant heterogeneity is a necessary, but insufficient condition of periodicity. As the results of the latent periodicity search, in automatic mode are undoubtedly heterogeneous sequences and some additional analysis is required in order to confirm their periodic structure, these results are refered to as heterogeneities.

References

1. Chaley M., Kutyrkin V. Model of perfect tandem repeat with random pattern and empirical homogeneity testing poly-criteria for latent periodicity revelation in biological sequences. *Mathematical Biosciences*, 2008, Vol. 211, Issue 1, pp. 186-204.

2. Chaley M.B., Nazipova N.N., Kutyrkin V.A. Statistical Methods for Detecting Latent Periodicity Patterns in Biological Sequences: The Case of Small-Size Samples. *Pattern Recognition and Image Analysis*. 2009, Vol. 19, No. 2, pp. 358-367.

3. Chaley M., Nazipova N., Teplukhina E., Tyulbasheva G., Kutyrkin V. Statistical Methods for Detecting Latent Periodicity in Biological Sequences: Solving a Problem of Small-Size Samples. In: *2009 IEEE International Conference on Bioinformatics and Biomedicine: the book of abstracts of the BIBM 2009 (Nov 1 – 4, 2009, Washington D.C.)*. Los Alamitos: IEEE Computer Society, 2009, pp. 92-96.

The logical structure of the Database

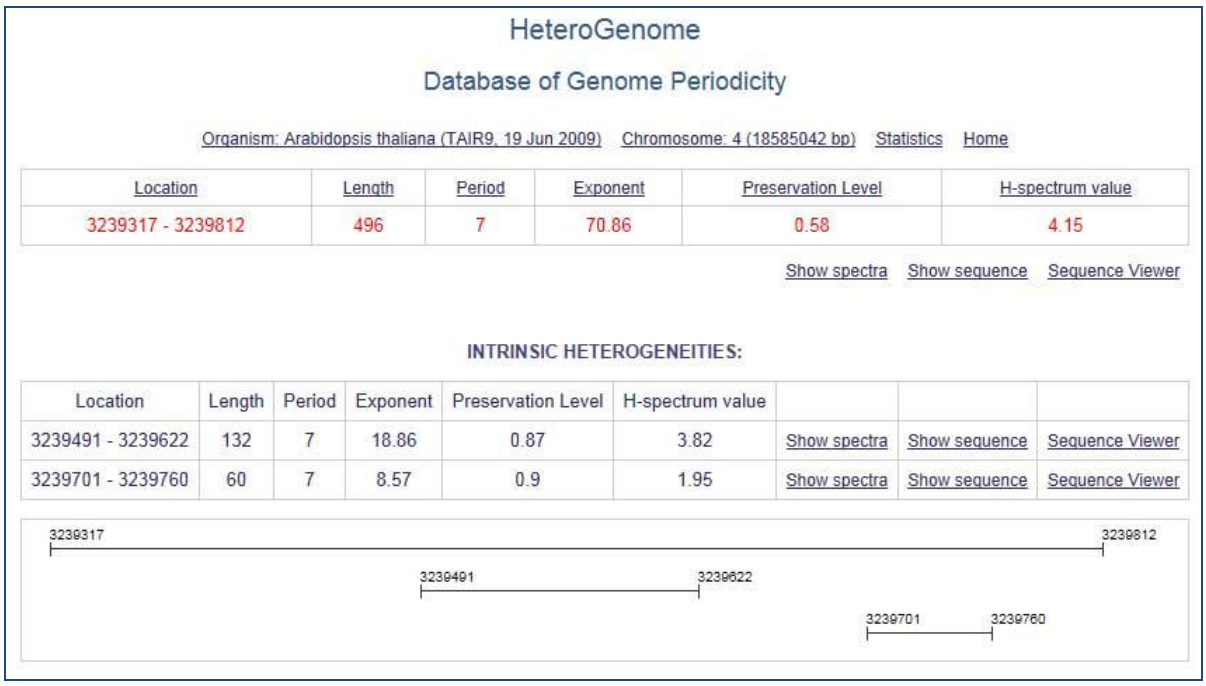


Figure 1. An example of two-level structure of logical record in HeteroGenome database is shown. The upper table shows the parameters of sequence presenting a group at the nonredundant level (highlighted with red). The lower table contains data for other sequences in the Group. A scaled scheme of the group is placed below.

A conception of logical record as a Group of overlapping regions of significant heterogeneity (latent periodicity) has been developed for non-redundant and visual presentation of the HeteroGenome content. The longest region serves as a group representer. Group members which lie inside the representer are regarded as its intrinsic heterogeneities. The latter ones either point to the most structured fragments of the region of latent periodicity, or, if revealed periods are neither equal to nor multiple of the period of Group representer, they facilitate correct data interpretation, offering to reconsider the Group as consisting of individual subgroups.

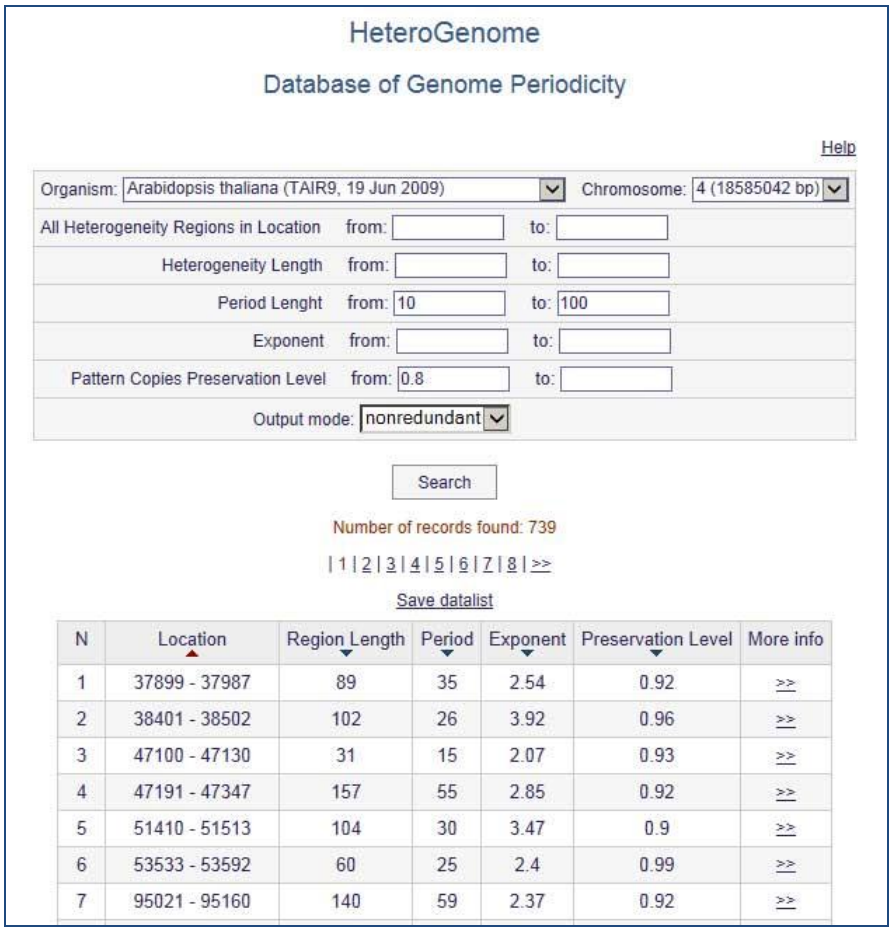
See Figure 1 for example of such a logical record. Two levels can be distinguished in the record. The first one corresponds to the sequence, representing the Group, the second level contains its intrinsic heterogeneities – the sequences of determinated periodic structure. The sequences of the first level are not intersecting and constitute non-redundant content of the HeteroGenome database.

Search in the Database

To simplify user's work the search system is organized trough all the fields of the database record. The Figure 2 shows an example of query on the chromosome IV of Arabidopsis. Minisatellites with high level of pattern copies preservation (>=0.8) are requested. The search is carried out at the first level over group representers (the parameter "output mode" = "nonredundant" is set). The pattern copies Preservation Level for them is in interval ranging from 0.8 to maximal available value 1.0 which is set by default.

The results of a query are produced as a table sorted by the field "Location". Fragments are listed according to their placement order along the chromosome. The data can be sorted by clicking on the arrow on the top of every column in descending (down arrow) or ascending (up arrow) mode of values in the column. The arrow on top of the column, according to which results are sorted, is marked with red.

To download results of a search query click on "Save datalist". To see the details of a certain region just click on the link "more info" in the corresponding line.



N	Location	Region Length	Period	Exponent	Preservation Level	More info
1	37899 - 37987	89	35	2.54	0.92	>>
2	38401 - 38502	102	26	3.92	0.96	>>
3	47100 - 47130	31	15	2.07	0.93	>>
4	47191 - 47347	157	55	2.85	0.92	>>
5	51410 - 51513	104	30	3.47	0.9	>>
6	53533 - 53592	60	25	2.4	0.99	>>
7	95021 - 95160	140	59	2.37	0.92	>>

Figure 2. A search query example in HeteroGenome database. The minisatellites (Period Length is ranging from 10 to 100) with high values of Pattern Copies Preservation Level (ranging from 0.8 to maximal available value 1.0 which is set by default) are requested on the chromosome IV of *Arabidopsis thaliana*. Only group representers are searched for (the parameter "output mode" = "nonredundant" is set).

What kind of information can a user obtain about a certain region of periodicity?

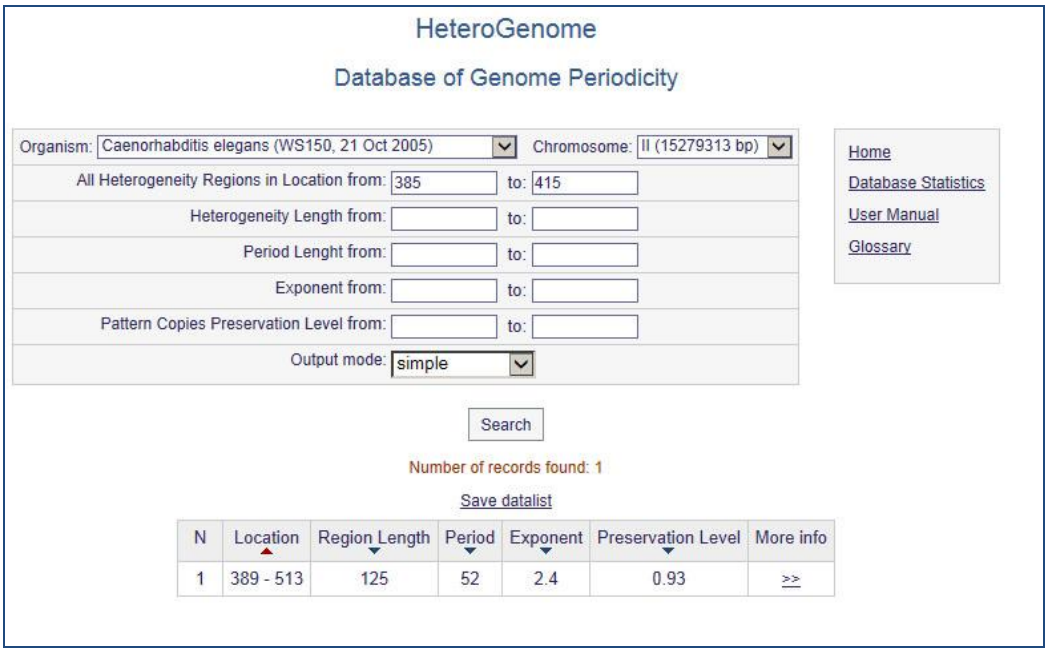


Figure 3. An example of a searching request for something lying in the specified area (between positions 385 and 415 from 5'-end) of chromosome II of *C. elegans*

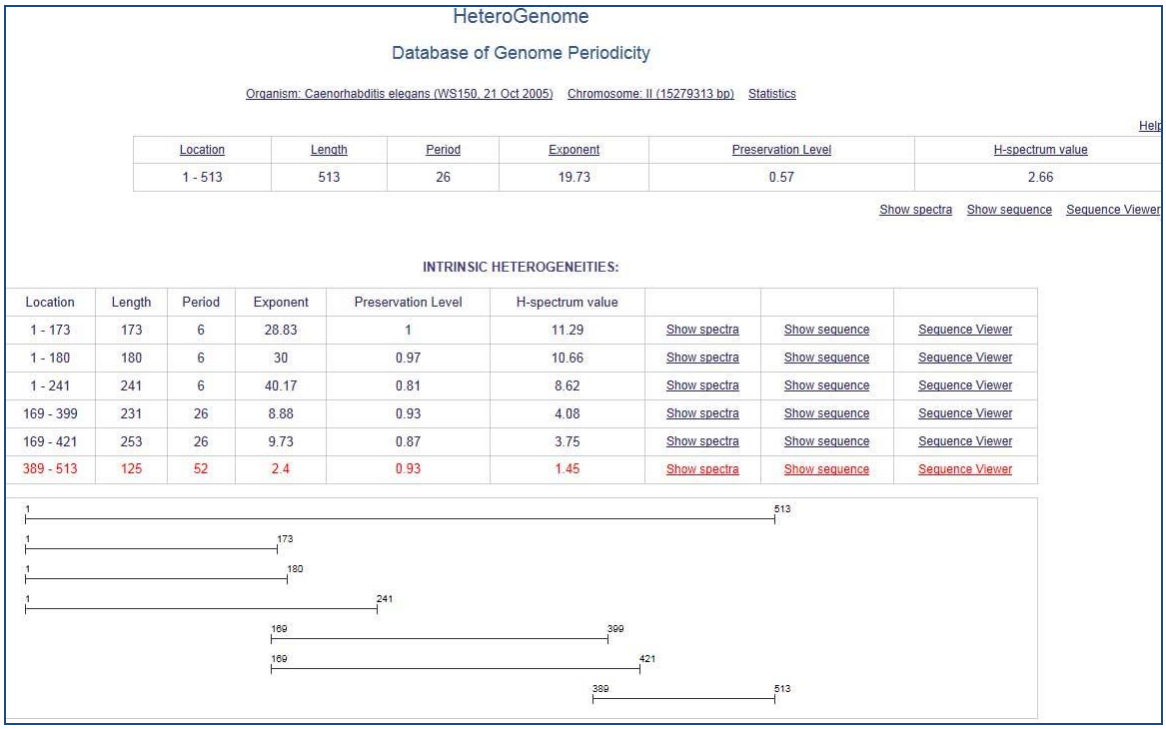


Figure 4. An example of information output showing the details for latent periodicity region on the chromosome II of *C. elegans,* corresponding to the link "more info" in the Figure 3.

For each sequence listed in the database search result (Figures 2, 3) a particular page is produced. Each page shows the contents of the logical record describing the region of user's interest (the region is highlighted with red). Any additional information can be obtained from the HeteroGenome database and the NCBI resources. Information from each source is placed in different windows according to the following links:

- "Organism" – query to the NCBI Genome database concerning the genome of a studied organism;
- "Chromosome" – request applying the NCBI Map Viewer for the chromosome map display;
- "Statistics" – request of summary statistics for the latent periodicity details on specified chromosome of the organism of interest;

- "show spectra" – display of the spectra values for two spectral-statistical parameters of DNA sequences with latent periodicity (see *pl*-spectrum and *H*-spectrum in Glossary), on the basis of which the periodicity pattern size has been estimated;
- "show sequence" - display of the sequence in the form of the sequence profile (column of the sequence substrings) where each substring is of length equaled to the pattern size estimate;
- "Sequence Viewer" - request applying the NCBI Sequence Viewer graphical interface for the genome annotation content of corresponding region of a chromosome.

Why are the inner regions collected in the HeteroGenome database?

In the result of automated procedure revealing the regions with significant heterogeneity (latent periodicity), the longest region and the regions with the dominant values of two spectral-statistical parameters (see *pl*-spectrum and *H*-spectrum in Glossary) are chosen from the pool of collocated regions with the same or multiple pattern length. As a divergence of periodic structure happens irregularly along a tandem repeat, ~~so~~ the inner regions point to repeat fragments with the most preserved structure.

What kind of information can a user obtain from analysis of the spectra for the two statistical parameters?

Period Length presented in the HeteroGenome database has been found by an automated procedure based on the analysis of two spectral parameters: the pattern copies preservation level (see *pl*-spectrum in Glossary) and the spectrum of heterogeneity manifestation (see *H*-spectrum in Glossary). The first dominating peak in the *pl*-spectrum, as a rule, points to an estimate of period length. The length of the interval between regularly repeating peaks in both spectra can also be considered as an estimate of period length. Moreover, in order to determine the accurate length of the period, corresponding values of the parameters of other members in group can be considered as probable estimations for period length as well. Anyway, some additional analysis should be done to verify such estimate, as described in the next section.

What kind of information can a user obtain on the web page "show sequence"?

To reinvestigate the length of the period, users can consider the sequence on the screen displayed in the form of a column of substrings with the length equaled to given period estimate. This kind of sequence presentation is called a profile. By changing the period length and varying the length of flanks of investigated region users can determine visually, which estimation is corresponding mostly to the periodicity of the region of interest.

Let us consider an example of an inner fragment of a group which covers 5'-beginning of the chromosome II of *C. elegans* (from 1 to 513 bp) and integrates sequences with the period lengths of 6, 26, 52 bps (see Figure 4).

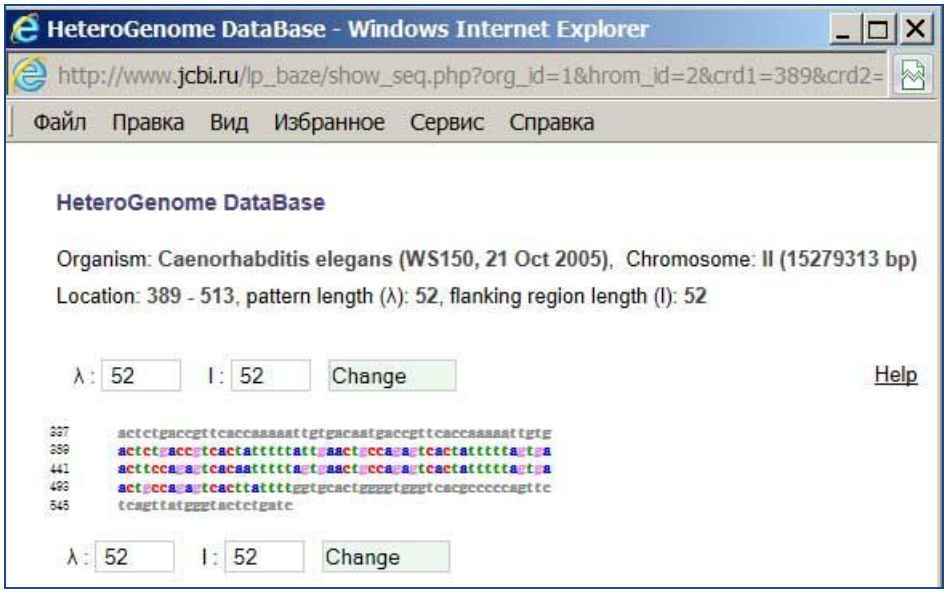


Figure 5. A sequence profile in the "show sequence" window. It corresponds to a fragment highlighted by red in the Figure 4.

As one can see in the Figure 5, the sequence of the given region (Location: 389 – 513) is displayed in the form of a column of substrings (a profile). Each of the substrings has the length equaled to an estimate of the length of periodicity pattern (λ). The region is shown together with flanking sequences of length l (which is, by default, equal to the length of the periodicity pattern l=λ). The flanking sequences are displayed in grey, and the nucleotides of the region are multicolored. The length of the period and the length of the flanks of the investigated region can be redefined and immediately tested. If λ=26 and l=400 are given, then by clicking on the "Change" button user will get the profile presented in the Figure 6.

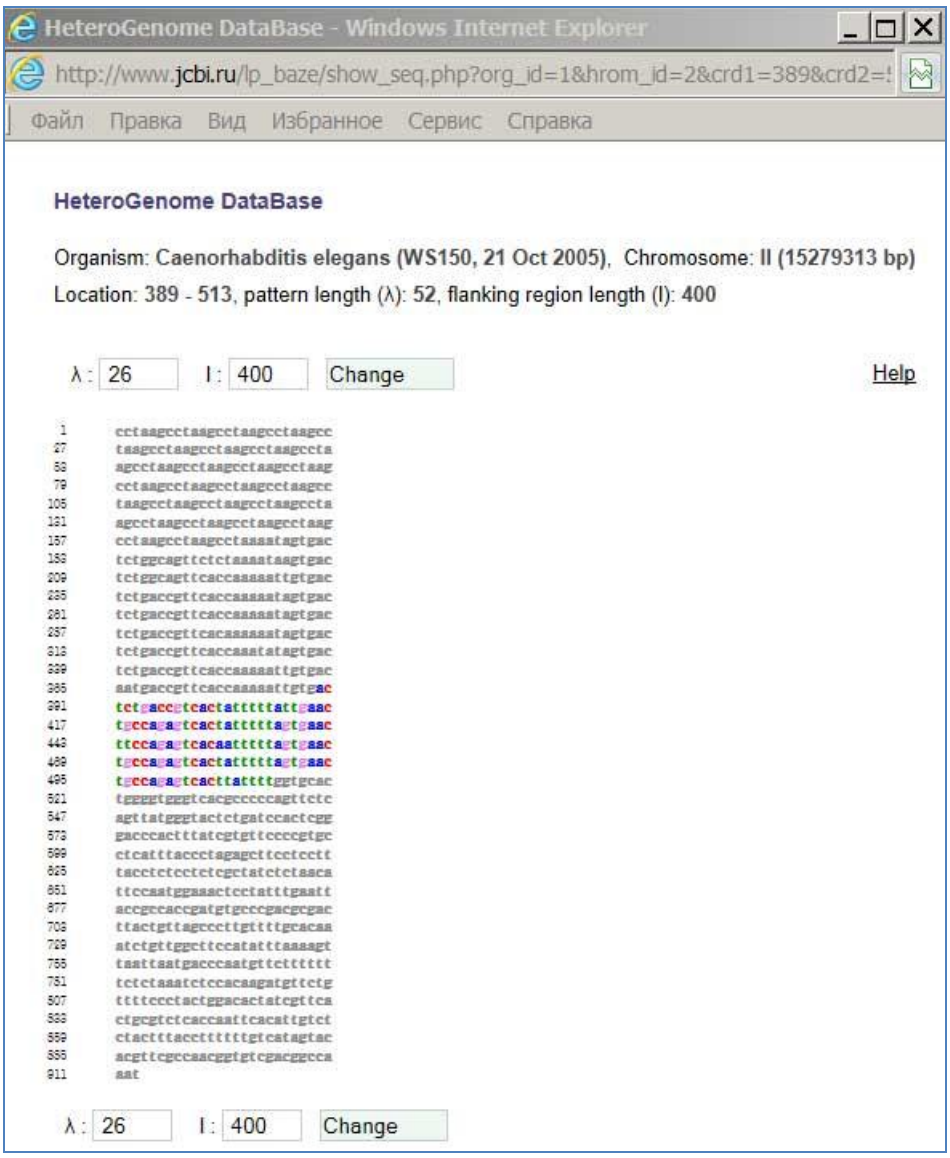


Figure 6. An example of another profile for the sequence in the Figure 5 with redefined period length λ=26 and flanks' length l=400.

As Figure 6 shows, three different tandem repeats with patterns listed in the Table 1 are located in the region (from 1 to 513 bp) of C. elegans chromosome II. These results were obtained due to the possibility of additional analysis for information (shown in the Figure 4). Finally, the structure of the considered fragment was more precisely defined. And three different tandem repeats were recognized inside the fragment with accurate repeats borders and their correct patterns (see Table 1).

Table 1.

Location	True Pattern Length	Pattern
1 - 168	6	cctaag
169 – 399	26	tcaccaaaaattgtgactctgaccgt
400 – 513	26	cactatttttagtgaactgccagagt

What kind of information can a user see on the page "Statistics"?

There are three levels of overview of the data collected in the HeteroGenome database. The first (intergenomic) level combines the data over all genomes, the second (genomic) level summarizes the data on different genomes separately and the last (chromosomal) level presents detailed description of database content concerning each chromosome in the genome.

At the intergenomic level (see Figure 7) a user can see summary results describing non-redundant genome coverage by significant heterogeneity regions (latent periodicity coverage). The genomes coverage is considered from three viewpoints. The number of the sequences forming the genome coverage (number of Groups) and the summary length of genome coverage by latent periodicity sequences are presented. The share of the genome coverage in the whole genome is shown as well. The data for micro- mini- and megasatellites, are also shown, How rich a repertory of Pattern Lengths is in each genome is demonstrated.

Organism		*A. thaliana*	*C. elegans*	*D. melanogaster*	*S. cerevisiae*	*S. cerevisiae* MT
Length, bp		119146348	100269917	120381546	12070900	85779
Genome Coverage by Latent Periodicity Sequences, %	micro*	1.66	2.56	3.13	2.49	18.17
	mini	1.05	3.03	0.57	0.75	4.66
	mega	0.82	1.23	0.41	0.29	1.39
	total	3.54	6.81	4.11	3.53	24.22
Genome Coverage by Latent Periodicity Sequences, bp	micro	2001025	2532147	4003842	291478	15583
	mini	1235737	2966952	587316	88266	3999
	mega	1014809	1201040	518539	19396	1194
	total	4251571	6700139	5109697	399140	20776
Number of the Sequences in Genome Coverage	micro	27293	27178	67150	3693	100
	mini	6770	11459	5365	389	31
	mega	502	692	256	12	1
	total	34565	39329	72771	4094	132
Number of Units in Repertory of Pattern Lengths		407	709	326	66	28

*Correspond to revealed period length L, for micro- ($2 \le L \le 10$), mini- ($10 < L \le 100$) and mega- ($100 < L \le 2000$) satellites

Figure 7. Genome coverage by significant heterogeneity regions (latent periodicity coverage).

On the genomic level one can see the diagrams of structural content for heterogeneity (latent periodicity) regions in each genome (see Figure 8). They correspond to the revealed period length *L*, for micro- ($2 \le L \le 10$), mini- ($10 < L \le 100$) and mega- ($100 < L \le 2000$) satellites. The coverage (as a percentage) of the genome by heterogeneity (latent periodicity) regions with various values of preservation level is shown as separate histograms. The preservation level value in interval $0.4 \le pl \le 0.7$ (red) corresponds to highly diverged tandem repeats; those in interval $0.7 < pl \le 0.8$ (green) corresponds to moderately diverged tandem repeats; that in $0.8 < pl \le 0.9$ (yellow) corresponds to slightly diverged tandem repeats and that in $0.9 < pl \le 1.0$ (blue) corresponds to perfect tandem repeats.

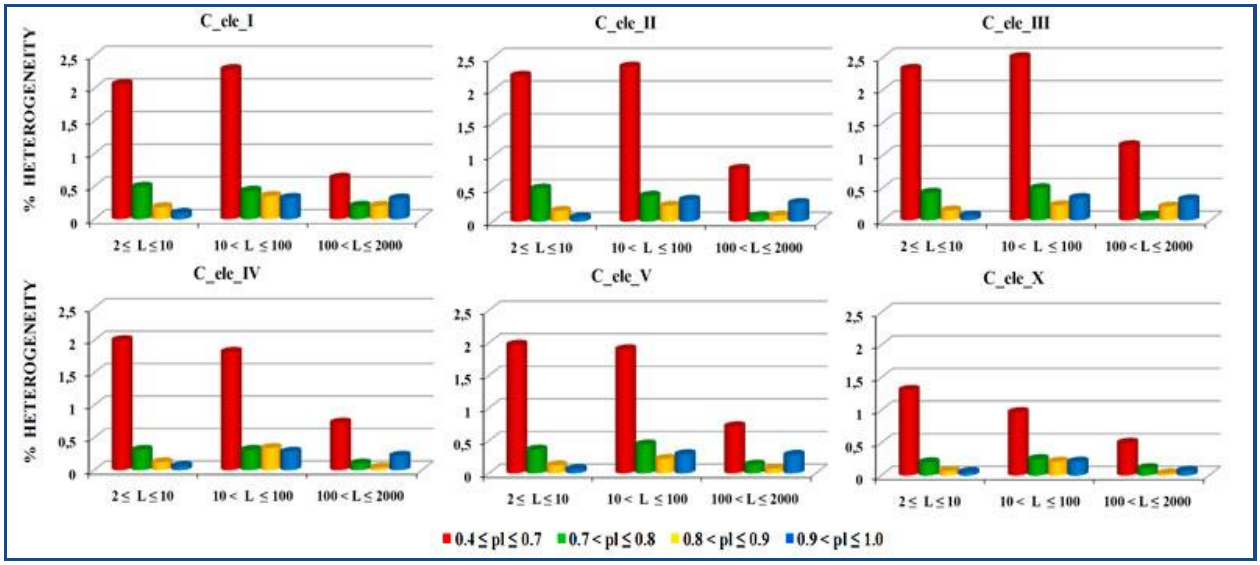


Figure 8. An example of the coverage (as a percentage) of *C. elegans* genome by heterogeneity (latent periodicity) regions with various values of preservation level *pl.*

Moreover, the histograms of density distribution of latent periodicity regions along the chromosomes are demonstrated (see Figure 9). Each chromosome was divided into consecutive fragments of the length equal to 0.5% of the whole chromosome length. Each bar corresponds to the percentage of the sum of latent periodicity regions (in

nucleotides), falling inside the fragment, from the chromosome length. Summarizing all bar heights in histograms gives an estimate of general percentage of latent periodicity on chromosome.

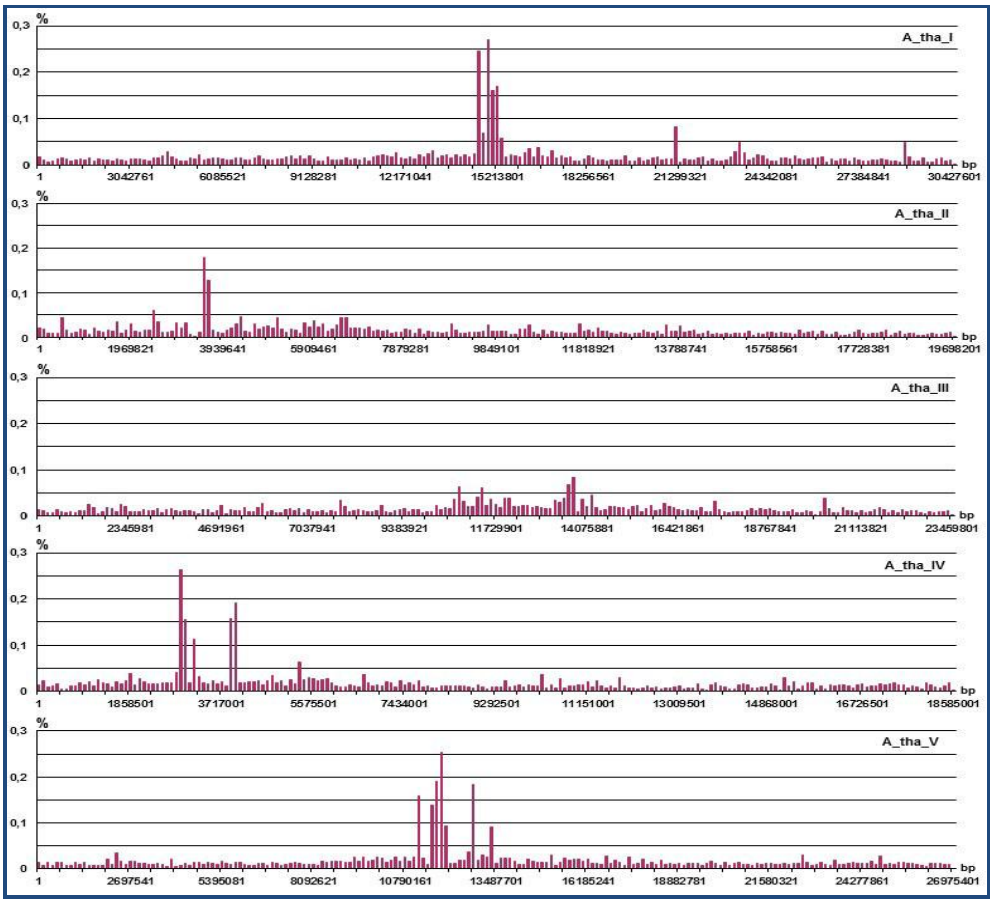


Figure 9. An example of the histograms showing density distribution of latent periodicity regions along the chromosomes of *A. thaliana*.

On the chromosomal level histogram of density distribution of latent periodicity regions along the specified chromosome is presented (see Figure 10). For each type of periodicity (for micro- ($2 \le L \le 10$), mini- ($10 < L \le 100$) and mega- ($100 < L \le 2000$) satellites) the corresponding histogram is shown in blue.

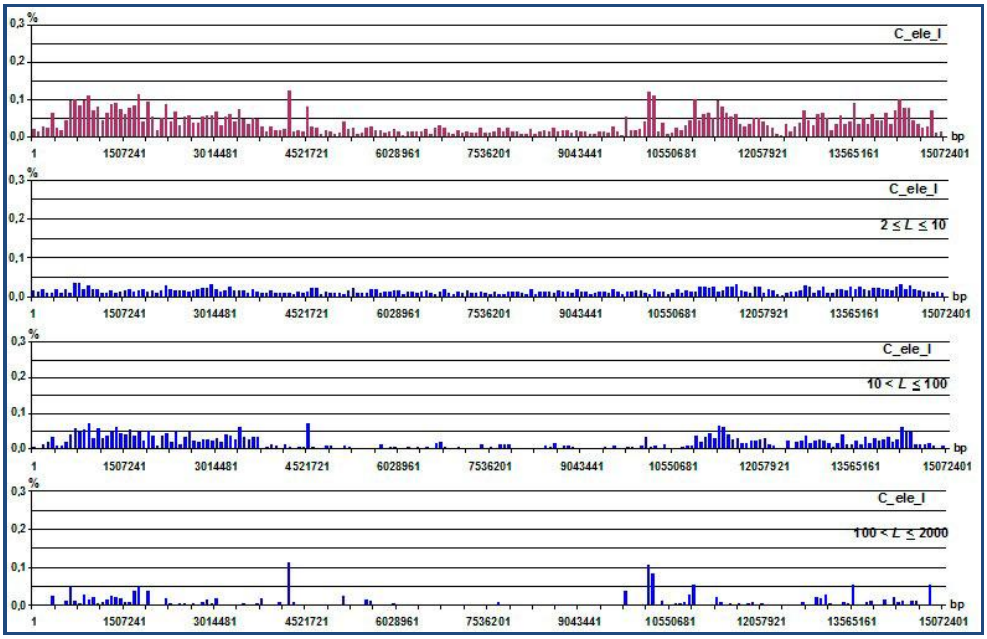


Figure 10. An example of a whole density distribution of the latent periodicity regions on *C. elegans* chromosome I (at the top) and its decomposition in accordance with three types of periodicity – micro- ($2 \le L \le 10$, mini- ($10 < L \le 100$) and mega- ($100 < L \le 2000$) satellites.