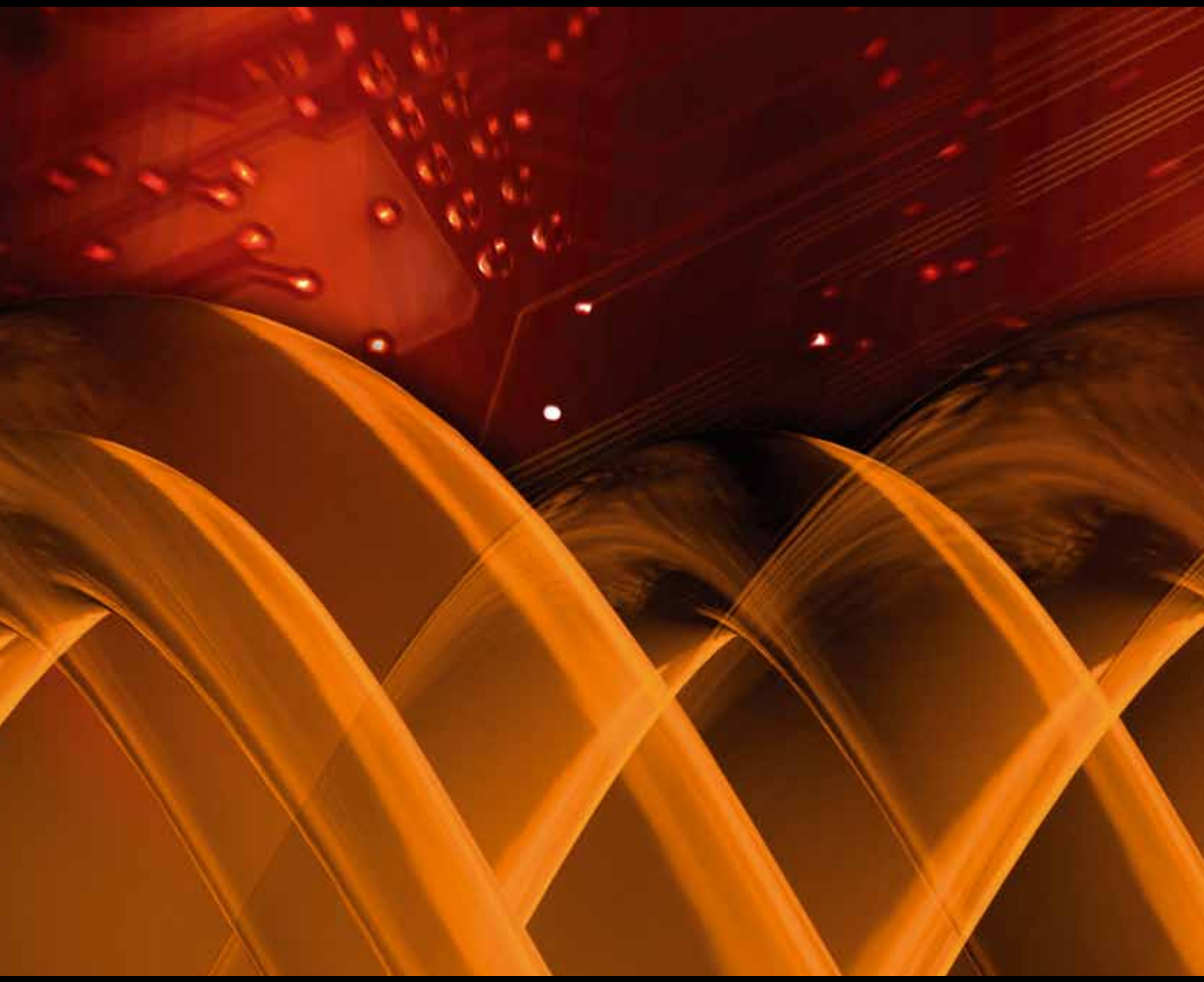
International Journal of Genomics

EXTRACTING EVOLUTIONARY INSIGHTS USING BIOINFORMATICS

Guest Editors: Dmitry Sherbakov, Yuri Panchin, and Ancha Baranova



Extracting Evolutionary Insights Using Bioinformatics

International Journal of Genomic

Extracting Evolutionary Insights Using Bioinformatics

Guest Editors: Dmitry Sherbakov, Yuri Panchin, and Ancha Baranova

Editorial Board

Contents

Hindawi Publishing Corporation
International Journal of Genomics
Volume 2013, Article ID 376235, 2 pages
http://dx.doi.org/10.1155/2013/376235

Editorial

Extracting Evolutionary Insights Using Bioinformatics

Dmitry Sherbakov,[1,2] Yuri Panchin,[3,4] and Ancha Baranova[5,6]

[1] *Laboratory of Molecular Systematics, Limnological institute, Ulan-Batorskaya Str. 3, Irkutsk 664033, Russia*
[2] *Department of Biology, Irkutsk University, Sukhe-Bator 5, Irkutsk 664003, Russia*
[3] *Department of Mathematical Methods in Biology, Belozersky Institute, Moscow State University, Vorbyevy Gory 1-40, Moscow 119991, Russia*
[4] *Institute for Information Transmission Problems, Russian Academy of Sciences, Bolshoi Karetny Pereulok 19-1, Moscow 127994, Russia*
[5] *Research Centre for Medical Genetics, Russian Academy of Medical Sciences, Moscow 115478, Russia*
[6] *School of System Biology, George Mason University, Fairfax, VA 22030, USA*

Correspondence should be addressed to Ancha Baranova; abaranov@gmu.edu

Received 22 October 2013; Accepted 22 October 2013

According to much quoted Russian/American geneticist Theodosius Dobzhansky, "*Nothing in biology makes sense except in the light of evolution*" [1]. Coined in 1973, this famous phrase remains up to date. Fascinating variety of the phenotypes seen as either visible or subtle differences in species, populations, and individual organisms inhabiting our planet attracts an unwavering inquiry. The results of these inquiries are shedding some light both at intricate molecular programs continuously executed within living bodies and at checks and balances within homeostatic ecosystems. In fact, every transformative technology so far invented to improve human medicine or other areas of applied science was immediately embraced by traditional biologists and contributed to the extraction of evolutionary insights. Some years ago, that happened with PCR. Recently, we entered systems biology driven turn of an inquiry spiral. This turn is enabled by an advent of NextGen sequencing. Importantly, a few final steps on this spiral became computationally heavy. Hence, the aid of bioinformatics was summoned, and computationally derived conclusions came in droves.

While the mining of completed genomes remains an obvious source is information on evolutionary processes, we also see the emergence of a novel kind of the full-genome studies, ones that aim at the tinkering with the evolutionary theory itself. By their scale and potential, these studies may be compared to the two previous technological developments that changed the history of evolutionary studies. The first of these methodological revolutions is a comparative analysis of traits that was started by Lamarck [2] and Darwin [3] and still is a source of novel evolutionary inferences. The second one was the introduction of the molecular analysis (see [4]) that approximately coincided with the establishment of the synthetic theory of evolution. Since that, the methods of the analysis of molecular traits kept progressing. By now, it is clear that molecular-based inquiry opened whole new field of questions which may be asked and answered, for example, one that relates to the evolution of the intracellular parasites, microbial communities, or populations of the cells within human body burdened with a tumor. The introduction of NextGen sequencing dramatically increased the amount of information that may be extracted from the sample of biological material and added the whole new level of integration of organismal data, thus enabling the systems biology in its true sense.

Nowadays, one may be certain that widespread adoption of NextGen sequencing will uplift the evolutionary theory to new heights, however, where will it bring us even in next few years is substantially harder to predict. We should not forget that evolutionary genomics is still an emerging field in biology. For now, it does not have a road map, a master plan, or even a template—one who starts the analysis of data collected from the study of living system never knows where

this analysis might end. This volume is yet another attempt to provide a bird's view at bioinformatics-driven forays into the field of evolutionary biology.

However, it is an opinion of the Editors that presented eclectic collection of papers is not without its merits. In case of existing data, even a mere push to extract an evolutionary insight enables the scientist to see the pattern that would be very difficult to discern otherwise. So much better it is for the studies specifically designed to solve one or another biological enigma. In this issue, we are glad to present an eclectic collection of papers that cover a range of topics from analysis of intron evolution across kingdoms to the study of the divergence in salmonid species and to comparative genomics of cancer-specific genes. These manuscripts are a sampler of future insights yet to come from the troves of genome sequences thoroughly dissected with bioinformatics tools. Enjoy!

Dmitry Sherbakov
Yuri Panchin
Ancha Baranova

References

[1] T. Dobzhansky, "Nothing in biology makes sense except in the light of evolution," *American Biology Teacher*, vol. 35, no. 3, 1973.

[2] J. B. Lamarck, *Zoological Philosophy*, Macmillan, London, UK, 1914.

[3] C. Darwin, *On the Origin of Species by Means of Natural Selection, Or the Preservation of Favoured Races in the Struggle for Life*, John Murray, London, UK, 1st edition, 1859.

[4] J. C. Avise, *Molecular Markers, Natural History and Evolution*, Sinauer Associates, London, UK, 2nd edition, 2004.

Hindawi Publishing Corporation
International Journal of Genomics
Volume 2013, Article ID 173616, 6 pages
http://dx.doi.org/10.1155/2013/173616

Research Article

Comparative Analysis of Context-Dependent Mutagenesis in Humans and Fruit Flies

Sofya A. Medvedeva,[1] Alexander Y. Panchin,[2] Andrey V. Alexeevski,[1,3,4] Sergey A. Spirin,[1,3,4] and Yuri V. Panchin[2,3]

[1] *Department of Bioengineering and Bioinformatics, Moscow State University, Vorbyevy Gory 1-73, Moscow 119992, Russia*
[2] *Institute for Information Transmission Problems, Russian Academy of Sciences, Bolshoi Karetny pereulok 19-1, Moscow 127994, Russia*
[3] *Department of Mathematical Methods in Biology, Belozersky Institute, Moscow State University, Vorbyevy Gory 1-40, Moscow 119991, Russia*
[4] *Department of Mathematics, Scientific-Research Institute for System Studies, Russian Academy of Sciences, Nakhimovskii prospekt 36-1, Moscow 117218, Russia*

Correspondence should be addressed to Alexander Y. Panchin; alexpanchin@yahoo.com

Received 1 April 2013; Accepted 7 July 2013

Academic Editor: Dmitry Sherbakov

In general, mutation frequencies are context-dependent: specific adjacent nucleotides may influence the probability to observe a specific type of mutation in a genome. Recently, several hypermutable motifs were identified in the human genome. Namely, there is an increased frequency of T>C mutations in the second position of the words ATTG and ATAG and an increased frequency of A>C mutations in the first position of the word ACAA. Previous studies have also shown that there is a remarkable difference between the mutagenesis of humans and drosophila. While C>T mutations are overrepresented in the CG context in humans (and other vertebrates), this mutation regularity is not observed in *Drosophila melanogaster*. Such differences in the observed regularities of mutagenesis between representatives of different taxa might reflect differences in the mechanisms involved in mutagenesis. We performed a systematical comparison of mutation regularities within 2–4 bp contexts in *Homo sapiens* and *Drosophila melanogaster* and found that the aforementioned contexts are not hypermutable in fruit flies. It seems that most mutation contexts affect mutation rates in a similar manner in *H. sapiens* and *D. melanogaster*; however, several important exceptions are noted and discussed.

1. Introduction

The average rates of point mutations in multicellular eukaryotic genomes are usually between 10^{-7} and 10^{-10} mutations per nucleotide per generation [1, 2]. However, the rates of point mutations may be dramatically altered by their genomic context. In some cases, this context-dependent change in mutation frequency can be attributed to known molecular mechanisms involved in mutagenesis. For example, the increased frequency of C>T mutations in the word CG in humans (and other vertebrates) is attributed to the methylation of cytosines by context-specific DNA methyltransferases [3]. This mutation regularity is absent in *D. melanogaster* [4], in which cytosine methylation occurs, but appears to be restricted to early embryonic development and is not specific to cytosines followed by guanines [5]. Many other examples of context-dependent mutagenesis have been reported [4, 6–9].

Recently, an increased rate of T>C mutations in the second position of the words ATTG and ATAG and an increased rate of A>C mutations in the first position of the ACAA word were reported in the human genome [10]. This was achieved by calculating the values called "minimal contrast" and "mutation bias" for 2–4 bp mutation contexts to evaluate if the addition of specific nucleotides to the 5′ or 3′ end of 1–3 bp words increases the probability of observing certain mutations in fixed positions. Mutation bias indicates the total excess (or deficiency) of mutations within a given

context. Minimal contrast indicates the excess (or deficiency) of mutations within a given context that cannot be explained by the excess (or deficiency) of mutations in one of its subcontexts.

H. sapiens and *D. melanogaster* are perspective model organisms for this kind of studies because of the vast amount of data on genetic variation that is available for them. The goal of our study was to compare the mutation regularities of *H. sapiens* and *D. melanogaster* in terms of "minimal contrast" and "mutation bias."

2. Methods

We searched for single nucleotide variable positions in intergenic sequences of 37 individual *D. melanogaster* genomes (multiple alignments obtained from http://genome.ucsc.edu/ [11]). *Drosophila sechellia* (droSec1, Oct. 2005) and *Drosophila erecta* (droEre2, Feb. 2006) genomic sequences were used as outgroups to reconstruct the ancestral states for the variable positions. *D. melanogaster* genome (dm3, Apr. 2006) was used as the reference.

2.1. Mutation Data. We assume that a mutation with a known direction within a known context has occurred in a specific position of the *D. melanogaster* genome if the following conditions are met.

(1) *D. sechellia* and *D. erecta* genomes have the same nucleotide aligned to this position (this nucleotide will be referred to as the "ancestral nucleotide").

(2) Among the 37 *D. melanogaster* genomes, some contain the ancestral nucleotide in this position, while some other genomes contain a different nucleotide.

(3) Only 2 genetic variants are present in this position for the 37 *D. melanogaster* genomes.

(4) The 3 bp upstream and downstream positions from these positions in the multiple alignment do not contain any substitutions or gaps.

Mutation bias and minimal contrasts for *D. melanogaster* were calculated for 2–4 bp mutation contexts using the methods described in [10]. Mutation bias, contrasts, and other data for *H. sapiens* were taken directly from [10].

2.2. Mutation Context and Subcontext. We denote the mutation context of mutation mut in position pos of the word W as {mut | pos, W}. For example, {C>T | 1, CG} represents a C>T mutation in the first position of the word CG. Mutation context $\{\text{mut} \mid \text{pos}', \text{W}'\}$ is called a subcontext of the context {mut | pos, W} if W' is a subword of W, and any mutation mut occurring in position pos of the word W is at the same time a mutation occurring in position pos' of the word W'. For example, {C>T | 1, CG} is a subcontext of {C >T | 2, ACG}.

2.3. Contrast. For each pair of context {mut | pos, W} and its subcontext $\{\text{mut} \mid \text{pos}', \text{W}'\}$, the value of contrast is given by the formula

$$\text{Contrast}\left(\{\text{mut} \mid \text{pos}, \text{W}\}, \{\text{mut} \mid \text{pos}', \text{W}'\}\right) = \frac{P_{\{\text{mut}|\text{pos},\text{W}\}}}{P_{\{\text{mut}|\text{pos}',\text{W}'\}}}. \quad (1)$$

Here, $P_{\{\text{mut}|\text{pos},\text{W}\}}$ and $P_{\{\text{mut}|\text{pos}',\text{W}'\}}$ are the conditional probabilities of observing mutation mut in the position pos of the word W and position pos' of word W', respectively, in a given dataset. Although these probabilities cannot be explicitly calculated without assumptions of the general probability of mutation per nucleotide in the genome, their ratio can be estimated by the following formula:

$$\frac{P_{\{\text{mut}|\text{pos},\text{W}\}}}{P_{\{\text{mut}|\text{pos}',\text{W}'\}}} = \frac{N_{\{\text{mut}|\text{pos},\text{W}\}}/P_{\text{W}}}{N_{\{\text{mut}|\text{pos}',\text{W}'\}}/P_{\text{W}'}}. \quad (2)$$

Here, P_{W} and $P_{\text{W}'}$ are the observed frequencies of words W and W', respectively, among all words of the same length. $N_{\{\text{mut}|\text{pos},\text{W}\}}$ and $N_{\{\text{mut}|\text{pos}',\text{W}'\}}$ are the observed numbers of mutation mut in position pos of word W and position pos' of the word W', respectively.

The ratio $P_{\text{W}}/P_{\text{W}'}$ estimates the probability for W' to be extended to W. This ratio coincides with the expected ratio $N_{\{\text{mut}|\text{pos},\text{W}\}}/N_{\{\text{mut}|\text{pos}',\text{W}'\}}$ under the hypothesis that mutations rates are the same in the context {mut | pos, W} and its subcontext $\{\text{mut} \mid \text{pos}', \text{W}'\}$. Therefore, if Contrast ({mut | pos, W},{mut | pos', W'}) is greater than 1, it indicates an increased mutation rate in the context {mut | pos, W} compared with the subcontext $\{\text{mut} \mid \text{pos}', \text{W}'\}$; while if Contrast ({mut | pos, W},{mut | pos', W'}) is less than 1, it indicates a decreased mutation rate.

2.4. Minimal Contrast. For a given context {mut | pos, W}, let us consider all of its subcontexts $\{\text{mut} \mid \text{pos}', \text{W}'\}$. The minimal contrast is the value MC = Contrast ({mut | pos, W},{mut | pos', W'}) such that the absolute difference |MC−1| is the lowest among all subcontexts $\{\text{mut} \mid \text{pos}', \text{W}'\}$. We did not study discontigvous contexts such as CNG and CNNG.

2.5. Mutation Bias. For any context {mut | pos, W}, there exists only one subcontext $\{\text{mut} \mid \text{pos}', \text{W}'\}$ such that the length of W' is equal to 1 (i.e., W' is the one-letter word consisting of the mutated letter). The mutation bias is the contrast of the given context and this subcontext.

2.6. Word Frequencies. We used two measures of *D. melanogaster* word frequencies. The first measure was obtained using complete aligned sequences of 37 *D. melanogaster*, the *D. sechellia*, and *D. erecta* genomes. For the second measure, we used conserved regions in which the ancestral nucleotide matches at least one of the *D. melanogaster* genetic variants, and no gaps or unread sequences are present in the multiple alignment. Word frequencies from the conserved regions were used for calculating mutation biases and contrasts.

Table 1: Comparison of nucleotide composition of complete alignments and conserved regions of *D. melanogaster*.

Nucleotide	Nucleotide fraction within all positions	Nucleotide fraction within conserved positions	Difference, %
a	0.2979	0.2901	2.6
t	0.2978	0.2899	2.7
c	0.2022	0.2100	−3.9
g	0.2021	0.2100	−3.9

3. Results and Discussion

The nucleotide composition of complete alignments and conserved regions (see Section 2) of *D. melanogaster* were similar (Table 1). We decided to use word frequencies within conserved regions of *D. melanogaster* for calculations of contrast and mutation bias.

Previous studies have shown that the representation of mutation data on a plot of mutation bias versus minimal contrast is useful for identifying important mutation contexts [10]. Mutation bias and minimal contrasts of mutation contexts in *D. melanogaster* are shown in Figure 1. The {A>C | 2, CACC} and {A>C | 3, CCA} mutation contexts have the highest minimal contrast values in *D. melanogaster*. Interestingly, the addition of C or G nucleotides to either end of the word CCA increases mutation bias of the A>C mutation, while the addition of A or T nucleotides to these words decreases mutation bias.

As shown in Table 2, mutation patterns differ between *D. melanogaster* and *H. sapiens* at the single nucleotide scale: *D. melanogaster* has a lower transition/transversion ratio. Moreover, the G>T (C>A) transversion in *D. melanogaster* comprises a much larger fraction of mutations than the A>G (T>C) transition, which is consistent with previous findings [4].

One of the mechanisms by which G>T (C>A) transversions occur is through the formation of 8-Oxoguanine [12] caused by reactive oxygen species [13] or ultraviolet irradiation [14]. In eukaryotes, the damaged DNA is repaired with the help of DNA glycosylase OGG1. This enzyme removes the 8-oxoguanine, forming a DNA apurinic-apyrimidinic site, which is then recognized by other proteins of the DNA repair system. If further reparation does not occur, the apurinic-apyrimidinic site will be complemented with an adenine nucleotide during DNA replication, resulting in a C>A mutation. Another protein with DNA glycosylase activity for 8-hydroxyguanine, called dOgg1, was also described in *D. melanogaster* [15].

Another factor that might be responsible for increased G>T (C>A) transversion rates in *D. melanogaster* is aflatoxin B1. Aflatoxin B1 is known to induce base substitutions in DNA [16, 17], especially G>T (C>A) transversions. It is a product of a fungus from the *Aspergillus* genus, which grows on fruits and grains in a humid climate; thus, it is quite possible that *D. melanogaster* is exposed to this toxin.

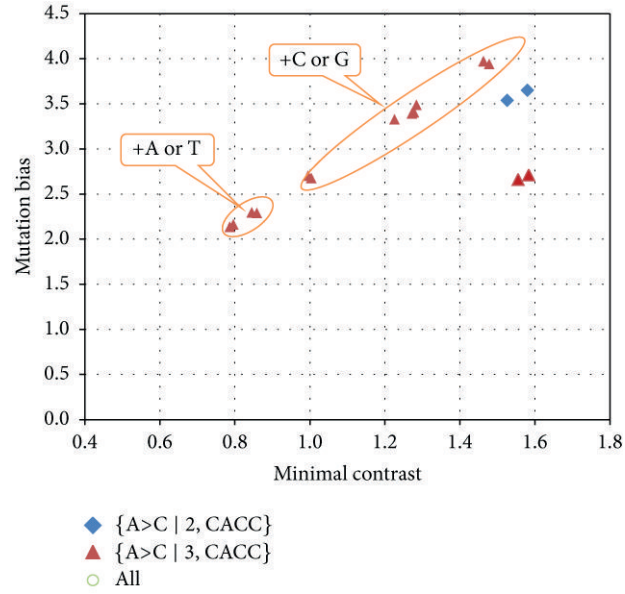


Figure 1: Mutation bias and minimal contrasts of mutation contexts in *D. melanogaster*. Each dot represents a mutation context. Triangles represent the {A>C | 3, CCA} (as well as complementary contexts) and contexts that had this context as a subcontext. Most dots are in pairs because complementary contexts have similar mutation bias and minimal contrast values.

Table 2: Comparison of single nucleotide mutations in *D. melanogaster* and *H. sapiens*. Transitions are italic, while transversions are bold.

D. melanogaster		*H. sapiens*	
Mutation	Fraction	Mutation	Fraction
A>C	**0,044**	**A>T**	**0,031**
T>G	**0,047**	**T>A**	**0,031**
C>G	**0,047**	**A>C**	**0,037**
G>C	**0,048**	**T>G**	**0,038**
A>T	**0,057**	**C>G**	**0,051**
T>A	**0,058**	**G>C**	**0,051**
A>G	*0,063*	**G>T**	**0,058**
T>C	*0,064*	**C>A**	**0,058**
G>T	**0,118**	*T>C*	*0,118*
C>A	**0,121**	*A>G*	*0,118*
C>T	*0,166*	*C>T*	*0,204*
G>A	*0,167*	*G>A*	*0,204*
Transversions	**0,540**	**Transversions**	**0,355**
Transitions	*0,460*	*Transitions*	*0,645*

D. melanogaster and *H. sapiens* mutageneses are also strikingly different for several 2–4 bp contexts, as shown in Figure 2. The {C>T | 1, CG}, {T>C | 2, ATTG},

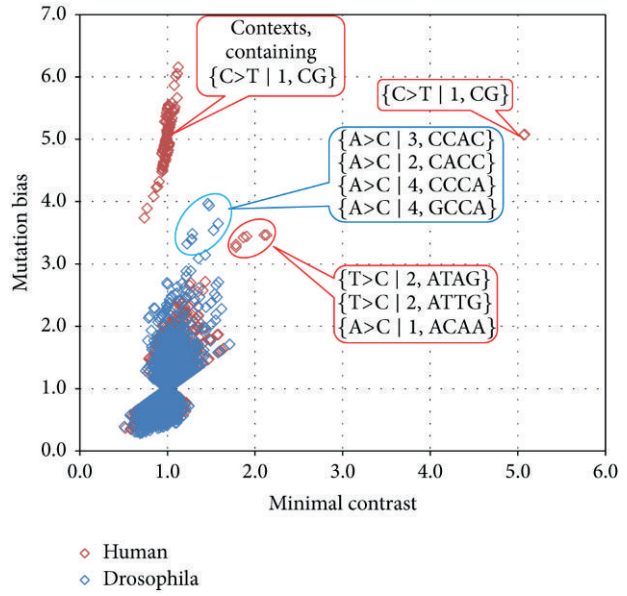


Figure 2: Mutation bias and minimal contrast for *D. melanogaster* and *H. sapiens*. Each dot represents a mutation context (blue in *D. melanogaster*, red in *H. sapiens*). Dots are overlapping and are usually in pairs because complementary contexts have similar mutation bias and minimal contrast values.

{T>C | 2, ATAG}, and {A>C | 1, ACAA} mutation contexts appear to have excessive mutation frequencies in *H. sapiens* but not in *D. melanogaster*. Interestingly, the CAATT sequence (contains the ATTG word on the reverse strand) appears to be a mutation hotspot for the human DNA polymerase eta [18]. Also, the CCAAT (contains the ATTG word on reverse strand) motif is a known target site for enhancer-binding proteins [19]. The increased number of ATTG>ACTG mutations might be partially due to selection against enhancers sequences in nontranscribed regions of the genome.

On the other hand, several mutation contexts seem to have increased mutation bias in *D. melanogaster*. The differences between different mutation contexts in *D. melanogaster* and *H. sapiens* are shown in more detail in Figure 3.

In a previous study, we compared the over- and underrepresentation of 1–7 bp nucleotide words in the genomes of 139 complete eukaryotic genomes, including *H. sapiens* and *D. melanogaster* [20]. Table 3 contains a part of this comparison for several words in *H. sapiens* and *D. melanogaster* related to the previously discussed mutation contexts. The word CG has a strong underrepresentation in *H. sapiens* (by 76.37% from the expected genomic frequency) while in *D. melanogaster* it is only slightly underrepresented (by 5.93% from the expected genomic frequency). The derived word TG is overrepresented by 20.1% and by 10.67% in *H. sapiens* and *D. melanogaster*, respectively. The {C>T | 1, CG} mutation context seems to be the only example of a mutation context that has remarkably affected the genomic word composition in *H. sapiens* compared to *D. melanogaster*. The absence of such effects for words related to other mutation contexts might be due to us not taking into account the rates of other mutations in these words or mutations that produce these words.

Table 3: Over- and underrepresentation of genomic frequencies for several words in *H. sapiens* and *D. melanogaster*. Data is taken from a previous study [20] supplementary table (available at http://mouse.genebee.msu.ru/words/Supple3(contrast_k).xls). The numbers represent the value C = [(Obs (W) − Exp (W))/Exp (W)] · 100%, where Obs (W) is the observed word frequency and Exp (W) is the expected word frequency (based on the frequencies of all of its subwords).

	Genomic word over- and underrepresentation in	
	H. sapiens	*D. melanogaster*
Words containing a mutation context with increased mutation bias in *H. Sapiens*		
CG	−76.37%	−5.93%
ATAG	−0.79%	4.38%
ATTG	−7.07%	−2.35%
ACAA	1.62%	3.75%
Words derived from mutation contexts with increased mutation bias in *H. Sapiens*		
TG	20.10%	10.67%
ACAG	1.51%	−4.94%
ACTG	−2.07%	−0.46%
CCAA	−6.17%	−1.61%
Words containing mutation contexts with increased mutation bias in *D. melanogaster*		
CCAC	0.19%	1.52%
CACC	1.18%	−4.24%
CCCA	5.63%	0.09%
GCCA	−2.77%	3.63%
ACC	2.28%	−2.39%
CCA	14.82%	9.90%
Words derived from mutation contexts with increased mutation bias in *D. melanogaster*		
CCCC	−5.10%	2.19%
GCCC	1.66%	−1.41%
CCC	−12.66%	−7.78%

4. Conclusions

The regularities of mutagenesis are different in *D. melanogaster* and *H. sapiens*. However, these differences may be attributed to a rather small number of mutation contexts that behave in a different manner in these two species. First, there is an increased frequency of G>T (C>A) transversions in *D. melanogaster*. Several possible molecular mechanisms for this have been proposed. Second, there is an increased frequency

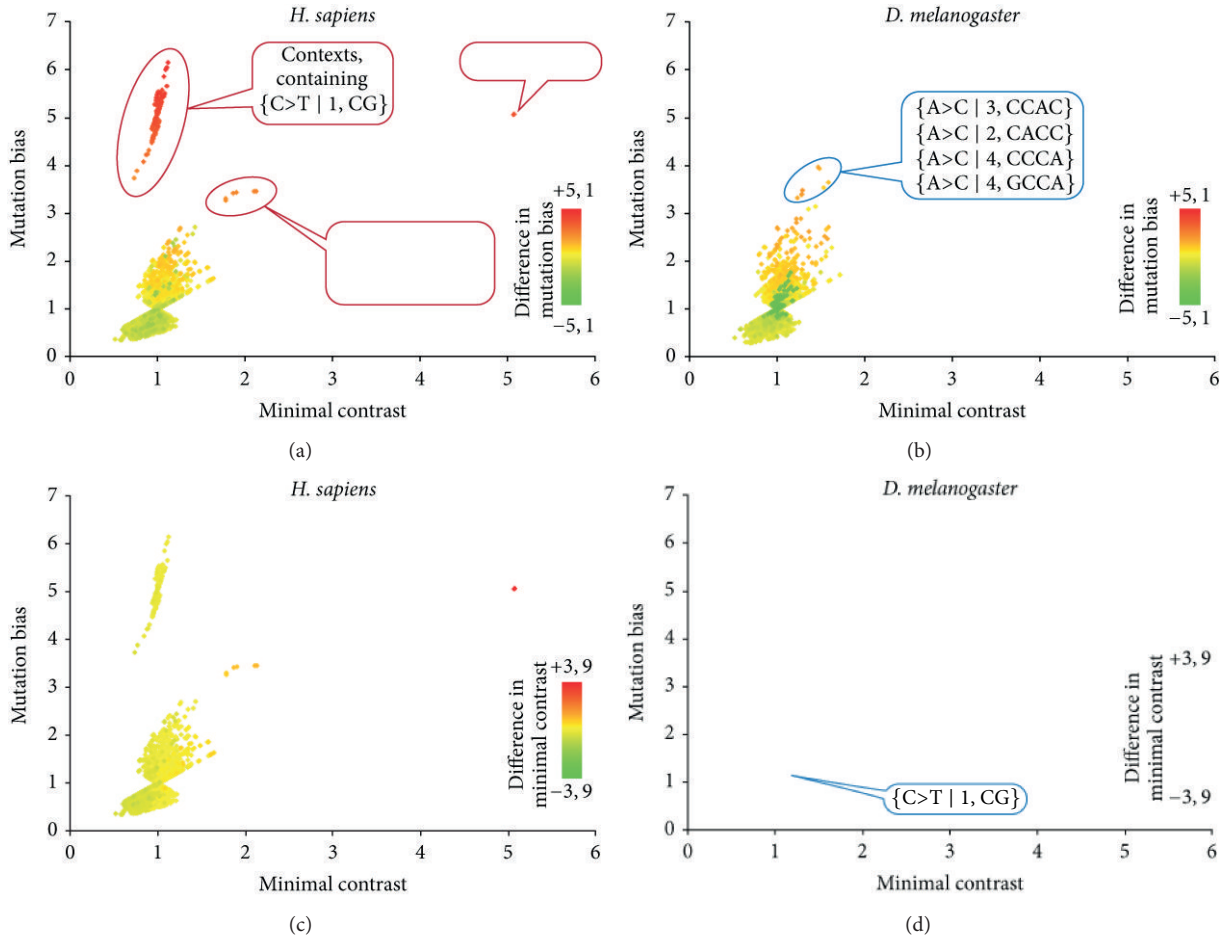


Figure 3: The difference between *H. sapiens* and *D. melanogaster* mutation bias ((a) and (b)) and minimal contrast ((c) and (d)) for 2–4 bp mutation contexts. Each dot represents a mutation context. The *X* axis represents the contexts minimal contrast values, and the *Y* axis represents the contexts mutation bias. The minimal contrast and mutation bias values are given for *H. sapiens* ((a) and (c)) and for *D. melanogaster* ((b) and (d)), and the color scheme indicates the difference between minimal contrasts. Thus, red dots on (a) and (c) represent contexts that are hypermutable in humans comparing to drosophila, while green dots represent contexts that are hypermutable in *D. melanogaster* comparing to *H. sapiens*. This scheme is reversed for (b) and (d).

of C>T mutation in the word CG in *H. sapiens*. This is probably explained by the fact that human germline methylation is abundant and CpG specific, while *D. melanogaster* is not. Third, there is an increased frequency of T>C mutations in the second position of the words ATTG and ATAG and an increased frequency of A>C mutations in the first position of the ACAA word in *H. sapiens* but not in *D. melanogaster*. And finally, there is an increased A>C mutations rate in {A>C | 2, CACC} and {A>C | 3, CCA} mutation contexts in *D. melanogaster* but not in *H. sapiens*.

Acknowledgments

This work was supported by Russian Ministry of Science and Education State Contracts 8494 and 8100 of the Federal Special Program "Scientific and Educational Human Resources of Innovative Russia" for 2009–2013 and the Russian Foundation for Basic Research Grants 12-04-91334, 11-04-91340, 13-07-00969, 12-04-31071, and 11-04-01511.

References

[1] C. F. Baer, M. M. Miyamoto, and D. R. Denver, "Mutation rate variation in multicellular eukaryotes: causes and consequences," *Nature Reviews Genetics*, vol. 8, no. 8, pp. 619–631, 2007.

[2] A. Kong, M. L. Frigge, G. Masson et al., "Rate of de novo mutations and the importance of father's age to disease risk," *Nature*, vol. 488, no. 7412, pp. 471–475, 2012.

[3] D. N. Cooper and M. Krawczak, "Cytosine methylation and the fate of CpG dinucleotides in vertebrates genomes," *Human Genetics*, vol. 83, no. 2, pp. 181–188, 1989.

[4] N. D. Singh, P. F. Arndt, A. G. Clark, and C. F. Aquadro, "Strong evidence for lineage and sequence specificity of substitution rates and patterns in drosophila," *Molecular Biology and Evolution*, vol. 26, no. 7, pp. 1591–1605, 2009.

[5] F. Lyko, B. H. Ramsahoye, and R. Jaenisch, "DNA methylation in Drosophila melanogaster," *Nature*, vol. 408, no. 6812, pp. 538–540, 2000.

[6] N. Arnheim and P. Calabrese, "Understanding what determines the frequency and pattern of human germline mutations," *Nature Reviews Genetics*, vol. 10, no. 7, pp. 478–488, 2009.

[7] A. Hodgkinson, E. Ladoukakis, and A. Eyre-Walker, "Cryptic variation in the human mutation rate," *PLoS Biology*, vol. 7, no. 2, Article ID e1000027, 2009.

[8] R. D. Blake, S. T. Hess, and J. Nicholson-Tuell, "The influence of nearest neighbors on the rate and pattern of spontaneous point mutations," *Journal of Molecular Evolution*, vol. 34, no. 3, pp. 189–200, 1992.

[9] D. G. Hwang and P. Green, "Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 39, pp. 13994–14001, 2004.

[10] A. Y. Panchin, S. I. Mitrofanov, A. V. Alexeevski, S. A. Spirin, and Y. V. Panchin, "New words in human mutagenesis," *BMC Bioinformatics*, vol. 12, article 268, 2011.

[11] R. M. Kuhn, D. Karolchik, A. S. Zweig et al., "The UCSC genome browser database: update 2009," *Nucleic Acids Research*, vol. 37, no. 1, pp. D755–D761, 2009.

[12] K. C. Cheng, D. S. Cahill, H. Kasai, S. Nishimura, and L. A. Loeb, "8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes G → T and A → C substitutions," *Journal of Biological Chemistry*, vol. 267, no. 1, pp. 166–172, 1992.

[13] D. Wang, D. A. Kreutzer, and J. M. Essigmann, "Mutagenicity and repair of oxidative DNA damage: insights from studies using defined lesions," *Mutation Research*, vol. 400, no. 1-2, pp. 99–115, 1998.

[14] T. Douki, D. Perdiz, P. Gróf et al., "Oxidation of guanine in cellular DNA by solar UV radiation: biological role," *Photochemistry and Photobiology*, vol. 70, no. 2, pp. 184–190, 1999.

[15] C. Dherin, M. Dizdaroglu, H. Doerflinger, S. Boiteux, and J. P. Radicella, "Repair of oxidative DNA damage in Drosophila melanogaster: identification and characterization of dOgg1, a second DNA glycosylase activity for 8-hydroxyguanine and formamidopyrimidines," *Nucleic Acids Research*, vol. 28, no. 23, pp. 4583–4592, 2000.

[16] P. L. Foster, E. Eisenstadt, and J. H. Miller, "Base substitution mutations induced by metabolically activated aflatoxin B1," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 80, no. 9, pp. 2695–2698, 1983.

[17] Y. Trottier, W. I. Waithe, and A. Anderson, "Kinds of mutations induced by aflatoxin B1 in a shuttle vector replicating in human cells transiently expressing cytochrome P450IA2 cDNA," *Molecular Carcinogenesis*, vol. 6, no. 2, pp. 140–147, 1992.

[18] T. Matsuda, K. Bebenek, C. Masutani, I. B. Rogozin, F. Hanaoka, and T. A. Kunkel, "Error rate and specificity of human and murine DNA polymerase η," *Journal of Molecular Biology*, vol. 312, no. 2, pp. 335–346, 2001.

[19] S. Osada, H. Yamamoto, T. Nishihara, and M. Imagawa, "DNA binding specificity of the CCAAT/enhancer-binding protein transcription factor family," *Journal of Biological Chemistry*, vol. 271, no. 7, pp. 3891–3896, 1996.

[20] S. I. Mitrofanov, A. Y. Panchin, S. A. Spirin, A. V. Alexeevski, and Y. V. Panchin, "Exclusive sequences of different genomes," *Journal of Bioinformatics and Computational Biology*, vol. 8, no. 3, pp. 519–534, 2010.

Hindawi Publishing Corporation
International Journal of Genomics
Volume 2013, Article ID 629543, 16 pages
http://dx.doi.org/10.1155/2013/629543

Research Article

Comparative Study of Genome Divergence in Salmonids with Various Rates of Genetic Isolation

Elena A. Shubina,[1,2] Mikhail A. Nikitin,[1] Ekaterina V. Ponomareva,[1] Denis V. Goryunov,[1] and Oleg F. Gritsenko[3]

[1] *Belozersky Institute for Physico-Chemical Biology of Lomonosov Moscow State University, Leninskie gory, 1, Moscow 119991, Russia*
[2] *Biological Department of Lomonosov Moscow State University, Leninskie Gory 1, Moscow 119991, Russia*
[3] *Russian Federative Research Institute of Fisheries and Oceanology, 17A V. Krasnoselskaya Street, Moscow 107140, Russia*

Correspondence should be addressed to Elena A. Shubina; shubina@genebee.msu.su

Received 15 October 2012; Revised 22 March 2013; Accepted 15 May 2013

Academic Editor: Dmitry Sherbakov

The aim of the study is a comparative investigation of changes that certain genome parts undergo during speciation. The research was focused on divergence of coding and noncoding sequences in different groups of salmonid fishes of the Salmonidae (*Salmo, Parasalmo, Oncorhynchus*, and *Salvelinus* genera) and the Coregonidae families under different levels of reproductive isolation. Two basic approaches were used: (1) PCR-RAPD with a 20–22 nt primer design with subsequent cloning and sequencing of the products and (2) a modified endonuclease restriction analysis. The restriction fragments were shown with sequencing to represent satellite DNA. Effects of speciation are found in repetitive sequences. The revelation of expressed sequences in the majority of the employed anonymous loci allows for assuming the adaptive selection during allopatric speciation in isolated char forms.

1. Introduction

In view of the biological concept of Mayr [1] the process of speciation in the organisms with the sexual reproduction involves accumulation of differences sufficient to set the barrier of partial or complete incompatibility. According to Dobzhansky [2] it implies for the process of unlimited genetic recombinations within the species and the lack of gene flow between the species. Meanwhile, as repeatedly noted by many researches, for example, Mallet, Garside and Christie, Svardson, Wolf et al., Gross et al., and Scribner et al. ([3–7], review [8]), hybridization between species is known to occur both in the wild and under artificial conditions, and the hybrid forms exist along with the parental species. The fate of such interspecific hybrids sporadically occurring in the wild and their contribution in the genetic structure of populations are still under question, as Coyne and Orr and Hudson et al. [9, 10] showed.

Repetitive DNA sequences are convenient for the studies of the genome evolution [11–13]. According to Ohno [14], this fraction originates in the process of gene duplications and has a potential for large-scale rearrangements, because they are not subjected to the pressing of the natural selection.

From the directly obtained experimental data, phylogenetic reconstructions for the lower taxa on the basis of the repetitive DNA sequences yield better results than the other nuclear sequences for both animals and plants as Chase et al., Thompson et al., and Warburton and Willard [15–17] wrote. As mentioned by Ohta, [18], the factors of intragenomic homogenization counteract intragenomic differentiation of the fraction of repeats. These sequences become peculiar specific markers. The process of concerted evolution has been previously shown by Zimmer et al., Jeffreys et al., Gray et al., and Elder and Turner [19–22] to involve highly repetitive DNA sequences (satellite and satellite-like).

A well solution method of comparative studies of genomic eukaryotic DNA according to distribution of the sites of digestion by restriction endonucleases was proposed by Fedorov et al. [23]. The method, called taxonomic fingerprinting or taxonoprint, is a modification of the approach initially developed for the analysis of the mitochondrial DNA [24]. Investigation of about 50 animal species of various taxa

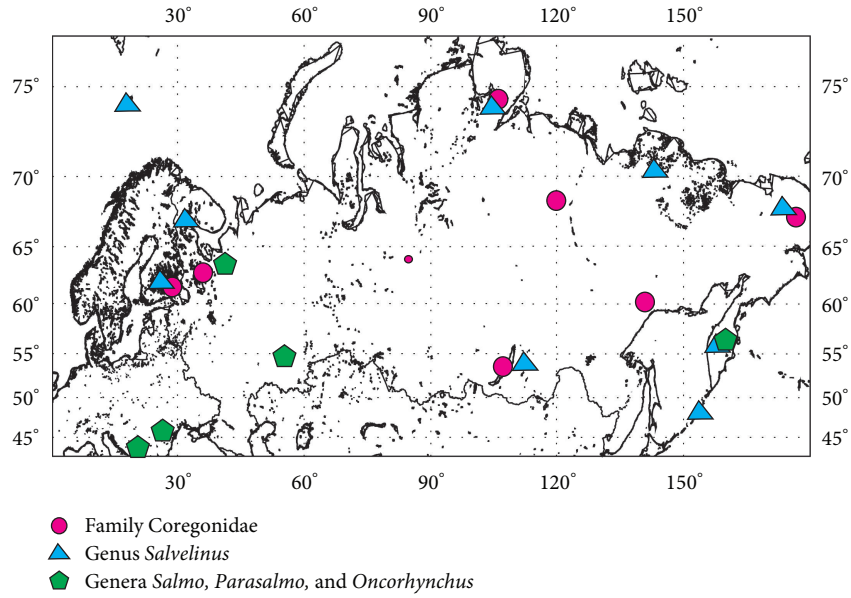


Figure 1: Sampling localities in Russia, Armenia, Kazakhstan, Finland, and Norway.

[25] has shown species specificity of the band patterns along with the absence of individual, sexual, and interpopulation polymorphism. Dominating contribution of high-copy, relatively long tandem repetitive sequences in "taxonoprints" was revealed by Roudykh et al. [26]. This method seems to us to be sufficiently appropriate for studying the molecular aspects of speciation.

We have shown previously that these general principles of the concert evolution were fully pronounced in the evolution of repetitive sequences of the salmonids of *Salmo*, *Parasalmo,* and *Oncorhynchus* genera—Mednikov et al. [27]. "Homing" common to salmonids resulted in reliable reproductive isolation with the subsequent divergence of the populations in accordance with the morphological and molecular characters. Situation with the repetitive DNA sequences in the organisms with the less strict genetic isolation seemed to be worth being analyzed. Whitefishes of Coregonidae family are one of the largest groups with interspecific hybridization; according to many experts, application of Mayr's biological species concept [1] to these fishes is limited (e.g., see discussion of the problem [28–30]).

The study was aimed at the comparative investigation of alteration of some genome fractions under differentiation of salmonid species and forms belonging to various groups and demonstrating various rates of reproductive isolation. Whitefishes of Coregonidae family, true salmons of the *Salmo*, *Parasalmo* (*Oncorhynchus*, after Smith and Stearley [31]) and *Oncorhynchus* genera, and a number of forms and species of *Salvelinus* genus were studied. The rates of isolation in whitefish and true salmon have polar characteristics due to extensive hybridization in some species and strict homing in the other. Geographic isolates and insular populations of anadromous chars occupy intermediate position.

Two methods of multilocus DNA analysis were applied in the phylogenetic and taxonomic studies of the Coregonidae and Salmonidae families. The methods were based both on the comparison of the extensive repetitive sequence [26] and collation of the anonymous PCR products (RAPD amplification in modification of Welsh and McClelland [32] and Williams [33]) and their subsequent sequencing.

2. Materials and Methods

Most of the salmon tissue samples were collected by the specialists of the Department of Ichthyology, Moscow State University, in 1984–2004 and Russian Federal Research Institute of Fisheries and Oceanography in 2000–2004. Arctic chars from the lakes of Finland were collected by M. Kaukkoranta. Whitefish from Lake Como (Canada) were kindly provided by Yu. S. Reshetnikov. Collecting sites are mapped on Figure 1.

In the early experiments, DNA samples of salmons and whitefish were extracted from gonads at III-IV stages of maturity preserved in alcohol with the method of phenol-chloroform extraction [34]. In some cases, additional purification with CsCl in the presence of ethidium bromide [35] and additional precipitation with cetyltrimethylammonium bromide [36] were performed. The length of DNA was checked by electrophoresis in 0.6% agarose gel. Concentrations of the obtained DNA preparations were estimated on SP-800 spectrophotometer (UK) and adjusted.

At a later stage, either Silica method [37] or traditional procedure with the use of proteinase K and organic solvents of Sambrook et al. [38] was applied. Prior to amplification, all samples were additionally purified with PEG-8000. DNA pellets were dissolved in TE buffer and stored frozen.

For restriction analysis DNA aliquots were digested by *Msp*I, *Taq*I, *Csp*6I (*Rsa*I), *Tru*9I (*Mse*I), *Hin*6I (*Hha*I), and *Mbo*I tetranucleotide restriction endonucleases and two isoschizomers sensitive to the presence of methylated bases

Table 1: The RAPD-PCR primers sequences.

No.	Designation	Sequence 5′-3′	Length [nucleotides]
1	I	CGT TGG AAG ACA GAC CTC CG	20
2	II	ATT CCC TGT CAA AGT AGG GT	20
3	III	GAG CAC TTT CTT GCC ATG AG	20
4	IV	GAA GCT GCT ATG CTT CGT AT	20
5	VI	CAT AAA TTG CTT TAA GGC G	19
6	VII	TCA TCT TCT TCC TCT TCT TC	20
7	1	TGT GAC TGC TTG TAG ATG GC	20
8	2	TGG AGC TGT GTA AGA AGT AC	20
9	3	AAA AGA CAT GAA GAC TCA GG	20
10	5	TGG ACA GTA CGG TGA ATG C	19
11	6	CCA CAA ACC AAT ATC TCT C	19
12	7	CTC AGA GTC CAA CCT GGG TAG	21

(*Bsp143*I and *Sau3A*I), as well as *Cfr13*I and *Bcn*I degenerate pentanucleotide restriction endonucleases (*Fermentas*, Lithuania; *Sibenzym*, Russia). Reactions were performed overnight under conditions corresponding to the manufacturer's recommendations. The fragments of hydrolysis were labeled at cohesive 3′ ends using Klenov's fragment of DNA-polymerase I *E. coli* and using [a-^{32}P] dNTPs (Institute for Physics and Power Engendering, Obninsk, Russia). Prior to electrophoresis, minor labeled oligonucleotides were removed from hydrolysate by gel filtration through Sefadex G-50 (medium) (Sigma, USA) during centrifugation according to Maniatis et al. [39]. This procedure improved resolution of radioautographs. Electrophoresis in nondenatured 10% polyacrylamide gel (20 × 40 cm) was performed manually according to the method described in Fedorov et al. [23]. pBR322 DNA-*Msp*I digest was used as a marker of molecular weight.

For RAPD-PCR experiments 19- or 20-mer oligonucleotides in various combinations were used as the primers (Tables 1 and 2).

Amplification reaction with two arbitrary primers was performed in 25 μL of 0.01 M tris-HCl PH 8.3 buffer containing 0.05 M NaCl, the mixture of four dNTP (0.2 mM each), $MgCl_2$ (5 mM), two arbitrary primers in various combinations (5 μm each), DNA-polymerase *Thermus aquaticus* (2.0 units per sample) (*Dialat*, Russia), and appr. 20 ng of DNA. PCR conditions: 94°C, 2.0 min, (88°C—1 min, 92°—1 min) × 1; (94°C—45 sec, 50°C—30 sec, 72°C—30 sec) × 3; (94°C—45 sec, 60°C—30 sec, 72°C—30 sec) × 35, 72°C—10 min. No DNA was added to the check sample. Compounding ingredients were prepared and mixed in accordance with published recommendations. Electrophoretic fractionation of PCR products was performed in 2% agarose gel (1.5% low melted, Sigma, +0.5% type II, Sigma), dyed with ethidium bromide, and photographed in UV light. Under the described conditions the reactions were stable and replicable. Photographs of gels containing DNA PCR-RAPD electrophoretic patterns were taken in UV light in Kodak EDAS 290 Electrophoresis Documentation and Analysis System. Adobe Photoshop 7.0 and Gel-Quant (Free Trial) software was used for further editing. Subsequent compilation of gels and construction of the binary matrix of the characters was performed manually. The matrix was compiled according to the "presence or absence of the fragment" principle; only stable major bands were considered.

Table 2: Combinations of oligonucleotides, used in pairwise RAPD-PCR.

Pair no.	Combination
1slv	1 + I*
4sm	1 + IV
9slv	2 + II
11slv	2 + IV*
12slv	2 + V
14sm	2 + VII
15slv	3 + I*
16sm	3 + II
18sm	3 + IV
30sm	5 + II**
31sm	5 + III**
32sm	5 + IV**
32′sm	5
35slv	5 + VII*
37sm	6 + II**
43sm	7 + I
44sm	7 + II
45sm	7 + III
56sm	2 + 3
58slv	2 + 5*
69sm	5 + 7*
71sm	I + II
76slv	I + VII*
77sm	II + III

(*) Pairs, used for the genus *Salvelinus* species and forms (**)—Pairs, used with all the rest fishes DNA.

Cluster analysis with constructing dendrograms was performed by the distant Neighbor Joining (NJ) approach of Saitou and Nei [40] and UPGMA. Pairwise genetic distances

between patterns were calculated according to the method of Nei and Li [41] and Nei [42]. Distant trees were constructed using the TREECON 1.3b program of Van De Peer and De Wachter [43]. Node stability was tested by bootstrap analysis according to Felsenstein [44], and not samples but characters were estimated in both cases.

DNA fragments were extracted following electrophoretic separation of the PCR products in the agarose gel in columns with GFX PCR: DNA and Gel Band Purification kit (Amersham Biosciences Inc., USA) according to the manufacturer's recommendation. For sequencing of heterogeneous matrices and fragments amplifying from a single primer, corresponding products were cloned in *E. coli* with InsTAclone PCR Cloning Kit (Fermentas, Lithuania) following their extraction and freeing from agarose. The kit included the so-called T-vector, that is, pTZ57R/T plasmid with the extended "sticky" T-end. The colonies of white color containing inserts were grown after screening in liquid medium. In total 160 clones were examined. The difference between lengths ranged from 30 to 50 bp in different cases.

DNA sequencing was performed using ABI PRISM BigDye Terminator v. 3.1 kit with subsequent analysis of the reaction products on ABI PRISM 3100-Avant Genetic Analyzer (Life Technologies/Applied Biosystems, USA). CHROMAS, DNA, and BioEdit programs were used for interpretation of chromatograms. Homologous sequences were searched for in GenBank with the use of dbEST bases from NCBI resources; BLAST software was used for the search.

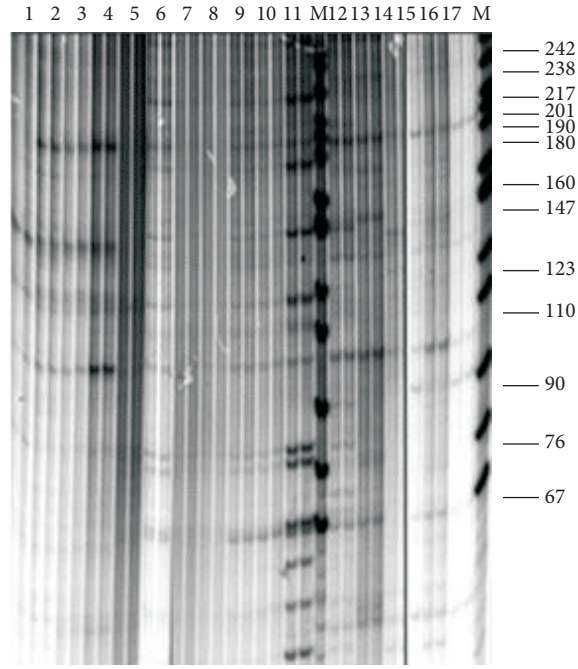


Figure 2: Electrophoretic separation of DNA *Taq*I digest from (1) *Salmo salar*; (2) *S. trutta*; (3) *S. trutta caspius*; (4) S. *ischchan*; (5) *Parasalmo mykiss* (steelhead, anadromous form); (6) *P. mykiss* (with traits of *P. clarkii*); (7) *P. mykiss* (freshwater steelhead); (8) *P. mykiss* (introduced rainbow trout); (9) *P. mykiss* (steelhead, North America); (10) *P. clarkii*; (11). *Oncorhynchus masou*; (12) *O. keta*; (13) *O. gorbuscha*; (14) *O. nerka*; (15) *O. kisutch*; (16) *O. tshawytscha*. M-marker: *Msp*I-digested pBR322 DNA.

3. Results and Discussion

3.1. Restriction Endonuclease Analysis. Restriction analysis of salmonid repetitive DNA has revealed some peculiarities of electrophoretic patterns in whitefishes. The first of them refers to the total number of bands of low intensity, which is extremely high. The matrices are not presented in the paper because of their large size and low information value. We have also found this phenomenon in the other salmonids (Atlantic and Far-Eastern salmons, trout, and chars), which is probably associated with their polyploidy origin. We used sperling (*Osmerus* sp.) as an outer group; in our opinion, notably lower degree of DNA banding patterns in this species could testify in behalf of this assumption.

The second characteristic trait is the high degree of similarity of the repetitive DNA in the studied fishes, which not only are "good" morphological species but also belong to different genera. On the basis of this characteristic whitefishes differed from all previously studied organisms [25], the fishes of Salmonidae family (genera *Salmo*, *Parasalmo*, *Oncorhynchus*, and *Salvelinus*—Figures 2 and 3) among them; see also [27] and [45]. An example of whitefishes' restriction pattern is shown on Figure 4.

The results obtained indicate that each whitefish species is characterized, under chosen experimental conditions, by a number of major and minor bands ranging between 600 and 40 bp, but only some of them are species specific. For example, major band of 230 bp appearing after digestion by *Taq*I restriction endonuclease is present in broad whitefish *Coregonus nasus*, the least cisco *C. sardinella,* and three North American species: Arctic cisco *C. autumnalis*, broad whitefish *C. nasus*, and lake cisco *C. artedi*. This band is notably less pronounced in *Msp*I cutting *C. lavaretus* from Lake Anetti, Finland (predator, high-gillraker form) and European cisco *C. albula*. The rest of species under study lack this band. The band of 150 bp is common to all whitefish species except for round whitefish *Prosopium cylindraceum*. Specific bands correspond to 300 bp in Canadian Arctic cisco and 80 and 60 bp in Finnish high-gillraker whitefish. Distinct species-specific band of 170 bp appears in round whitefish after digestion by *Msp*I; the rest major bands of 460 bp, 310 bp, and 90 bp and smaller are common for all studied DNA. The sites for *Sau3A*I and *Tru9*I restriction endonuclease are not polymorphic at all. *Cfr13*I (*Asu*I) reveals species-specific bands 180 bp and 190 bp in round whitefish; the rest major bands (350 bp, double 300 bp, 240 bp, 150 bp, 135 bp, double 100–110 bp and smaller) are identical in all studied DNA.

Thus, taxonoprint analysis revealed surprising homogeneity of the fraction of high-copy DNA repeats in different species and even genera of whitefish (with round whitefish being the single exception mentioned above). Most of the major bands in the patterns of all used restriction endonucleases were absolutely identical. Slightly pronounced polymorphism was found only in the families of sequences with small number of copies. Reliability of the relative positions

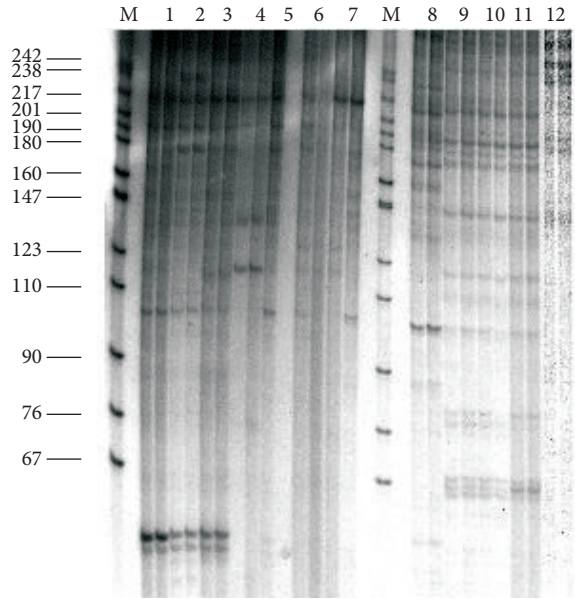


Figure 3: Electrophoretic separation of DNA *Taq*I digests from: (1) *Salvelinus alpinus*; (2) *S. drjagini*; (3) *S. malma*; (4) *Salvethymus svetovidovi*; (5) *Salvelinus elgyticus*; (6) *S. boganidae*; (7) *S. confluentus*; (8) *S. leucomaenis*; (9) *Parasalmo clarkii*; (10) *P. mykiss*; (11) *Salmo irideus*; (12) *Oncorhynchus masou*. M—marker: *Msp*I-digested pBR322 DNA.

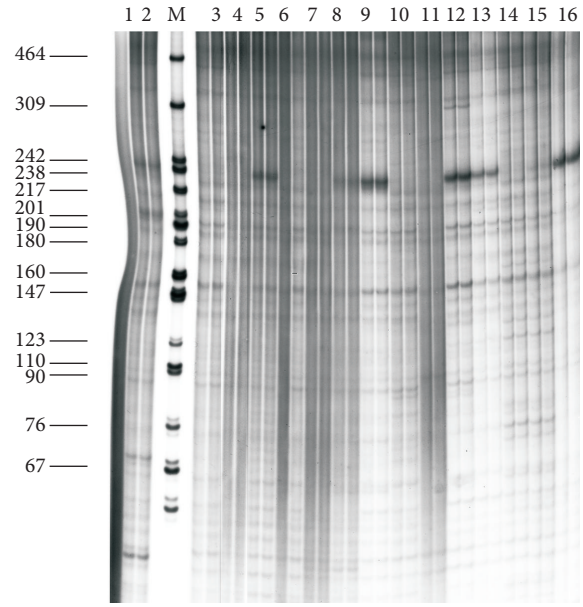


Figure 4: Electrophoretic separation of DNA *Taq*1 digests from (1) *Osmerus* sp. (outgroup); (2) *Coregonus lavaretus* lacustrine form; (3) *C. lavaretus* anadromous form; (4) *C. autumnalis migratorius*; (5) *C. nasus*; (6) *C. lavaretus montchegor*; (7) C. *muksun*; (8) *C. lavaretus pidschiani*; (9) *C. albula*; (10) *C. sardinella*; (11) *Stenodus leucichthys nelma*; (12) *Prosopium cylindraceum*; (13) *C. autumnalis* (Como Lake); (14) *C. clupeaformis* (Como Lake); (15) *C. nasus* (Como Lake); (16) *C. artedi* (Huron Lake).

of the branches on computer-generated phylogenetic trees appeared to be low.

Taxonoprint analysis of *Salmo*, *Parasalmo,* and *Oncorhynchus* genera specimens (Figure 2) showed distinct division of the samples under study into four groups of generic rank: (1) Atlantic salmon *Salmo salar*; (2) Brown trout *Salmo trutta*, Caspian trout *Salmo trutta caspius*, and closely related trout *Salmo ischan* from Lake Sevan; (3) Kamchatka rainbow trout *Parasalmo mykiss*, its American freshwater form *P. gairdneri*, and cutthroat trout *P. clarkii*; (4) other species of the Pacific salmons *Oncorhynchus* genus. The bands of 242 and 240 bp, 175 and 140, 120, 110, and a number of bands lower than 70 bp revealed by the use of *Taq*I (Figure 2) are genus specific for all *Salmo sensu stricto* species. The bands of 240 and 150, 120 and 76, and short bands of 10, 20, and 30 bp absent in Atlantic *Salmo* and in *Oncorhynchys sensu stricto* [27] were present in the American and Kamchatka trout. The number of species-specific bands appeared to be very small, with coho salmon *Oncorhynchus kisutch* being the sole exception. However, the DNA pattern of masu *Oncorhynchus masou* has nothing in common with the patterns of Pacific trout and Atlantic salmons, except for the family-specific bands of 67 bp, 110 bp, and 510 bp (Figure 2); the rest *Oncorhynchus* species differ from them to even greater degree.

NJ dendrogram constructed on the basis of taxonoprints of repetitive DNA of salmonid fishes of *Salmo*, *Parasalmo,* and *Oncorhynchus* genera is shown in Figure 5. All six species of the *Oncorhynchus* genus in its classical interpretation form, with high reliability, a cluster isolated from the American and Kamchatka trouts.

Distance analyses of endonuclease restriction data and trees reconstruction were done with UPGMA [43]. Figure 5 depicts the genetic variability within salmon (a), chars (b), and whitefishes (c).

Dendrogram (a) of the Salmonidae family represents a robust phylogeny of species with perfect reproductive isolation. Although position of some branches is controversial, the divergence of the Pacific trout *Parasalmo* and the salmon *Oncorhynchus* is firmly established. The diversification of major nodes is dated by 5–15 Mya [46, 47].

Within the genus *Salvelinus* (b), all debatable species of the *Salvelinus alpinus* – *Salvelinus malma* complex (including the North American Dolly Varden *S. confluentus*) are separated by small genetic distances. Their phylogenetic relationships cannot be established, as evident from low bootstrap support values. The compact *Salvelinus alpinus* – *Salvelinus malma* complex separates from *Salvethymus svetovidovi* and the "good" species *S. leucomaenis*, although the genus rank of *Salvethymus* [45, 48] cannot be confirmed with these data. The genus *Salvelinus* diverged at least 10 Mya [49], although the *Salvelinus alpinus* – *Salvelinus malma* complex is of glacial or postglacial origin, that is, of age less than 1 My.

On the whitefish phylogeny (c) the unresolved node contains all European and Asian species and forms of *Coregonus*. The North American representatives of the genus form a robust monophyletic clade. However this group is also compact, and the distance between the two contained taxa is less than 0.2 scale units. The basal branch leads

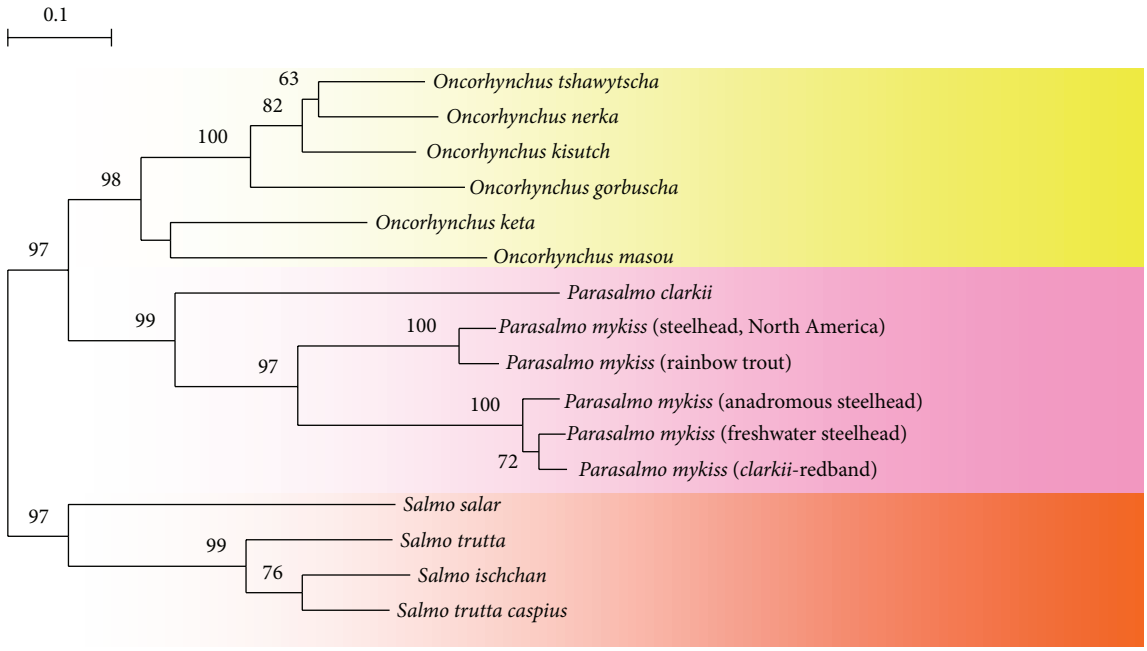


Figure 5: Salmon fishes NJ unrooted tree from restriction analysis data.

to the second coregonid genus represented by *Prosopium cylindraceum*. Recent whitefish diverged in the Pliocene 2–5 Mya or, perhaps, even earlier [50]. The genus *Prosopium* diverged in the early Miocene [51] dating the origin of round fishes to 20 Mya. This is the only reproductively isolated species of the Coregonidae highly supported as a distinct lineage by repetitive DNA restriction data.

Genetic variability of highly repetitive DNA in the three groups (Figures 6(a)–6(c)) is in good agreement with dating of the clades and the level of reproductive isolation: high within true salmons (clades diverged about 15 Mya; species are strictly isolated), intermediate within chars (the genus itself diverged about 10 Mya, the *Salvelinus alpinus – Salvelinus malma* complex—less than 1 Mya; forms are genetically close, albeit their hybridization is spatially prevented), and low within whitefishes (the main diversity established 2–5 Mya; hybridization is commonly observed across all Eurasian forms).

The presented evidence experimentally corroborates the hypothesis of concerted evolution of long satellite DNA.

3.2. RAPD-PCR. As could be expected, data on PCR-RAPD of whitefish were more diversified (Figure 7), though in this case bootstrap indices also appeared to be not high when the phylogenetic trees of Coregonidae were constructed (Figure 8).

Changes in algorithms of calculation of the genetic distances resulted in alteration of the tree topology. With all obtained dendrograms being considered, we can probably discuss only certain tendencies in relationships of whitefish. Round whitefish *P. cylindraceum* stands apart from all other species in the family Coregonidae and tends to occupy the position of the outer group. It shows up in all trees constructed by the UPGMA method when the outer group is chosen automatically as the most distant one. That is to say, remoteness of round whitefish *P. cylindraceum* from the other studied whitefishes is comparable to that of the sperling *Osmerus* sp. Inconnu (nelma) *Stenodus leucichthys* forms cluster with cisco *Coregonus albula*, whereas Baikal omul-arctic cisco (*C. autumnalis migratorius*) should be separated from the Arctic cisco *C. a. autumnalis* as an independent species. The latter is very close to the Bering cisco *C. laurettae* and has probably originated from the *Clupeaformis arthedi* group of numerous whitefish species inhabiting the Great Lakes. This assumption matches data on allozyme analysis [52]. The Baikal Arctic cisco *C. autumnalis* have been already proposed to be segregated as a separate species [53]. Our data suggest that *Leucichthys* is a composed diphyletic subgenus.

Experiments on RAPD analysis of *Salvelinus* chars were aimed at the assessment of the genetic diversity of the wild populations of *Salvelinus malma* inhabiting Paramushir and Onekotan Islands. Electrophoretic pattern is shown in Figure 9.

Population genetics of *Salvelinus* is of particular interest because of extensive processes of form and species generation occurring now in the genus [54]. Nominal species of the genus are polymorphic and represented by anadromous and freshwater forms as well as the geographic isolates characterized by particular morphological traits. Despite great commercial interest in salmonid fishes, the rate of genetic isolation of various char forms in the wild has been poorly studied so far. DNA of five available "good" morphological species (*S. alpinus, S. leucomaenis, S. fontinalis, S. namaycush, and S. malma*) was used for selection of the primers that allow discrimination between the Salvelinus species. The layout of electrophoretic patterns of the reactions with the pairs of primers NN 1, 9, 11, 12, 15, 30, 31, 32, 35, 37, 58, 69, and 76 (Tables 1 and 2) was used. The listed markers

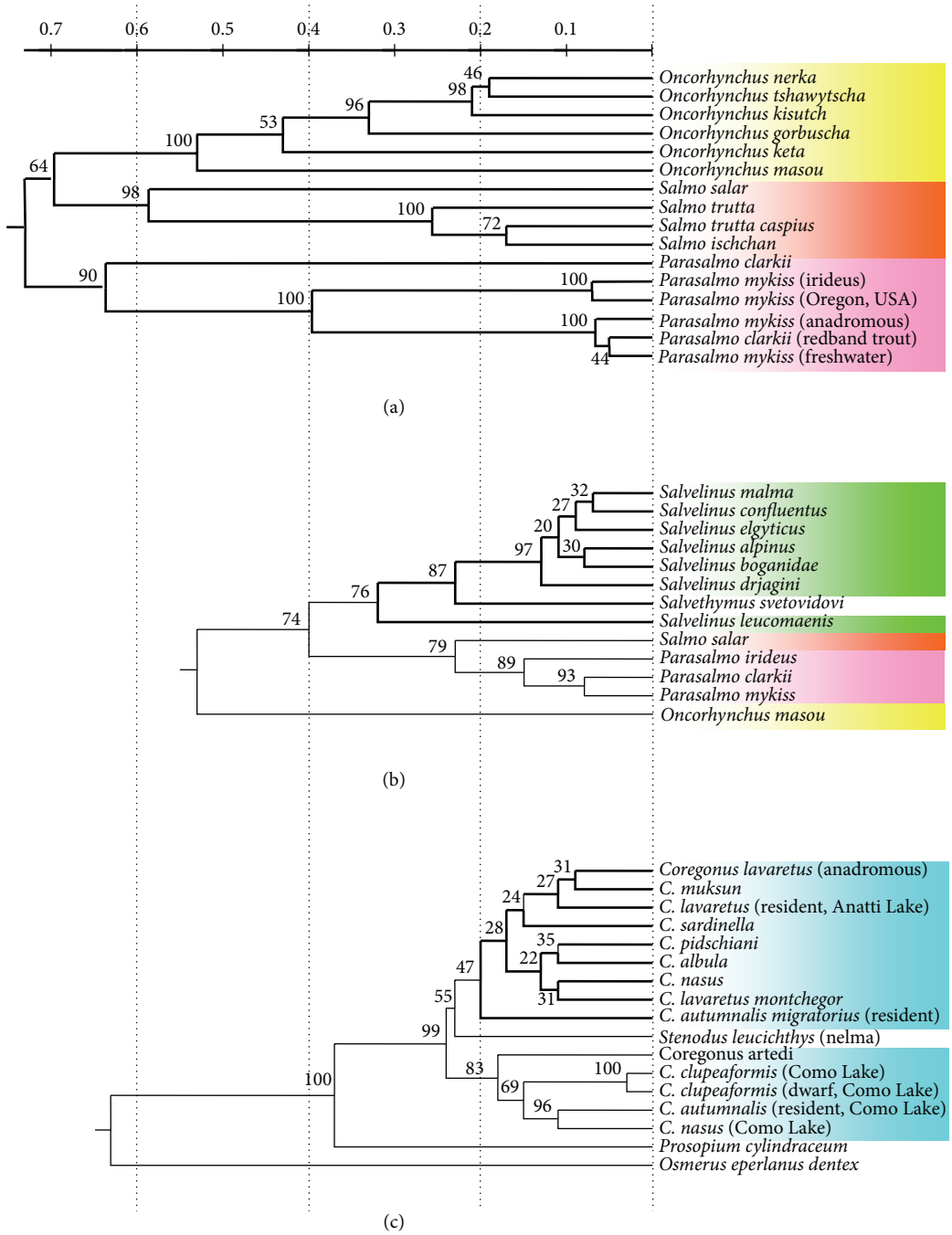


Figure 6: The extent of genetic variability within salmon observed from restriction endonuclease digestion data by UPGMA. (a) Salmonidae, (b) *Salvelinus*, and (c) Coregonidae. For (a) and (c) the proportion of acrylamide/bisacrylamide in PAGE was 29 : 1; for (b) this proportion was 19 : 1, which resulted in slightly lower numbers of detected bands and lower absolute genetic distances.

had sufficient polymorphism and the difference between the species is evident from the pattern of the major bands as well. In the course of binary matrix compilation, all bands in the electrophoretic pattern of PCR products were taken into consideration.

A total of 170 DNA samples representing 23 forms and species of chars were analyzed for the tree construction. Many analyzed isolated forms have been described [55] as species on the basis of various phenotypic characteristics, though some experts [54, 56] consider them to represent a single complex species *S. alpinus* complex with circumpolar distribution and low rate of genetic distinction [57, 58]. On this particular stage of the study we attempted to use genetic distances in a restricted system of the markers as a measure of taxonomic status of the certain forms. *Salmo salar* was used as an outer group. We took advantage of the matrix of pairwise distances, generated by TREECON for NJ tree construction to comprise the absolute values of genetic distances between

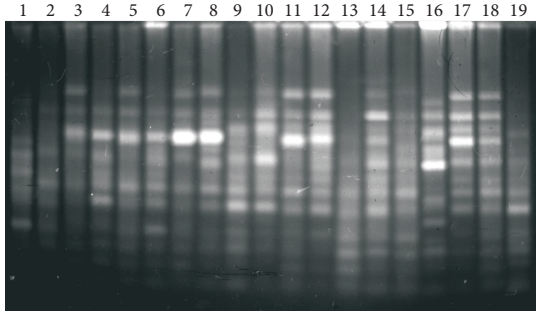


Figure 7: PCR-RAPD electrophoretic pattern of coregonid fishes DNA with pairwise combination of primers no. II + IV (Table 2). (1) *Osmerus* sp.; (2) *Coregonus lavaretus* (resident high-gillraker form, Anatty Lake); (3) *C. lavaretus* (anadromous form); (4) *C. lavaretus* (resident form); (5) *Coregonus lavaretus montchegor*; (6) *C. lavaretus pidschiani*; (7) *C. muksun*; (8) *C. autumnalis migratorius* (Baykal Lake, resident form); (9) *C. autumnalis autumnalis*; (10) *C. autumnalis* (Como Lake, resident form); (11) *C. nasus*; (12) *C. nasus* (Como Lake); (13) *C. albula*; (14) *C. sardinella*; (15) *Stenodus leucichthys*; (16) *Prosopium cylindracium*; (17) *C. clupeaformis* (Como Lake); (18) *C. clupeaformis* (dwarf); (19) *C. arthedi.*

commonly acknowledged "good" as well as disputable species and isolates (Table 3). On the tree as well as in the table the disputable species are designate by quotation marks.

Analysis of the genetic distance dendrogram (Figure 10) allows us to draw the following conclusions. The tree consists of three clusters. The first of them brings together all forms and species of *Salvelinus alpinus* complex under study. The stability of this node is 74% of bootstrap. The entire assembly of Dolly Varden (*S. malma*) is included into this cluster as its component. It is heterogeneous in the used system of the markers. Anadromous Dolly Varden from Bering Island appeared to be proximal to the long-finned char *Salvethymus svetovidovi* from Lake Elgygytgyn, though this grouping is uncertain. On the whole, the entire node combining the malma trout and the forms of the Arctic char seems unsolved. Kamchatka predatory char (*S. malma*) forms complex with the other Kamchatka and Kuril Dolly Varden (40% bootstrap support). UPGMA method even more robustly adds predatory char to the cluster containing Dolly Varden (tree is not presented). The same common cluster, along with the *S. malma*, contains also long-finned char "*Salvethymus*," described as a specimen of a separate genus [46].

Since at this stage of the work resolving power adequate to the species or close to it taxonomic level was used as a criterion for selection of the markers, 46% support for the inner cluster suggests the Dolly Varden being, more likely, of intraspecific status with respect to *S. alpinus* complex, as Berg supposed [59]. The second supported cluster (67% of bootstrap) is formed by white-spotted char *S. leucomaenis*. Reliability of the specific status is undisputed in this case. Although genetic distance between the South Kuril and Kamchatka white-spotted chars is relatively high, they form a monophyletic tree cluster corresponding to the single species *S. leucomaenis*. At least, two American species, *Salvelinus fontinalis* and *Salvelinus namaycush,* form the third branch supported by 68%. Alteration of the outer group (*Salmo salar* changed for *Osmerus*) or application of the alternative methods to construction of the trees (UPGMA or parsimony) did not change the topology in the basal part. In fact, two North American species represented an additional outer group for all other members of genus *Salvelinus*.

Within three discriminated clusters, the branch points are in most cases uncertain and could hardly be used for determination of the relationships of the separate populations. The dendrogram shows that anadromous Arctic char *Salvelinus alpinus*, a type species of the genus *Salvelinus*, forms a compact group that includes both anadromous and freshwater forms from Inary and Sayma Lakes (Finland), anadromous char from Svalbard, and Drjagin's char from Taimyr. According to Table 3, intrasample distances in this group range between 11.4 and 23.5 Nei's distance units ($\times 10^{-2}$) and averaged 18.03 Nei's units. In parsimonial construction the cluster shows maximal number of synapomorphies (not shown).

This group was used as a reference point for the estimation of the average genetic distances between chars of different taxonomic status. The distances between the samples of Dolly Varden (24.8×10^{-2} Nei's units on average) were estimated likewise. Subsequent analysis of the correspondence of the genotypic data to the position of particular forms in the system of genus *Salvelinus* was performed with the use of the table of pairwise distances. Data presented in Table 3 allows comparison of the absolute genetic distances between various forms of chars with the disputable taxonomic status and the "good" species: arctic char *Salvelinus alpinus*, white-spotted char *S. leucomaenis*, brook trout *S. fontinalis*, and lake trout *S. namaycush*. According to the table, the distances between *S. alpinus* and *S. leucomaenis*, and *S. fontinalis* and *S. namaycush* exceed 40×10^{-2} Nei's units. The distance between all the rest chars and *S. alpinus* could be defined as corresponding to the intraspecific taxonomic level with the different rate of advance. The longest distances are characteristic of the isolated forms of chars inhabiting the river and lake basin of the Taimyr Peninsula and Transbaikalian endemic species [60]. Although the data of this table are by no means the absolute indicators of the taxonomic status of particular forms, they provide an idea about correspondence of the phenotypic and genotypic data.

Identification of the nucleotide sequences of the fragments generated in PCR in the used system of primers with subsequent search for the homologous sequences in GenBank showed that most of the fragments forming RAPD-PCR electrophoretic pattern had homologies in dbEST base of NCBI resources with the "Salmonidae" or "zebrafish" as a filters (Table 4). Only extensive homological sequences with low "expect" value (low probability of the random coincidence) were inserted into the table.

RAPD fragments for the sequencing were chosen from North Kuril Islands populations *Salvelinus malma* DNA. From 160 clones analyzed for 10 insert DNA the homologies were not found ("junk" DNA?) and for four inserts homologies were found in nucleotide collection (microsatellites). The great part of the salmon fishes EST resources where homologies for variable RAPD fragments have been found was developed by Koop et al. [61]. The detailed analysis of

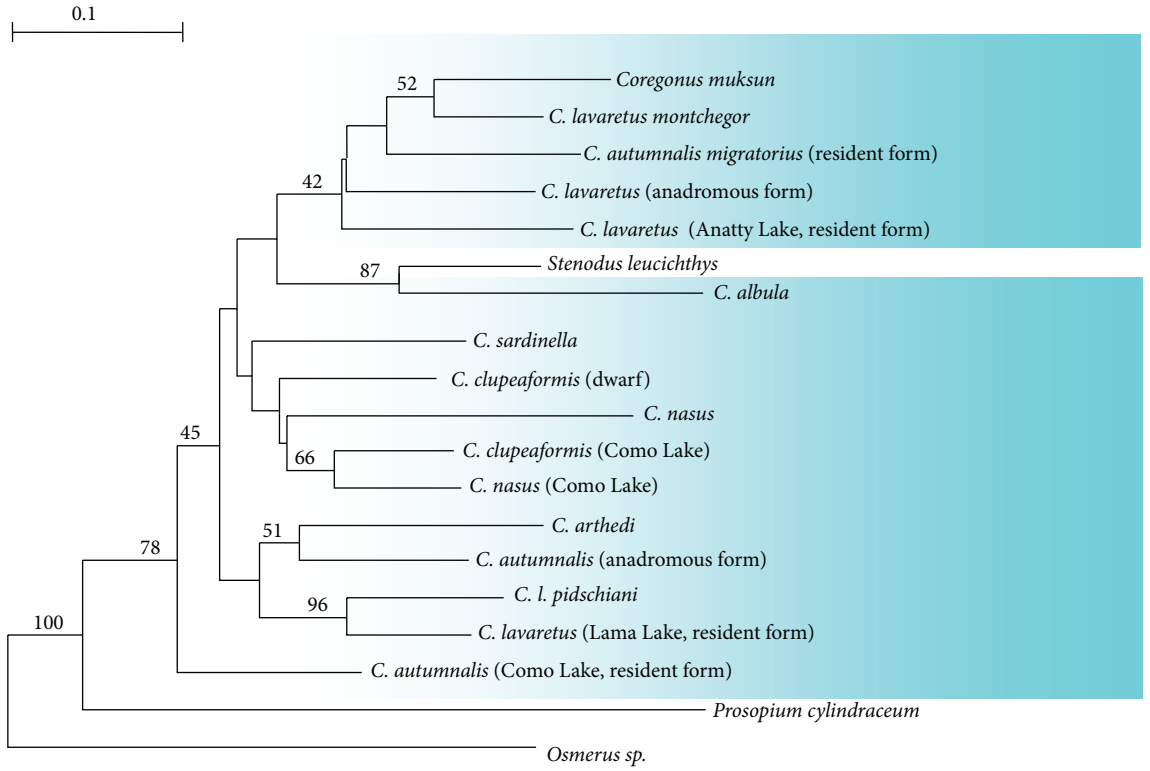


Figure 8: NJ tree of coregonid fishes from PCR-RAPD (more detailed description of species and forms are at the legend of Figure 6).

Table 3: The absolute values of genetic distances Nei between different forms of chars and the valid species.

Species and forms	*S. alpinus**	*S. leucomaenis**	*S. fontinalis*	*S. namaycush*
Genetic distances [Nei × 10^{-2}]				
*S. alpinus**	0.0	40.35	43.0	41.6
S. alpinus (Tuulispaa Lake)	25.8	41.6	48.1	42.1
S. alpinus Black char (Lama Lake)	34.6	44.4	53.3	50.0
"*S. boganidae*"	27.55	38.35	40.3	40.4
S. alpinus (Barents Sea basin)	27.1	44.55	50.0	46.1
S. alpinus Davatchan (Baikal Lake basin)	33.4	46.35	56.5	53.3
S. alpinus Mountain char (Lama Lake basin)	34.17	44.4	47.2	49.6
S. malma Predator (Kamchatka Peninsula)	27.85	40.1	41.9	36.4
S. alpinus Goggle-eyed (Lama Lake basin)	34.7	4.9	46.4	48.8
"*S. neiva*" Neiva (Sea of Okhotsk basin)	30.3	45.3	40.5	48.3
"*Salvethymus svetovidovi*" (Elgigitgin Lake, Chukotka Peninsula)	26.7	46.6	41.5	40.2
*S. malma** (Kuril Islands)	29.26	40.6	40.3	41.9

The matrix of distances was generated by TREECON for construction of NJ tree. For the chars that abbreviate with (*) the averaged pairwise distances between all samples analyzed are shown. The more detailed description of sample is shown at Figure 10.

these results is still not completed now, but major pattern appears to be followed. It was found that the majority markers used were either entire exons or contained exon fragments, abundantly exhibited in cDNA libraries, and portions of introns or intergenic DNA without detectable homologies in any library. The sequences with open frame encoded conservative protein domains in individual cases were found for the molecular markers with exon origin. For example, a significant and extensive homology with the hypothetical variable protein of *D. rerio* with the length of 659 amino acids (acc. no. XP 001332830.1), 52% identity, and very low probability of coincidence (2^{-26}) was found after the nucleotide matrix of 12.3 fragment had been converted into translated protein sequence by reading from frame 3.

Another fragment with the length of 905 bp was found to be the most homologous to *D. rerio* (zebrafish): clone DKEY-182H7 in linkage group 7 (acc. no. BX 663609.29) and mRNA of predicted protein, similar to norepinephrine (analogue of

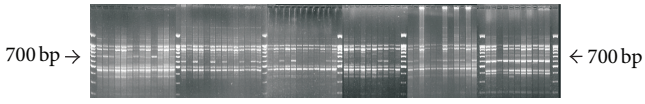


Figure 9: PCR-RAPD electrophoretic pattern of chars genus *Salvelinus* DNA with pairwise combination of primers II + 3. The samples are (1) Dolly Varden—*Salvelinus malma* (Kol' River, Kamchatka Peninsula); (2) Dolly Varden (Chernoe Lake, Onekotan Island); (3) Dolly Varden (Fontanka Stream, Onekotan Island); (4) Dolly Varden (Shelekhovka River, Paramushir Island); (5) Dolly Varden (Gol'tsovyi Stream, Onekotan Island); and (6) white-spotted charr—*S. leucomaenis* (South Kuril Islands). M—marker ladder 100 bp + 1.5 kb.

noradrenalin) carrier (acc. no. XM 689046.3 in GenBank). All the remaining homologies found for this sequence were to certain extent connected with noradrenalin protein carriers.

As was pointed out above, all PCR-RAPD reactions were repeated two times as minimum and reproduced absolutely. May be it is the result of the RAPD markers connections with the conservative protein coding genome areas in used primers system. We analyzed the monomorphous, generating the structure of electrophoretic patterns as well as polymorphic fragments. Essentially, it seems likely that introns or repetitive "junk" DNA with no homologies in dbEST or nucleotide databases are most informative for estimating genetic distances. For some of them the homologies in NCBI dbTSA were found.

Only in the single cases the homological sequences belonged to mtDNA with high evolution rate that should be able to correspond with the divergence in intraspecific taxonomical level. For example for reading from frame 4 sequence of 254 bp length (Table 4) the homology was found with Cytb of Atlantic salmon *Salmo trutta* (acc. no. ACO57211). However, it is impossible to exclude the implication of adaptive nuclear sequences in divergence of model groups studied.

Since the fish's groups selected differ considerably in basic stages of evolution and divergence time [30, 47, 51, 62, 63], integrated genomes assessment allows reducing the errors of disparity caused by difference in molecular evolution of different genes [64].

In the row containing almost all species and isolated forms of chars (disputable allopatric "species"), absolute genetic distances were used as a criterion—see Thorpe [65]. Correctness of this approach was confirmed by 100% homology of nucleotide sequences of RAPD fragments with identical electrophoretic mobility obtained in a single reaction.

According of our results the well-identified species, the lake trout and the brook trout, form a single cluster. No information on the level of phylogenetic relationships between *S. namaycush* and *S. fontinalis* is available to us. However, existence of hybrid forms traced up to F_4 by Berst et al. [66] and entering reproductive relations indicate close relations between these two North-American species. Molecular data by Westrich et al. [67] also are in favor of this assumption. Sister relations between *S. namaycush* and *S. fontinalis* oppose the proposal to consider *S. namaycush* as a specific genus "*Cristivomer*" [68]. This species should be placed in genus *Salvelinus*.

The second conclusion is connected with extremely low values of genetic distances between all samples of Dolly Varden and forms of the Arctic char. The degree of these distances is the same as that between the anadromous and isolated forms of the Arctic char, and even lower in some cases. If we rely on the distances only, we should refer all studied populations of Dolly Varden to the Arctic char. In other words, specific status of *Salvelinus malma* could be contested in case it is based only on the absolute values of genetic distances.

Restriction analysis of the Atlantic and Pacific salmons has revealed specific sets of DNA fragments in every species and even in the morphological forms of these fishes. Along with it, the results indicate pronounced isolation of the Pacific trout from the members of the genus *Oncorhynchus*. In addition, taxonoprints considerably and reliably differ in the species with overlapping ranges (sympatric species). It is true for Pacific salmons of the genus *Oncorhynchus* (chum salmon, pink salmon, sockeye salmon, chinook salmon, cherry salmon, and coho salmon), the chars of the genus *Salvelinus* (*S. alpinus* and *S. leucomaenis*), and two chars from Lake Elgygytgyn (long-finned char *Salvethymus svetovidovi* and Arctic char *S. alpinus*).

The main part of these results was received with anonymous DNA sequences before the GenBank recourses became available, but they do not seem contrary to Crespi and Fulton [69] strong results with employment of a powerful tool of genomics (with the exception of taxonomic relations of *O. masou*).

Analysis of whitefishes' restriction data yielded the results not completely matching those generally accepted in the modern systematics of whitefishes [50, 70, 71]. From the point of view of systematics, Coregonidae is one of the most complex and intricate groups. Great variability and polymorphism of the whitefishes are the reason for the differences in conclusions about phylogenetic links between species based on different approaches, for example, Bernatchez et al. [72]; Smith and Todd [73]; Bodaly et al. [52]; Frolov [74], Turgeon and Bernatchez [30]. Investigation of genetic structure of the species and identification of closely related species from various sites of the range also cause difficulties. Up to a hundred intraspecific categories were described for a whitefish type species *C. lavaretus* from the Russian water basins only [50].

Interspecific and even, in some cases, intergeneric hybridization between the representatives of Coregonidae family yielding viable hybrids is a well-known fact, described by Garside and Christie a long time ago [4]. Casual relations between the diversity of the forms of whitefishes and introgression have been discussed more than once, for example, Svärdson [75] and Bernatchez et al. [76]. The possibility of exchange of genetic information in whitefish could be considered proved. It could be due to this that the index of genetic distances in them is usually lower than that in the taxa of the same level in other animals as reported by Bodaly et al. [52] and Kartavtsev [77]. Such phenomenon as

Table 4: *Salvelinus malma* DNA PCR-RAPD fragment sequences searched for NCBI.

Pairs of primers	PCR RAPD cloning fragment length**	Genbank Accession Numbers of homological sequences*		
		Microsatellites 5′ → 3′	dbEST*	Protein product***
IV + IV	555 bp		EG849746	
1 + IV	568 bp		DY73125	
IV + IV	463 bp		EG827041	BT_071991
1 + 1	304 bp		EG827041	
IV + 1	439 bp		EG930580	NP_001133464
IV + 1	676 bp		BX086452	
IV + IV	446 bp		CX357234	
IV + 1	395 bp		EG935616, DW006459	XP_003198377
1 + 1	418 bp	Microsatellite (CA)n 173–331		
1 + 1	905 bp		BX663609	XM 689046
2 + VII	409 bp	Microsatellite (GT)n 5–32		
3 + 3	549 bp		EG923517	NM_001102593
3 + 3	384 bp		DQ156149	XP_001332830
3 + 3	384 bp		CB511135.	
II + 3	122 bp		EU621899	XP_001336520
II + II	959 bp	Microsatellite (GT)n 226–294	AU081124	
3 + II	283 bp		CX723014	
3 + II	453 bp		CU062733	
3 + II	190 bp		GE828193	
3 + II	130 bp		GU552297	ADV31329
IV + 3	108 bp		EG911815	XP_003385009
3 + IV	236 bp		EG911136	ACH85273
3 + IV	389 bp.	Microsatellite (CA)n 276–329		
3 + IV	413 bp		EG792115	CBX11156.
3 + IV	724 bp		CA353611	
IV + 3	109 bp		EG911815	XP_001195378
3 + 3	801 bp		DW556963	ACI33792
3 + 3	496 bp		CB486060	
3 + 3	412 bp		EG831541	NP_001154053
3 + 3	322 bp		BX861631	XP_003197666
IV + IV	554 bp		EG849746	NP_001167305
3 + IV	290 bp		EG915402	NP_001135251
3 + 3	638 bp		CB509929	ACO08436
3 + IV	275 bp		CA388004	NP_001133389
3 + 3	685 bp		CK898369	NP_001167187
3 + 3	414 bp		FF839690	
IV + IV	254 bp		CB490887	ACO 57211 AAP58348
3 + IV	236 bp		EG911136	ACI66028
3 + 3	414 bp		EG792114	XP_001335224
3 + 3	490 bp		CB486060	ACI66769
7 + I	330 bp		EG 760735	XP_003200023
3 + IV	281 bp		BX861631	NP_001187967

*Only the first homological sequence from different libraries is exhibited.
**The lengths of cloning PCR fragments are given without considering primers.
***Translation performed by EMBOSS transeq (Sequence Translation Sites) of EBI-EMBL.

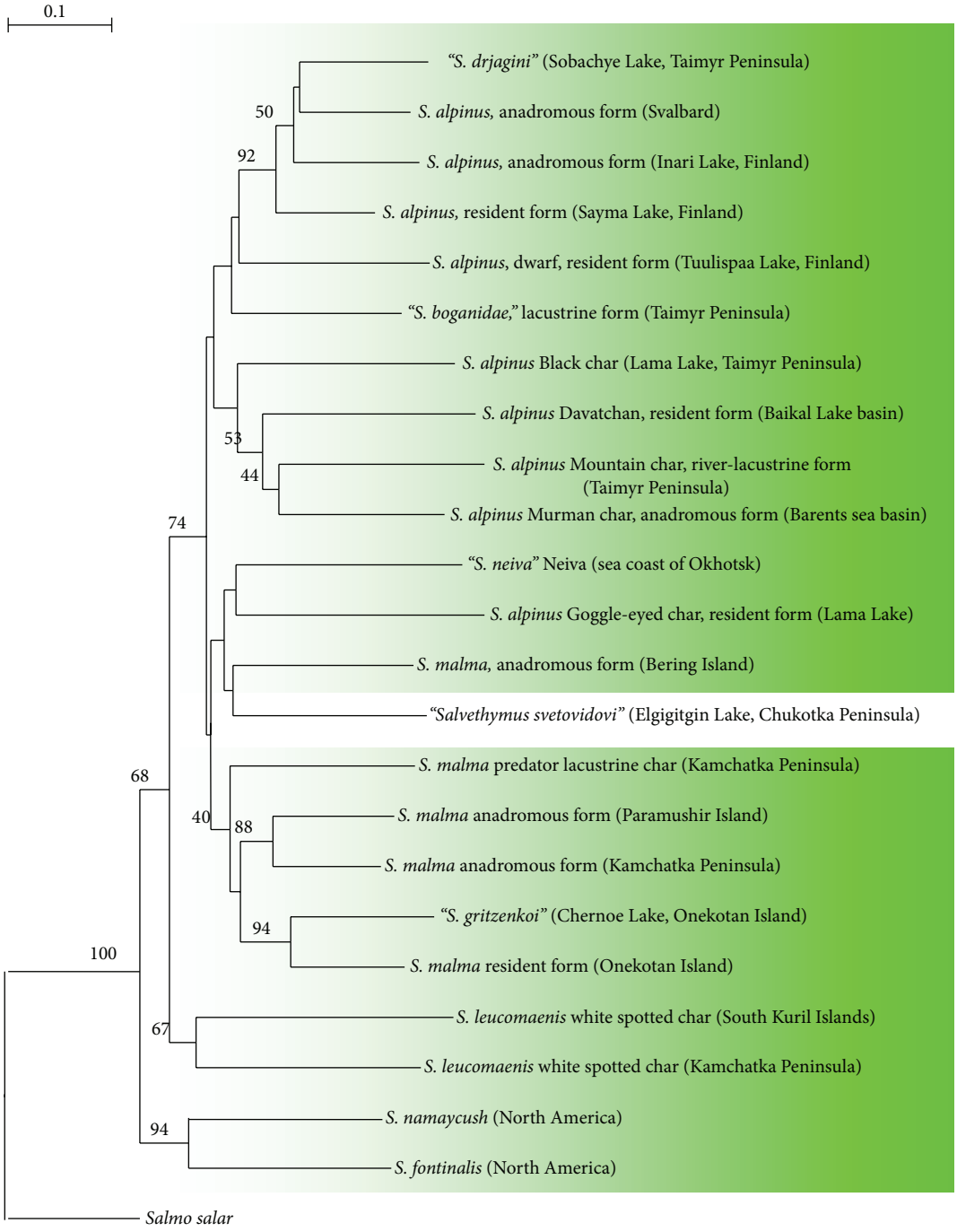


FIGURE 10: NJ tree of the g. *Salvelinus* forms and species, constructed in accordance with Nei genetic distances. The figures at the node indicate bootstrap indexes that exceed 40%.

"genetic parasitism," when the smaller and more numerous form replaces the much larger one, has been found and documented by Svardson in whitefishes [78]. Geographic isolation and isolation of other kinds as well as relatively young historical age (the main diversification of *Coregonus* species dates to Pleistocene glaciations—about 15,000 years in accordance with Behnke [51]) must have prevented various Coregonidae species from developing specific families of repetitive sequences, as it is common to noninterbreeding populations. On the contrary, many families of repetitive sequences are homogenous within all these forms, which indicate the intensive gene flow between all these species allowing molecular drive to adjust them.

All above mentioned enlightens us about greater rate of the differences in the experiments on amplification with arbitrary primers. They are the markers of the loci of expressed sequences and introns that are not subjected to molecular drive. However, no differences sufficient for reliable

discrimination of the species have been accumulated in these fractions either. Taking all the aforesaid into consideration, we can state that all these facts summed up vividly indicate a peculiar way of evolution and genetic structure of the Coregonidae population. Contrary to the common divergent evolution characteristic of the majority of animals, whitefishes demonstrate the elements reminding of reticulate evolution ("evolution via hybridization"), as supposed by Todd and Smith [29], Turgeon, Bernatchez [30], and Svardson [79], described for many plants, for example, Grant [80]. This pattern of evolution implies alternation of the divergent and hybridization (conversion) stages. External phenotypical differences in this case are determined by a small number of genes acting as morphological triggers and switching morphogenesis to a certain direction; the final studies of the process are known as discrete described forms written by Renaut and Bernatchez [81]. Accumulated cross breeding of the hybrids with prevailing parental species could become one of the possible mechanisms supporting morphological independence in the course of interspecies exchange of genetic information. However, any interpretation based on genetic isolation of Coregonidae species in the wild encounters the following unsolvable inconsistency. In the framework of such traditional hypothesis, one had to explain why the mutations in DNA repetitive sequences no longer occurred and were not preserved in this group, which seems incredible. Thus, the results obtained with the use of restriction analysis of highly repetitious DNA revealed great differences between electrophoretic patterns of the salmons with the high degree of reproductive isolation and the whitefishes. In the latter, introgressive hybridization has probably occurred, and the species still hybridize retaining their independent status.

Geographic isolation of the species is one of the mechanisms preventing interspecific mating. Over a long time, the view on allopatric speciation as a process of gradual accumulation of gene mutations of adaptive character located mostly in the coding DNA sequences has been commonly accepted (see, e.g., [82, 83]), and noncoding repetitive sequences were thought to be "junk" DNA. Now indirect evidence of involvement of the repetitive sequences (mostly mobile elements) in the adaptive evolution has been suggested by many researches—see for review Schmidt and Anderson [84] and Osada and Wu [83]. Studies of two sister *Drosophila* species have shown that heterochromatic region of X chromosome plays a great role in the establishment of the reproductive barrier [85]. Our results support this viewpoint.

4. Conclusion

The families of salmon fishes with different levels of reproductive isolation were compared using two strategies of multilocus fragment analysis. The compared lineages are (1) Salmonidae, possessing almost perfect homing and absolute reproductive isolation; (2) chars of g. *Salvelinus* (f. Salmonidae), possessing "good" and "difficult" species (reproductive isolation is often of spatial nature); (3) g. *Coregonus* (f. Coregonidae), containing recognized taxonomic species with common interspecies hybridization but distinct species phenotypes. Each sampling contained intraspecies forms and/or "disputable" species, "good" species, interbreeding species, and representatives of a sister genus.

Genetic distances were compared for lineages of the same taxonomic rank and juxtaposed with their divergence times; bootstrap support values were verified for corresponding phylogenetic clades. As a markers two types of sequences were chosen: (1) the long regions of satellite DNA, and (2) anonymous loci containing in 70% cases conserved exon and intron (or intergenic) regions. Genetic distances and clades robustness were shown to correlate well with the level of reproductive isolation in both marker systems.

The hypothesis of concerted evolution of satellite DNA is experimentally corroborated. The stronger is reproductive isolation between forms and species; the more species-specific band patterns are found in satellite DNA. Among whitefishes, the round fish *Prosopium cylindraceum* is the only reliably distinguished species separated from the unresolved *Coregonus* clade by a genetic distance comparable to those between individual genera in Salmonidae.

Molecular markers were used to clarify particular questions in salmon taxonomy and systematics: (1) *Salvelinus fontinalis* and *Salvelinus namaycush* are genetically close to each other; (2) *Salvethymus svetovidovi* cannot be considered a separate genus; (3) Dolly Varden *Salvelinus malma* is genetically identical to *Salvelinus alpinus*; (4) all species of the genera *Salmo*, *Parasalmo,* and *Oncorhynchus* are reliably distinguished, with the genus *Parasalmo* being sister to *Oncorhynchus*.

Conflict of Interests

The authors declare that they have no competing interests related to any of the mentioned products.

Acknowledgments

This study has been performed thanks to Boris M. Mednikov ideas. The authors thank K. A. Savvaitova (the Ichthyology Department of Lomonosov Moscow State University), Yu. S. Reshetnikov (A.N. Severtsov Institute of Ecology and Evolution Russian Academy of Sciences), M. Kaukorranta for sampling and important consultations, and A. Lomov for kindly placed some DNAs. The authors are grateful to Dr. Aleshin and Dr. Roussine for the assistance in preparation of the paper. The constructive remarks of two anonymous referees and referee for English corrections are acknowledged. The study was supported by Russian Federative Research Institute of Fisheries and Oceanology, the Russian Fond of Basic Researches: Grants 97-04-49753 and 01-04-48613, and Federal Program of the Ministry of Science and Education (the program Scientific and Scientific-Pedagogical Personnel of Innovative Russia).

References

[1] E. Mayr, *Animal Species and Evolution*, Harvard University Press, Cambridge, Mass, USA, 1963.

[2] T. Dobzhansky, *Genetics and the Origin of Species*, Columbia University Press, New York, NY, USA, 1937.

[3] J. Mallet, "Perspectives Poulton, Wallace and Jordan: how discoveries in *Papilio* butterflies led to a new species concept 100 years ago," *Systematics and Biodiversity*, vol. 1, no. 4, pp. 441–452, 2004.

[4] E. T. Garside and W. J. Christie, "Experimental hybridization among three coregonine fishes," *Transaction of the American Fishery Society*, vol. 9, no. 2, pp. 196–200, 1962.

[5] G. Svardson, "The Coregonid problem. VII. The isolation mechanism in sympatric species," *Reports of the Institute of Freshwater Research*, vol. 46, pp. 95–123, 1965.

[6] J. B. W. Wolf, J. Lindell, and N. Backström, "Speciation genetics: current status and evolving approaches," *Philosophical Transactions of the Royal Society B*, vol. 365, no. 1547, pp. 1717–1733, 2010.

[7] R. Gross, B. Gum, R. Reiter, and R. Kühn, "Genetic introgression between Arctic charr (*Salvelinus alpinus*) and brook trout (*Salvelinus fontinalis*) in Bavarian hatchery stocks inferred from nuclear and mitochondrial DNA markers," *Aquaculture International*, vol. 12, no. 1, pp. 19–32, 2004.

[8] K. T. Scribner, K. S. Page, and M. L. Bartron, "Hybridization in freshwater fishes: a review of case studies and cytonuclear methods of biological inference," *Reviews in Fish Biology and Fisheries*, vol. 10, no. 3, pp. 293–323, 2000.

[9] J. A. Coyne and H. A. Orr, "The evolutionary genetics of speciation," *Philosophical Transactions of the Royal Society B*, vol. 353, no. 1366, pp. 287–305, 1998.

[10] A. G. Hudson, P. Vonlanthen, R. Müller, and O. Seehausen, "Review: the geography of speciation and adaptive radiation in coregonines. Biology and management of coregonid fishes," *Archives of Hydrobiology Special Issue Advances in Limnology*, vol. 60, no. 2, pp. 111–146, 2007.

[11] J. M. Wright, "Nucleotide sequence, genomic organization and evolution of a major repetitive DNA family in tilapia (*Oreochromis mossambicuslhomorum*)," *Nucleic Acids Research*, vol. 17, no. 13, pp. 5071–5079, 1989.

[12] J. A. Shapiro and R. von Sternberg, "Why repetitive DNA is essential to genome function," *Biological Reviews of the Cambridge Philosophical Society*, vol. 80, no. 2, pp. 227–250, 2005.

[13] M. A. Biscotti, A. Canapa, E. Olmo et al., "Repetitive DNA, molecular cytogenetics and genome organization in the King scallop (*Pecten maximus*)," *Gene*, vol. 406, no. 1-2, pp. 91–98, 2007.

[14] S. Ohno, *Evolution By Gene Duplication*, Springer, New York, NY, USA, 1970.

[15] W. Chase, A. V. Cox, D. E. Soltis et al., "Large DNA sequences matrices phylogenetic signal and feasibility: an empirical approach," in *Proceedings of the American Society of Plant Taxonomists Annual Meetings*, Montreal, Canada, August 1997.

[16] J. D. Thompson, J. E. Sylvester, I. Laudien Gonzalez, C. C. Constanzi, and D. Gillespie, "Definition of a second dimeric subfamily of human α satellite DNA," *Nucleic Acids Research*, vol. 17, no. 7, pp. 2769–2782, 1989.

[17] P. Warburton and H. Willard, "Evolution of centromeric alpha satellite DNA: molecular organization within and between human and primate chromosomes," in *Human Genome Evolution*, M. S. T. Jackson and G. Dover, Eds., pp. 121–145, BIOS Scientific Publishers, 1996.

[18] T. Ohta, "On the evolution of multigene families," *Theoretical Population Biology*, vol. 23, no. 2, pp. 216–240, 1983.

[19] E. A. Zimmer, S. L. Martin, and S. M. Beverley, "Rapid duplication and loss of genes coding for the α chains of hemoglobin," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 77, no. 4, pp. 2158–2162, 1980.

[20] A. J. Jeffreys, V. Wilson, and S. L. Thein, "Hypervariable "minisatellite" regions in human DNA," *Nature*, vol. 314, no. 6006, pp. 67–73, 1985.

[21] K. M. Gray, J. W. White, C. Costanzi et al., "Recent amplification of an alpha satellite DNA in humans," *Nucleic Acids Research*, vol. 13, no. 2, pp. 521–535, 1985.

[22] J. F. Elder Jr. and B. J. Turner, "Concerted evolution at the population level: pupfish HindIII satellite DNA sequences," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 91, no. 3, pp. 994–998, 1994.

[23] A. N. Fedorov, L. V. Fedorova, V. V. Grechko et al., "Variable and invariable DNA repeat characters revealed by taxonprint approach are useful for molecular systematics," *Journal of Molecular Evolution*, vol. 48, no. 1, pp. 69–76, 1999.

[24] W. M. Brown, "Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 77, no. 6, pp. 3605–3609, 1980.

[25] V. V. Grechko, L. V. Fedorova, A. N. Fedorov et al., "Restriction endonuclease analysis of highly repetitive DNA as a phylogenetic tool," *Journal of Molecular Evolution*, vol. 45, no. 3, pp. 332–336, 1997.

[26] I. A. Roudykh, V. V. Grechko, D. G. Ciobanu, D. A. Kramerov, and I. S. Darevsky, "Variability of restriction sites in satellite DNA as a molecular basis of taxonoprint method: evidence from the study of Caucasian rock lizards," *Russian Journal of Genetics*, vol. 38, no. 8, pp. 937–941, 2002.

[27] B. M. Mednikov, E. A. Shubina, M. N. Mel'nikova et al., "The problem of the generic status of Pacific salmon and trout (Genetic Taxonomic Analysis)," *Journal of Ichthyology*, vol. 39, no. 1, pp. 10–17, 1999.

[28] L. Bernatchez and J. J. Dodson, "Allopatric origin of sympatric populations of lake whitefish (*Coregonus clupeaformis*) as revealed by mitochondrial-DNA restriction analysis," *Evolution*, vol. 44, no. 5, pp. 1263–1271, 1990.

[29] T. N. Todd and G. R. Smith, "A review of differentiation in Great Lakes ciscoes," *Polskie Archiwum Hydrobiologii*, vol. 39, no. 3-4, pp. 261–267, 1992.

[30] J. Turgeon and L. Bernatchez, "Reticulate evolution and phenotypic diversity in North American ciscoes, *Coregonus* ssp. (Teleostei: Salmonidae): implications for the conservation of an evolutionary legacy," *Conservation Genetics*, vol. 4, no. 1, pp. 67–81, 2003.

[31] G. R. Smith and R. F. Stearley, "The classification and scientific names of rainbow and cutthroat trouts," *Fisheries*, vol. 14, no. 1, pp. 4–10, 1989.

[32] J. Welsh and M. McClelland, "Genomic fingerprinting using arbitrarily primed PCR and a matrix of pairwise combinations of primers," *Nucleic Acids Research*, vol. 19, no. 19, pp. 5275–5279, 1991.

[33] J. G. K. Williams, M. K. Hanafey, J. A. Rafalski, and S. V. Tingey, "Genetic analysis using random amplified polymorphic DNA markers," *Methods in Enzymology*, vol. 218, pp. 704–740, 1993.

[34] F. E. Arrighi, J. Bergendahl, and M. Mandel, "Isolation and characterization of DNA from fixed cells and tissues," *Experimental Cell Research*, vol. 50, no. 1, pp. 40–47, 1968.

[35] S. E. Saunders and J. F. Burke, "Rapid isolation of miniprep DNA for double strand sequencing," *Nucleic Acids Research*, vol. 18, no. 16, pp. 4948–4950, 1990.

[36] M. G. Murray and W. F. Thompson, "Rapid isolation of high molecular weight plant DNA," *Nucleic Acid Research*, vol. 8, no. 19, pp. 4321–4325, 1980.

[37] R. Boom, C. J. A. Sol, M. M. M. Salimans, C. L. Jansen, P. M. E. Wertheim-Van Dillen, and J. Van Der Noordaa, "Rapid and simple method for purification of nucleic acids," *Journal of Clinical Microbiology*, vol. 28, no. 3, pp. 495–503, 1990.

[38] J. Sambrook, E. F. Fritsch, and T. Maniatis, *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Publishing, New York, NY, USA, 2nd edition, 1989.

[39] T. Maniatis, E. E. Fritch, and J. Sambrook, *Molecular Cloning. A Laboratory Manual*, Cold Spring Harbor Laboratory, 1982.

[40] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Molecular Biology and Evolution*, vol. 4, no. 4, pp. 406–425, 1987.

[41] M. Nei and W. H. Li, "Mathematical model for studying genetic variation in terms of restriction endonucleases," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 76, no. 10, pp. 5269–5273, 1979.

[42] M. Nei, "Genetic distance between populations," *American Naturalist*, vol. 106, no. 949, pp. 283–292, 1972.

[43] Y. Van De Peer and R. De Wachter, "Treecon for windows: a software package for the construction and drawing of evolutionary trees for the microsoft windows environment," *Bioinformatics*, vol. 10, no. 5, pp. 569–570, 1994.

[44] J. Felsenstein, "Confidence limits on phylogenies: an approach using the bootstrap," *Evolution*, vol. 39, no. 4, pp. 783–791, 1985.

[45] V. A. Maksimov, K. A. Savvaitova, B. M. Mednikov et al., "Alpine char-a new form of Arctic char (*Salvelinus*) from Taimyr," *Journal of Ichthyology*, vol. 35, no. 7, pp. 1–9, 1995.

[46] R. J. Behnke, *Native Trout of Western North America*, vol. 6, American Fisheries Society Monograph, Bethesda, Md, USA, 1992.

[47] A. G. Osinov and V. S. Lebedev, "Salmonid fishes (Salmonidae, Salmoniformes): the systematic position in the superorder Protacanthopterygii -the main stages of evolution, and molecular dating," *Journal of Ichthyology*, vol. 44, no. 9, pp. 690–715, 2004.

[48] I. A. Chereshnev and M. B. Skopets, "Salvethymus svetovidovi gen. at sp. nova, a new endemic fish from subfamily Salmonidae from Lake El'gygytgyn (the Central Chukotka)," *Journal of Ichthyology*, vol. 30, no. 2, pp. 201–213, 1990.

[49] T. M. Cavender, "Review of the fossil history of North American freshwater fishes," in *The Zoogeography of North American Freshwater Fishes*, C. H. Hocutt and E. O. Wiley, Eds., pp. 700–724, John Wiley & Sons, NewYork, NY, USA, 1986.

[50] Yu. S. Reshetnikov, *Ecology and Systematics of Coregonine Fishes*, Nauka, Moscow, Russia, 1980.

[51] R. J. Behnke, "The systematics of salmonid fishes of recently glaciated lakes," *Journal of Fisheries Research Board of Canada*, vol. 29, no. 5, pp. 639–671, 1972.

[52] R. A. Bodaly, J. Vuorinen, R. D. Ward, M. Luczynski, and J. D. Reist, "Genetic comparisons of New and Old World coregonid fishes," *Journal of Fish Biology*, vol. 38, no. 1, pp. 37–51, 1991.

[53] G. Kh. Shaposhnikova, "On the taxonomy of whitefishes from the USSR," in *Biology of Coregonid Fishes*, C. C. Lindsey and C. S. Woods, Eds., pp. 195–208, University Manitoba Press, Winnipeg, Manitoba, Canada, 1970.

[54] K. A. Savvaitova, "Patterns of diversity and processes of speciation in Arctic charr," *Nordic Journal of Freshwater Research*, vol. 71, no. 1, pp. 81–91, 1995.

[55] Yu. S. Reshetnikov, Ed., *Atlas of Russian Freshwater Fishes*, vol. 1, Nauka, Moscow, Russia, 2002.

[56] V. V. Barsukov, "About systematic of the Chukotka charr of the genus *Salvelinus*," *Journal of Ichtiology*, vol. 14, no. 1, pp. 3–17, 1960.

[57] A. G. Osinov and S. D. Pavlov, "Allozyme variation and genetic divergence between populations of Arctic char and Dolly Varden (*Salvelinus alpinus-Salvelinus malma* complex)," *Journal of Ichthyology*, vol. 38, no. 1, pp. 42–55, 1998.

[58] A. G. Osinov, "Evolutionary relationships between the main taxa of the *Salvelinus alpinus-Salvelinus malma* complex: results of a comparative analysis of allozyme data from different authors," *Journal of Ichthyology*, vol. 41, no. 3, pp. 192–208, 2001.

[59] L. S. Berg, "Fishes of the Amur River basin," *Memoires de l'Academie Imperial des Sciences Classe Physico-Math*, vol. 24, no. 9, pp. 1–270, 1909.

[60] S. S. Alekseev and M. Y. Pichugin, "A new form of charr *Salvelinus alpinus* (Salmonidae) from Lake Davatchan in Transbaikalia and its morphological differences from sympatric forms," *Journal of Ichthyology*, vol. 38, no. 4, pp. 292–302, 1998.

[61] B. F. Koop, K. R. von Schalburg, J. L. N. Walker et al., "A salmonid EST genomic study: genes, duplications, phylogeny and microarrays," *BMC Genomics*, vol. 9, p. 545, 2008.

[62] A. G. Osinov and V. S. Lebedev, "Genetic divergence and phylogeny of the Salmoninae based on allozyme data," *Journal of Fish Biology*, vol. 57, no. 2, pp. 354–381, 2000.

[63] P. A. Crane, L. W. Seeb, and J. E. Seeb, "Genetic relationships among Salvelinus species inferred from allozyme data," *Canadian Journal of Fisheries and Aquatic Science*, vol. 51, supplement 1, pp. 182–197, 1994.

[64] S. Kumar and S. B. Hedges, "A molecular timescale for vertebrate evolution," *Nature*, vol. 392, no. 6679, pp. 917–920, 1998.

[65] J. P. Thorpe, "The molecular clock hypothesis: biochemical evolution, genetic differentiation and systematics," *Annual Review of Ecology, Evolution, and Systematic*, vol. 13, pp. 139–168, 1982.

[66] A. H. Berst, A. R. Emery, and G. R. Spangler, "Reproductive behavior of hybrid charr (*Salvelinus fontinalis x S namaycush*)," *Canadian Journal of Fisheries and Aquatic Science*, vol. 38, no. 4, pp. 432–440, 1981.

[67] K. M. Westrich, N. R. Konkol, M. P. Matsuoka, and R. B. Phillips, "Interspecific relationships among charrs based on phylogenetic analysis of nuclear growth hormone intron sequences," *Environmental Biology of Fishes*, vol. 64, no. 1–3, pp. 217–222, 2002.

[68] S. U. Qadri, "Morphological comparisons of three populations of the lake charr, Cristivomer namaykush, from Ontario and Manitoba," *Journal of the Fisheries Research Board of Canada*, vol. 24, pp. 1407–1411, 1967.

[69] B. J. Crespi and M. J. Fulton, "Molecular systematics of Salmonidae: combined nuclear data yields a robust phylogeny," *Molecular Phylogenetics and Evolution*, vol. 31, no. 2, pp. 658–679, 2004.

[70] L. N. Ermolenko, "Genetic divergence in the family Coregonidae," *Polskie Archiwum Hydrobiologii*, vol. 39, no. 3-4, pp. 533–539, 1992.

[71] D. V. Politov, N. Y. Gordon, and A. A. Makhrov, "Genetic identification and taxonomic relationships of six Siberian species of *Coregonus*," *Advances in Limnology*, vol. 57, pp. 21–34, 2002.

[72] L. Bernatchez, F. Colombani, and J. J. Dodson, "Phylogenetic relationships among the subfamily Coregoninae as revealed by mitochondrial DNA restriction analysis," *Journal of Fish Biology*, vol. 39, supplement, pp. 283–290, 1991.

[73] G. R. Smith and T. N. Todd, "Morphological cladistic study of Coregonine fishes," *Polskie Archiwum Hydrobiologii*, vol. 39, no. 3-4, pp. 479–490, 1992.

[74] S. V. Frolov, "Some aspects of kariotype evolution in the Coregoninae," *Polskie Archiwum Hydrobiologii*, vol. 39, no. 3-4, pp. 509–515, 1992.

[75] G. Svärdson, "Significance of introgression in coregonid evolution," in *Biology of Coregonid Fishes*, C. C. Lindsey and C. S. Woods, Eds., pp. 33–59, University of Manitoba Press, Winnipeg, Manitoba, Canada, 1970.

[76] L. Bernatchez, S. Renaut, A. R. Whiteley et al., "On the origin of species: insights from the ecological genomics of lake whitefish," *Philosophical Transactions of the Royal Society B*, vol. 365, no. 1547, pp. 1783–1800, 2010.

[77] Yu. Kartavtsev, "Genetic variability and differentiation in Salmonid fish," *Nordic Journal of Freshwater Research*, vol. 67, pp. 96–117, 1992.

[78] G. Svardson, "The Coregonid problem. VI. The palearctic species and their intergrades," *Reports of the Institute of Freshwater Research Drottningholm*, vol. 38, pp. 267–356, 1975.

[79] G. Svardson, "Postglacial dispersal and reticulate evolution of Nordic Coregonids," *Nordic Journal of Freshwater Research Drottningholm*, vol. 74, pp. 3–32, 1998.

[80] V. Grant, *Plant Speciation*, Columbia University Press, New York, NY, USA, 2nd edition, 1981.

[81] S. Renaut and L. Bernatchez, "Transcriptome-wide signature of hybrid breakdown associated with intrinsic reproductive isolation in lake whitefish species pairs (*Coregonus* spp. *Salmonidae*)," *Heredity*, vol. 106, no. 6, pp. 1003–1011, 2011.

[82] K. Johannesson, "Parallel speciation: a key to sympatric divergence," *Trends in Ecology and Evolution*, vol. 16, no. 3, pp. 148–153, 2001.

[83] N. Osada and C.-I. Wu, "Inferring the mode of speciation from genomic data: a study of the great apes," *Genetics*, vol. 169, no. 1, pp. 259–264, 2005.

[84] A. L. Schmidt and L. M. Anderson, "Repetitive DNA elements as mediators of genomic change in response to environmental cues," *Biological Reviews of the Cambridge Philosophical Society*, vol. 81, no. 4, pp. 531–543, 2006.

[85] P. M. Ferree and D. A. Barbash, "Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in *Drosophila*," *PLoS Biology*, vol. 7, no. 10, Article ID e1000234, 2009.

Hindawi Publishing Corporation
International Journal of Genomics
Volume 2013, Article ID 963956, 13 pages
http://dx.doi.org/10.1155/2013/963956

Research Article

Periodic Distribution of a Putative Nucleosome Positioning Motif in Human, Nonhuman Primates, and Archaea: Mutual Information Analysis

Daniela Sosa,[1,2] Pedro Miramontes,[1,2] Wentian Li,[3] Víctor Mireles,[1,2] Juan R. Bobadilla,[4] and Marco V. José[4,5]

[1] *Facultad de Ciencias, Universidad Nacional Autónoma de México, 04510 CP, DF, Mexico*
[2] *Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, 04510 CP, DF, Mexico*
[3] *The Robert S. Boas Center for Genomics and Human Genetics Manhasset, Feinstein Institute for Medical Research, North Shore LIJ Health System, Manhasset, NY, USA*
[4] *Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, 04510 CP, DF, Mexico*
[5] *Centro Internacional de Ciencias, Cuernavaca, Morelos, Mexico*

Correspondence should be addressed to Marco V. José; marcojose@biomedicas.unam.mx

Received 13 February 2013; Accepted 29 April 2013

Academic Editor: Ancha Baranova

Recently, Trifonov's group proposed a 10-mer DNA motif YYYYYRRRRR as a solution of the long-standing problem of sequence-based nucleosome positioning. To test whether this generic decamer represents a biological meaningful signal, we compare the distribution of this motif in primates and Archaea, which are known to contain nucleosomes, and in Eubacteria, which do not possess nucleosomes. The distribution of the motif is analyzed by the mutual information function (MIF) with a shifted version of itself (MIF profile). We found common features in the patterns of this generic decamer on MIF profiles among primate species, and interestingly we found conspicuous but dissimilar MIF profiles for each Archaea tested. The overall MIF profiles for each chromosome in each primate species also follow a similar pattern. Trifonov's generic decamer may be a highly conserved motif for the nucleosome positioning, but we argue that this is not the only motif. The distribution of this generic decamer exhibits previously unidentified periodicities, which are associated to highly repetitive sequences in the genome. *Alu* repetitive elements contribute to the most fundamental structure of nucleosome positioning in higher Eukaryotes. In some regions of primate chromosomes, the distribution of the decamer shows symmetrical patterns including inverted repeats.

1. Introduction

It is generally accepted that the chromatin organization of eukaryotic DNA is strongly governed by a code inherent to the DNA sequence. Modulating the accessibility of individual DNA sequences involves many complex interactions, the most prevalent of which are the interactions between histone octamers and DNA in compacted chromosomes [1, 2]. The condensation of DNA into an ordered chromatin structure allows the cell to solve the topological problems associated with storing huge amount of information of chromosomal DNA within the nucleus. In Eukaryotes and Archaea, DNA is packaged into chromatin in orderly repetitive protein-DNA complexes called nucleosomes. Each nucleosome consists of approximately 146-147 bp of dsDNA wound 1.7-1.8 times around a histone octamer [3–5] to form the basic unit of chromatin structure, the nucleosome. Each octamer is composed of two H3-H4 histone dimers bridged together as a stable tetramer that is flanked by two separate H2A-H2B dimers [6]. Stretches of DNA called linker up to 100 bp, often with an increment of 10 bp, separate adjacent nucleosomes. Multiple nuclear proteins bind to this linker region, some of which may be responsible for the ordered wrapping of strings of nucleosomes into higher-order chromatin structures [7].

Histone proteins condense DNA into complex nucleosome structures both in Eukaryotes and Archaea [2, 8]. Nucleosomes were originally regarded as a distinguishing feature of Eukaryotes prior to identification of histone orthologs in Archaea [9, 10]. The underlying DNA sequence, sometimes called "nucleosome core sequence" or "nucleosome positioning sequence," acts to bias its own packaging in nucleosomes through preferential positioning of histone octamer. It can facilitate DNA wrapping by placing AA dinucleotides along the portion of the DNA helix that faces the histone core complex [11–13]. Thus, DNA sequences that favor nucleosome formation are enriched with AA dinucleotides spaced ~10 bp apart, resulting in a deficiency of TT dinucleotides at the same location and on the strand facing the histone [11–14]. Five to six nucleotides in either direction, where the complementary strand faces the histone core, the trend is reversed (TT enrichment and a deficit of AA). Two main classes of nucleosome positioning sequence (NPS) patterns have been described. In the first class, AA, TT, and other WW dinucleotides (W = A or T) tend to occur together (in phase) in the major groove of DNA closest to the histone octamer surface, while SS dinucleotides (S = G or C) are predominantly positioned in the major groove facing outward. In the second class, AA and TT are structurally separated (AA backbone near the histone octamer and TT backbone further away), but grouped with other RR (where R is purine A or G) and YY (where Y is pyrimidine C or T) dinucleotides. As a result, the RR/YY pattern includes counterphase AA/TT distributions [15].

In the literature, nucleosome positioning is widely regarded as being sequence specific, enabling them with features of regulation of the access of nonhistone proteins to DNA *in vivo* (e.g., [16]). Albeit, the sequence-dependency of nucleosome positioning is still under debate (see, e.g., [16–21]), the fact that histone proteins in Eukaryotes are highly conserved whereas the genome sequences and the positioning sequence motifs seem to be highly divergent among organisms opens an intriguing question.

Both DNA sequence and nucleosome positioning are important factors in gene regulation [22–24]. Accessibility of transcription binding sites crucially depends on the nucleosome positioning [25, 26]. Nucleosomes are distributed in a highly nonrandom fashion around transcription start sites [27, 28]. Replication is dependent on nucleosome positioning [29].

Yet the so-called chromatin code has not been fully determined. This code is a well hidden, weak periodical DNA sequence pattern that is recognized by histone octamers. However, the weak signal is not a problem for the histone octamer. It may select the best bendable segments in random DNA sequences. Additionally, as experimental nucleosome mapping indicates, most of the nucleosomes have only marginal stability [13, 29]. It does not mean, however, that their positions are fully uncertain [30, 31]—as much as 50% contribution may come from sequence itself to determine whether a region is covered by a nucleosome or not [16].

The original assumption that DNA sequence is the major factor in nucleosome positioning was first made as early as 1975 [32] and later in 1984 [33] and confirmed afterwards [34, 35]. However, the exact formulation of the positioning pattern remained elusive. Recently, Trifonov's group has provided a pattern that they claim to be an ultimate solution of the long-standing problem of sequence-based nucleosome positioning [36]. Two basic binary periodical patterns are well established: in purine/pyrimidine alphabet—YRRRRRYYYYYR and in strong/weak alphabet—SWWWWWSSSSSW (S/W). Their merger (shifted by 5 bases) in four-letter alphabet sequence coincides with the first complete matrix of nucleosome DNA bendability [37], which was derived from a large database of nucleosome core DNA sequences generated by micrococcal nuclease (MNase) digestion of *C. elegans* chromatin [38, 39]. The results from the bendability analysis indicate that the sequence CGGAAATTTC, called a CG/AT motif, with CG and AT elements 5 bases apart, is predominant in nucleosome cores at the centers of complementary symmetry of the consensus nucleosome-binding pattern derived from bendability data. A more inclusive, but consistent with all previous proposals, consensus nucleosome positioning pattern observed in *C. elegans* was $(YYYYYRRRRR)_n$. Note that on the reverse complementary strand, the motif is still YYYYYRRRRR (Y/R), but if shifted by 5 bases, it becomes RRRRRYYYYY (R/Y) [40].

The solution was claimed by Trifonov's group to be unique, hence universal, since the physics of DNA bendability should, in principle, be the same for all species [36]. The simple higher occurrence common consensus of the motifs is TTTCCGGAAA, which is identical to their CG/AT motif derived from *C. elegans* nucleosomes [25, 41]. None of other suggested motifs scores better when compared to the rest of the set. Indeed, the experimental data on *C. elegans* were convincingly consistent with the decamer YYYYYRRRRR in regard to its association to nucleosome positioning partly because the motif was derived from the *C. elegans* MNase digestion data. This alone is a good reason to believe that the CG/AT sequence, as well as the more general YYYYYRRRRR motif, is a universal DNA bendability pattern. Another reason is that this motif can be derived from simple DNA deformability considerations, by minimizing unstacking of bases and base pairs caused by DNA bending on the surface of the histone decamer [36].

Analysis of periodicities in 13 fully sequenced eukaryotic genomes [42] showed that weakly periodically positioned TA dinucleotides are detected only in *Saccharomyces cerevisiae*.

The rationale of our work is as follows. If the generic decamer possesses inherent stability properties making it a universal nucleosome positioning sequence throughout Eukarya, we hypothesized that this decamer signal, caused by a regular spacing of nucleosomes, could also be detected in Archaea, whereby vestiges of primitive nucleosome structures could be identified [10, 43], but lacking in Eubacteria where the nucleosome structure does not exist. The goal of this work was to test the universality hypothesis of the putative nucleosome motif YYYYYRRRRR. To this end, we used mutual information function (MIF) profiles of the generic decamer YYYYYRRRRR along the entire genomes of 3 primate species and 4 species of Archaea. We also tested the S/W motif in all organisms. We show that the overall MIF profiles for the Y/R decamer for each chromosome in each

primate species followed a similar periodic pattern, whereas the S/W motif is regular but only in a few chromosomes of primate species. In Archaea species the MIF profiles were different but showed conspicuous periodic features. Hence, with the assumption that an appropriate periodic signal is an indication of the regular spacing of nucleosomes, the Y/R decamer seems to be a highly conserved motif of nucleosome positioning. We used as controls genomes of 3 bacteria, in which there are no nucleosomes, to show that the periodic signal is absent.

On the other hand, the long distance of the regular spacing reflects a low density of the Y/R decamer in these genomes. One implication is that the decamer may not occupy positions at every helix turn, more likely at every nucleosome. Another implication is that other motifs beside this decamer may play a role in the nucleosome positioning.

To further test whether decamer Y/R was able to cast light upon the nucleosome positioning, we generated 10 random sequences of decamers preserving the 5 Ys and 5 Rs content for each chromosome. We found that the random decamers did not present clear-cut patterns in the MIF profiles along chromosomes in contrast to Trifonov's decamer.

Our work is consistent with the assumption that Trifonov's generic decamer is one of the nucleosome positioning motifs in primates and in Archaea, and nucleosomes are regularly spaced. However, this motif was derived by conditioning on CG (or AA, AT, TT) as the flanking dinucleotide with periodicity of 10 (CG-8-CG, or AA-8-AA, etc.), which excludes any nucleosome positioning motifs that do not have these periodicities—10 to start with. There may be other motifs that may be associated to nucleosome positioning. This statement comes from our observation that Trifonov's decamer is not found with the same frequencies along different regions of a given chromosome, different local regions within a gene, or GC-rich versus GC-poor segments, even when the DNA is indeed uniformly supercoiled. Actually, there are long stretches in which the generic decamer is absent.

For comparison purposes and for validation of the use of the MIF, here we also report the same analysis in the five chromosomes of *C. elegans*, for which experimental data are available and certain results are expected [41]. With our approach we found that this motif not only reflects well-known periodicities of the nucleosome positions but also there seems to be other previously unidentified periodicities both in primates and Archaea. We conclude that Trifonov's decamer is not the "one-and-only" universal nucleosome positioning motif. We give evidence that these periodicities are associated with highly repetitive sequences in primate genomes. In particular, we show that the Y/R motif is clearly associated to *Alu* repetitive elements in primate species.

2. Materials and Methods

2.1. Data Sources. Human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), and macaque (*Macaca mulatta*) complete genome sequences were downloaded from NCBI released, respectively, in March, October, and June of 2006 from ftp://ftp.ncbi.nih.gov/genomes/. In particular, the whole genomes of human, chimpanzee, and rhesus macaque were downloaded from: ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/; ftp://ftp.ncbi.nih.gov/genomes/Pan_troglodytes/, and ftp://ftp.ncbi.nih.gov/genomes/Macaca_mulatta/, respectively.

We selected the following Archaea which were also downloaded from the NCBI website: *Methanocaldococcus jannaschii, Sulfolobus solfataricus, Nanoarchaeum equitans, Archaeoglobus fulgidus,* with the corresponding accession numbers: NC_000909, NC_002754, NC_005213, NC_00917. The selected Eubacteria used as controls are: *Escherichia coli, Pseudomonas fluorescens*, and *Deinococcus radiodurans R1* with accession numbers: NC_000913, NC_004129, and NC_001263.1, respectively.

2.2. An Overview of the Mutual Information Function. Initially the mutual information (MI) was used to measure the difference between the average uncertainty in the input of an information channel before and after the outputs were received [44]. The MI is a general measure of correlation between discrete variables, analogous to the Pearson product-moment correlation coefficient for continuous variables. For symbolic sequences, MI between two symbols separated by a distance k is a function of k, called mutual information function (MIF) [45]. The MIF is particularly useful for analyzing correlation properties of symbolic sequences [45].

Let us denote by $A = \{a, t, g, c\}$ an alphabet and by $s = (\ldots, a_0, a_1, \ldots)$ an infinite string with $a_i \in A$, $i \in \mathbb{Z}$, where $\mathbb{Z}$ represents the set of all integer numbers and the values of a_i can be repeated. The MIF of the string s and an identical string shifted k positions upstream is defined as

$$I(k, s) = \sum_{\alpha \in A} \sum_{\beta \in A} P_{\alpha,\beta}(k, s) \log_2 \left[\frac{P_{\alpha,\beta}(k, s)}{P_\alpha(s)\, P_\beta(s)} \right], \quad (1)$$

where $P_{\alpha,\beta}(k, s)$ is the joint probability of having the symbol α followed k sites away by the symbol β on the string s and $P_\alpha(s)$ and $P_\beta(s)$ are the marginal probabilities of finding α or β in the string s. By choosing the logarithm in base 2, $I(k, s)$ is measured in bits. Both the joint probability and the marginal probabilities are estimated throughout the sequence as a global property. The function $I(k, s)$ can be interpreted as the average information over all positions that one can obtain about the actual value of a certain position in the string, given that one knows the actual value of the position k-characters away. The mutual information vanishes if, and only if, the events are statistically independent, that is, if all 16 joint probabilities $P_{\alpha,\beta}(k, s)$ factorize. Thus, the MIF is a function capable of detecting any deviation from statistical independence. It must be noted from (1) that $I(k, s) \geq 0$. Computing the MIF for a given sequence using different shifts of magnitude k provides an autocorrelation profile.

2.3. The MIF Profile. In this work, we calculated for each given sequence s the contribution made to $I(k, s)$ by the generic decamers YYYYYRRRRR and SWWWWWSSSSSW. For this purpose, we computed $I(k, s)$ of the sequence s and then marked s such that each occurrence of, say, the decamer YYYYYRRRRR appeared in upper case, thereby extending

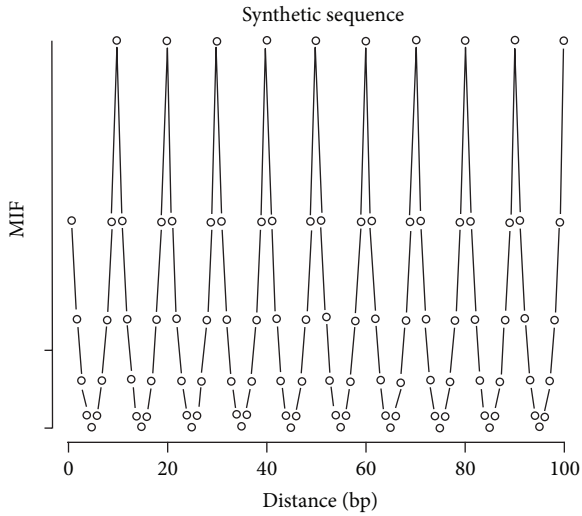


Figure 1: MIF profile of the decamer YYYYYRRRRR from a synthetic sequence. Note the 10-base periodicity.

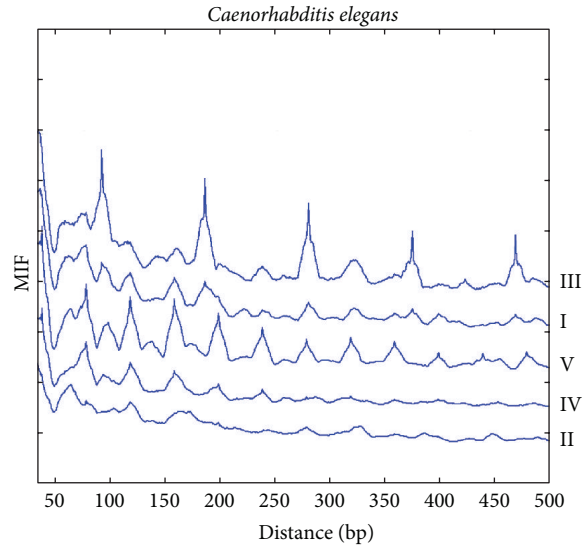


Figure 2: The MIF profiles of the decamer YYYYYRRRRR in 5 *C. elegans* chromosomes.

the alphabet to $A' = \{a, t, g, c, A, T, G, C\}$. If we call this marked sequence s', then the difference $I(k, s') - I(k, s)$ is a measure of how much additional information the decamer YYYYYRRRRR contributes to our prediction of the content of a position in the sequence k spaces away from a position whose information content is already known. This renders a brief description of how much of the correlations of a given chromosome are due specifically to the occurrences of the decamer YYYYYRRRRR. MIF, being similar to the autocorrelation function, is a method to detect periodicity in a sequence. A peak in MIF at spacing k indicates that the decamer prefers a spacing of k bases. In order to test our MIF profile, we generated a synthetic DNA sequence in which the decamer YYYYYRRRRR was placed at regular intervals (Figure 1). Note that the MIF profile clearly exhibits a 10-base periodicity.

For each chromosome the MIF was computed for k between 1 and 500. Besides this excess mutual information between symbol and symbol (base and base k-position away), an alternative measure of the decamer-decamer correlation is to convert a DNA sequence to a binary (0/1) sequence: 1 for an appearance of YYYYYRRRRR, 0 otherwise. These two methods lead to equivalent results.

Since tandem repeats of YYYYYRRRRR leads to periodicities at $k = 10, 20, \ldots$, in our MIF and since regular spacing of nucleosomes (e.g., 146 bp plus a 45 linker length corresponds to a spacing of 191 bp) leads to periodicity at, for example, 191, 382, ..., any periodicities at short (<150) and intermediate (>150 and <400) distances in MIF may indirectly confirm the role of YYYYYRRRRR in nucleosome positioning.

This strategy is played out at several levels; we expect to see a periodic presence (absence) of peaks in the MIF profile for genomes known to possess nucleosomes (in those known to have no nucleosomes). We expect to see peaks at both short and intermediate distances. Finally, any observations in contrast to our expectation may lead to new insight; for example, the absence of peaks when expected may point to other nucleosome positioning motifs not included in YYYYYRRRRR; or presence of peaks at unexpected distances may point to other roles of the YYYYYRRRRR motif.

3. Results

3.1. MIF Profiles of C. elegans. Since the decamer was derived from the *C. elegans* MNase digestion data, we expect periodicities to be present in the MIF profile, either due to the tandem repeats of the decamer or due to the regular spacing of the nucleosomes. The MIF profiles of the decamer on chromosomes I, III, and V, but not on chromosome II or IV, of *C. elegans* display a regular pattern of peaks that appear every 10, 20, 40, and 92–94 bp approximately (Figure 2), and they correspond to distance histograms (not shown). This pattern is even met by chromosome X (not shown). The MIF profile of chromosome V shows regular spacings of multiples of 20 bp (e.g., at 120, 160, 200, 240, 320, 360 400, and 480). Given a decamer, we would have expected that bumps (a lumped region like the top of a mountain different from an acute peak) would have a length of 10 bp but what we observed in both primates and *C. elegans* is that the larger the bumps the more repetitions of the decamer in those regions. The different patterns of the MIF profile observed in *C. elegans* imply that nucleosomes do not favor a universal structure even among chromosomes of the same species.

3.2. MIF Profiles of Homo Sapiens. In Figure 3, the MIF profiles of the decamer YYYYYRRRRR, for each human chromosome, are illustrated. These profiles, equivalent to correlation functions, correspond to the distribution of spacings between the generic decamer suggested to be associated to the nucleosome positioning. In general, they show rugged landscapes with several troughs (these spacings are avoided) and peaks (these spacings are preferred). The MIF profiles

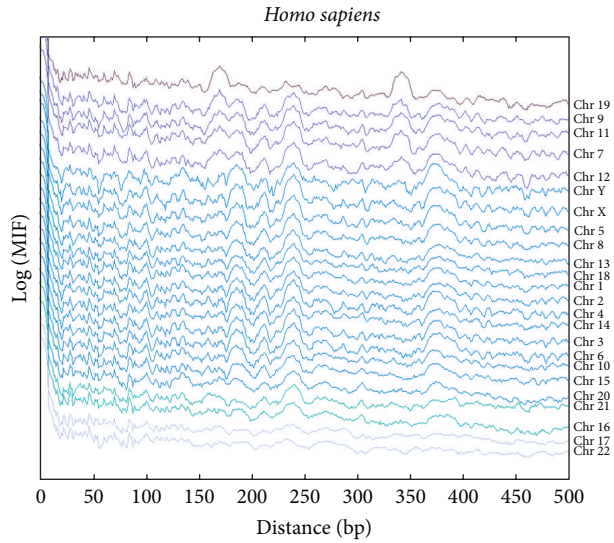


FIGURE 3: The MIF profiles of the decamer YYYYYRRRRR in all *Homo sapiens* chromosomes.

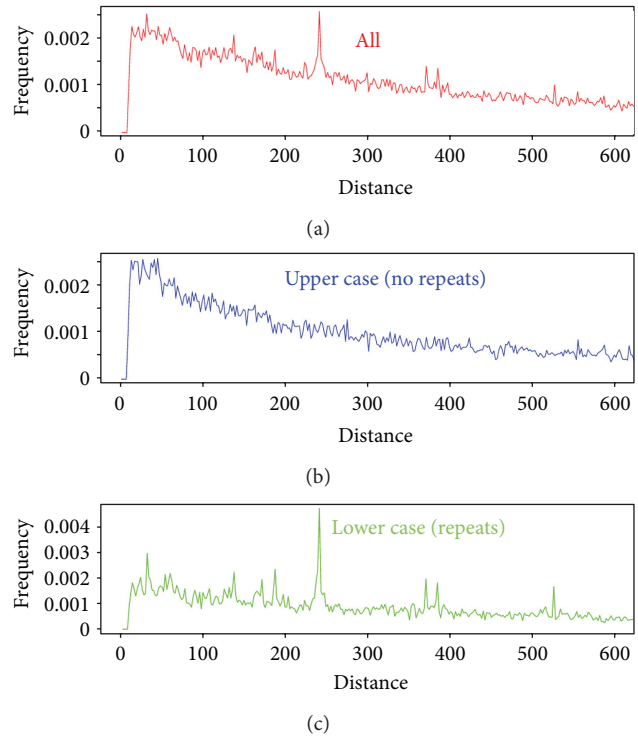


FIGURE 4: (a) Histogram of the spacing of the decamer YYYYYRRRRR in chromosome 21 of *Homo sapiens* (b) Histogram of the intact chromosome without repetitive elements; (c) Histogram of repetitive elements only.

of the decamers on the chromosomes were ordered in order to determine how similar (or different) they are among them. Each MIF profile was shifted upwards by successive integer multiples of 0.5 to facilitate visual inspection. At first glance, there seems to be 3 classes of profiles: class 1 comprises chromosomes 1 to 21 (except 17 and 19) and the sex chromosomes X and Y; class 2 includes chromosomes 17 and 22; and class 3 is represented by chromosome 19. Class 1 can still be subdivided into class 1a (chromosomes 1 to 21 excluding class 1b) and 1b (chromosomes 7, 9, 11, and 12), where the latter displays a bump at around 340 bp and two bumps in the range of 150 to 200 bp. It is widely recognized that the nucleosome has peaks at 80, 146, 165–167, and at around 240 bp [46].

A series of peaks up to −162 bp are clearly found in all chromosomes with the use of the MIF. At 10 bp, all chromosomes do not display a peak but they show a deviation in the falling trend. The observed periodicities occur at 31, 47, 62, 72, 84, 103, 110, 132, 136, and 162; bumps occur in regions 180–195, 225–255, and 365–395; long-range periodicities are found at 212, 240, 306, and 345.

Most chromosomes display a small peak at 165–167 bp, or 190 bp or 218 bp, which may reflect the periodic spacing between nucleosomes. With the exception of chromosomes 17, 19, and 22, all show a bump at around 240 bp due to repetitive elements as we will shortly illustrate (Figure 4). In addition to these peaks or hills, there are others like the ones found at 345 and 380, which might be considered as the spacing between next-nearest-neighbor nucleosomes.

This pattern from MIF is consistent with a direct measurement of histograms of the frequency distribution of spacing between the decamer along each chromosome except for the 10-base periodicity. The periodicities observed in the histogram occur at 10, 16, 20, 42, 55, 79, 93, 127, 146, 161, 178, 215, 230, 268, 287, 330, 360, 378, and 472 (not shown). Note that there is a great density of decamers at distances less than 500 bp, and at the same time there are specific peaks directly related to the more conspicuous ones of the MIF profile differences at these distances (Figure 3).

As close to 50% of the human genome consists of repetitive sequences, and we examine its contributions to the peaks seen in Figure 3. Figure 4 shows the histogram of spacing between the nearest decamer motifs for human chromosome 21 (Figure 4(a)), after masking all repetitive elements (Figure 4(b)), and finally, after masking all the sequence except repetitive elements (Figure 4(c)). A similar behavior is also seen in other human chromosomes (results not shown). Most peaks of intact chromosomes appear also in the histogram of repetitive sequences only. From Figure 4, it can be seen that the histogram of spacing of decamer YYYYYRRRRR for *H. sapiens* chromosome 21 shows peaks that appear in both the whole genome (Figure 4(a)) and in the only repetitive sequences (Figure 4(c)). In particular, this is true for the 240-241 peak. This means that in repetitive sequences there is a great deal of consecutive occurrences of the YYYYYRRRRR decamer spaced 240-241 bp apart. The biological meaning of this observation is still unknown. Due to the large proportion of repetitive sequences in the human genome, its potential function cannot be ignored. Our findings, as well as those in [46, 47], point to a potential role of repetitive elements in the nucleosome positioning. In Supplementary Information S1 available online at http://dx.doi.org/10.1155/2013/963956, we show a table of spacings between YYYYYRRRRR found at highly repetitive sequences in the human genomes.

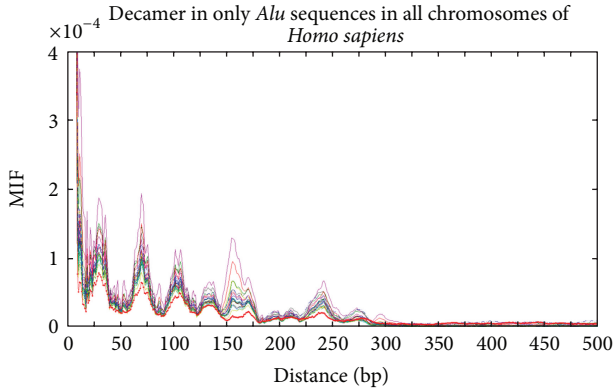


Figure 5: MIF profiles of the R/Y decamer in only *Alu* sequences in all *Homo sapiens* chromosomes.

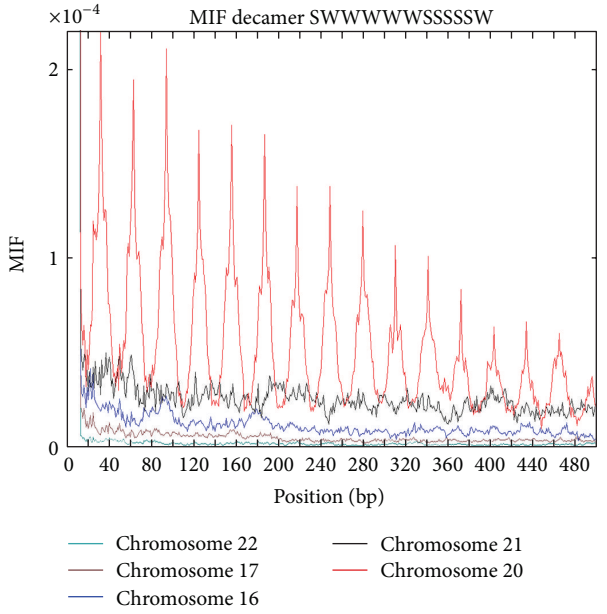


Figure 6: The MIF profiles of the 12-mer SWWWWWSSSSSW in some *Homo sapiens* chromosomes.

A possible relation between nucleosome positioning and one particular type of repetitive sequences, the *Alu* elements, has been suggested before [46]. It was observed that if one ignores the *Alu* repeats, several peaks in the Fourier spectra for AA/TT sequence (1 for AA or TT, 0 otherwise) disappear, but some peaks like the one found at about 165 bp still linger [46]. A similar observation was reported in [36]. Note that here we are analyzing a very different sequence (1 for YYYYYRRRRR, 0 otherwise), and repetitive sequences besides *Alu* are also masked. When only *Alu* sequences are considered, the MIF profiles of the decamer R/Y in all human chromosomes (Figure 5) display the same pattern in all chromosomes, indicating a strong association between the decamer Y/R (and R/Y) with *Alu* sequences. There are pronounced peaks at 32, 62, 110, 134, 160, and at 240 in all human chromosomes. There is a slight departure of this pattern in chromosome X (red curve).

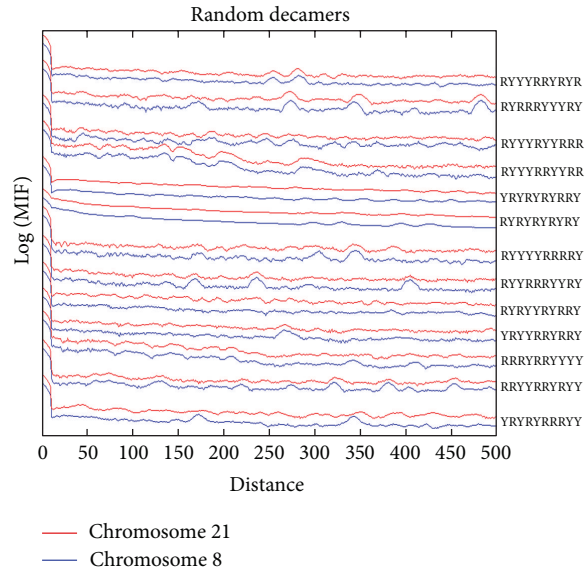


Figure 7: MIF profiles of random decamers with 5 purines and 5 pyrimidines along *Homo sapiens* chromosome 21, in order to compare the meaningful signal of YYYYYRRRRR as a binder nucleosome motif.

When the spacing of more specific decamers of the type YYYYYRRRRR (e.g., CGGAAATTTCCG) is analyzed, the periodic signal weakens considerably. The MIF profiles of the 12-mer SWWWWWSSSSSW in some chromosomes of *H. sapiens* are shown in Figure 6. Note that there is a regular behavior only in chromosome 20 in which there are peaks every 30 bp. A regular behavior is also observed in chromosome 12 whereas the remaining chromosomes exhibit a more irregular and nonuniform pattern.

If one selects at random a given decamer (preserving the number of Ys and Rs), not surprisingly, in most cases no prominent periodic signals are found as it is illustrated for the two chromosomes 21 and 8 of *H. sapiens* in Figure 7. The MIF profiles of the controls were not statistically similar among them (average correlation coefficient $r^2 = 0.56$) as the generic decamer in intact chromosomes do ($r^2 = 0.76$). The average correlation between the actual chromosome 8 with all random controls was $r^2 = 0.42$ whereas the average correlation between chromosome 21 with all random controls was $r^2 = 0.65$. Note that in the intact chromosomes we preserve the YYYYYRRRRR content and in the shuffled control we respect the nucleotide content but we disrupt the YYYYYRRRRR sequence. Therefore, the MIF profiles of the controls were not similar among chromosomes as the generic decamer do in intact DNA sequences.

3.3. Nonhuman Primate MIF Profiles. We also calculated the MIF profiles of the decamer on all available chromosomes of *Pan troglodytes* and *Macaca mulatta* (Figures 8 and 9). There is a consistency between MIF profiles of all chromosomes for each primate species, even though subtle differences exist. However, a more striking finding is that when two species are compared, a MIF profile for the decamer on a chromosome of a given species is more similar to that on the

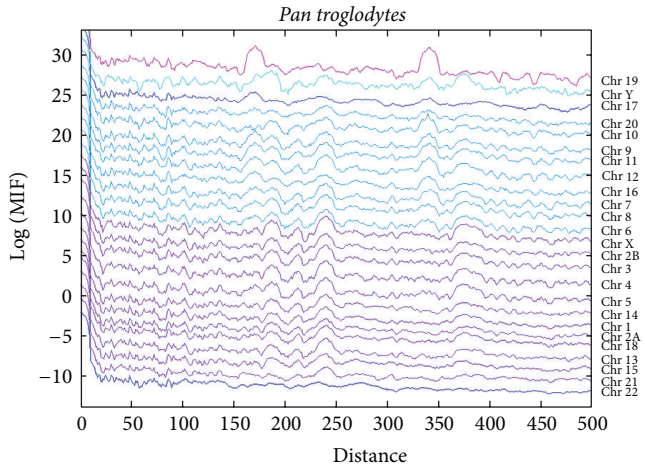


FIGURE 8: The MIF profiles of the decamer YYYYYRRRRR in all *Pan troglodytes* chromosomes.

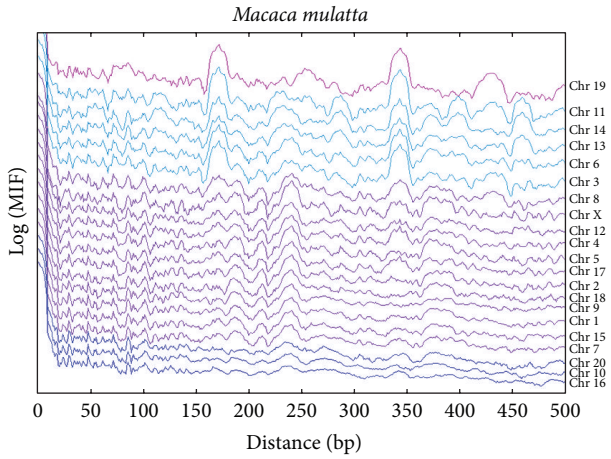


FIGURE 9: The MIF profiles of the decamer YYYYYRRRRR in all *Macaca mulatta* chromosomes.

same chromosome but in the other two species, than to the MIF profile on different chromosomes of the same species.

In general, it is clear that despite the different evolutionary histories of the 3 primate species, there is a common pattern in the MIF profiles of the generic decamer on a given chromosome.

For the same comparative purposes, the MIF profiles of the decamer on the chromosomes of *P. troglodytes* (Figure 8) can also be divided into the same three classes in which the *H. sapiens* chromosomes were divided. In the first class, there are chromosomes 1 to 21 and the sex chromosomes X and Y; in class 2, chromosomes 17 and 22 can even be subdivided given a conspicuous widening similar to a bump in the range of 150 to 175 that is present in chromosome 17. Class 3 is also represented by chromosome 19. But the first class can be subdivided into class 1a (chromosomes 1 to 21 excluding class 1b and 1c), class 1b with the same characterization that is, in human MIF profile (chromosomes 6 to 12, 16 and 22), and class 1c is represented by chromosome Y which has a different

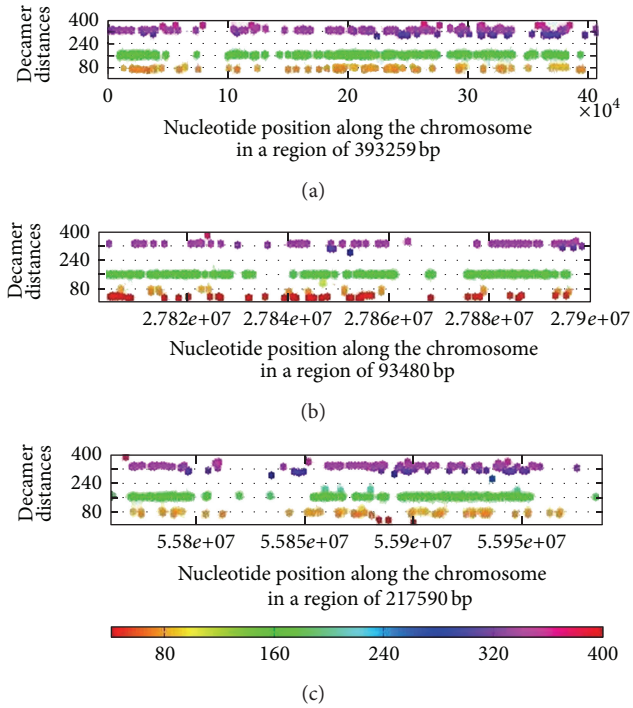


FIGURE 10: The three main regions of human chromosome 19, where distances around 80, 160, and 320 between the generic decamer are more highly concentrated than in the rest of the chromosomes are highlighted. Note that the clusters correspond to the peaks observed in their respective MIF profiles.

kind of bumps in the range of 64–84 bp and in the range of 164–200 bp (Figure 8).

The MIF profiles of the decamer on all chromosomes analyzed in *M. mulatta* (Figure 9) seem to pertain to only two classes. Class 1 can be subdivided by shorter amplitudes in the same bp signals between class 1a (chromosomes 10, 16, and 20) and class 1b (chromosomes 1, 2, 4, 5, 7 to 9, 12, 15, 17, and 18). With a similar profile than the two mentioned subclasses, class 1c presents highly conspicuous peaks at around 172 and 342 bp, and a bump in the range of 274 to 300 bp (chromosomes 3, 6, 11, 13, and 14). In class 1c, the chromosomes 11, 13 and 14 also have a peak at around 460 bp.

Class 2 is represented by chromosome 19 that, in contrast to chromosome 19 for *P. troglodytes* and *H. sapiens,* has a bump in the range of 400 to 450 bp which it shares only with chromosomes 3, 8, and 11, beside the features of its own profiles class (Figure 8). It is important to note that there are several common peaks among human, chimpanzee, and rhesus macaque at 31, 47, 62, 72, 84, 103, 110, 132, 136, and 162; even some bumps are shared among the three species at 180–195, 225–255, and 365–395 and some long-range periodicities at: 212, 240, 306, and 345.

It is important to mention that considering an alphabet of $A = \{A, T, G, C\}$, we calculated the MIF for the 3 species of primates masking all repeats (not shown) and all peaks disappear.

We estimated the similarities of the chromosomes within and between species based upon the cross-correlations of the

MIF profiles of the chromosomes for the 3 primate species. All Pearson's correlation coefficients within chromosomes of a given primate species display values which are in a rough agreement with the classes mentioned above (see correlations in S2). The Pearson correlation coefficients of chromosomes between the 3 primate species in general also support our visual inspection of the previous observations of heterogeneity between chromosomes within species, and uniformity among the same chromosome between species (see S2).

In general, it is clear that despite the subtle differences among chromosomes within species, there is a common pattern in the MIF profiles of the decamer. Given that this decamer is a consensus motif for nucleosome positioning sequence, the hypothesis that the statistical properties of the decamer can be translated to those of the nucleosome positioning can be put forward.

The distribution of the decamer YYYYYRRRRR along each chromosome is not uniform since there are regions in which clusters are crisply recognized whereas there are long stretches lacking this decamer (Figure 10). Since MIF and spacing histograms are averaged over all regions in a chromosome, Figures 3–10 do not show the heterogeneity information in regions deserted of this decamer. Therefore, other decamers or signals associated to the fine structure of the chromosomes cannot be ruled out.

To further examine the issue of heterogeneity, we show examples of physical maps of the location of the generic decamer along a given chromosome. In Figures 10 and 11, the location of the generic decamer along chromosome 19 of *H. sapiens* for different magnification scales is shown. The most striking observation is that the decamer positions are not random but they are not uniformly distributed along the chromosome either. The decamer distribution is clumped in certain regions but there are long stretches in which the decamer is plainly absent. For the remaining human chromosomes, nucleosomes are also more consistently positioned than expected by chance and many are organized in regularly spaced arrays that are enriched near active chromatin. Hence, nucleosome positions are also clearly influenced by DNA sequence. A striking example is an array of regularly spaced nucleosomes created by tandem repetition of sequences with strong nucleosome positioning properties across approximately 35,423 and 41,824 bp of chromosome 19 (Figures 10 and 11). Similar arrays can also be found in other chromosomes.

If we take a look at the distances between consecutive appearances of the decamer, there are regions of the chromosomes in which the decamer appear in a periodic manner. That is, there are two (or actually more) stretches of the chromosome which contain the same number of this decamer, but with the peculiarity that the first and second occurrences of the decamer are spaced by the same distance in both stretches, and so are the second and third, and so on (not shown). Another interesting feature is that there are arrangements of different distances in which the order of distances of the generic decamer can also be encountered in some downstream regions but exactly in reverse order

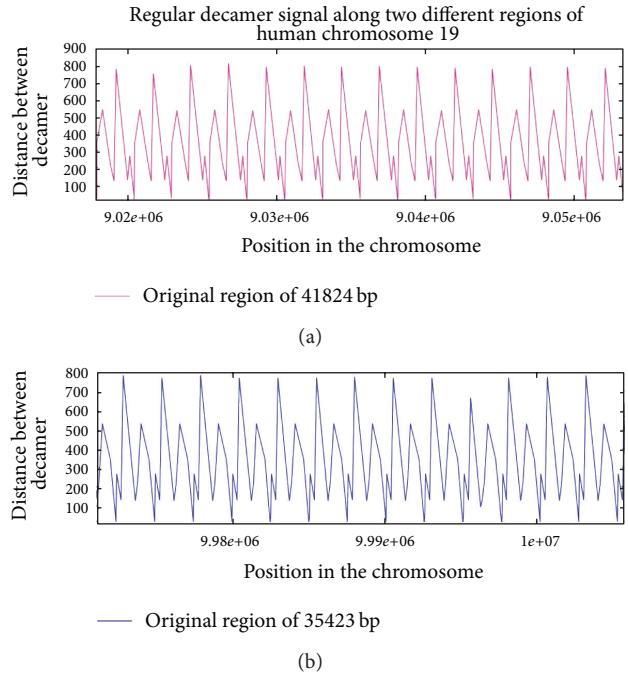


Figure 11: Plot of the distances between the generic decamer (ordinate) along two regions of chromosome 19 (abscissa) of *Homo sapiens*. Note the inverted repeat sequence.

of those distances. In other words, the distribution of the decamer exhibits an inverse symmetry (see Figure 11). It is worth mentioning that in this region there are genes of cadherin, beta-catenin and zinc fingers. This is consistent with a recent finding about rare roughly symmetrically positioned nucleosomes such as the zinc-finger containing protein that showed roughly symmetrically positioned nucleosomes [48].

3.4. MIF Profiles of Archaea and Eubacteria Species. We examine the MIF profiles of the decamer for the following Archaea species: *Methanocaldococcus jannaschii*, *Archaeoglobus fulgidus*, *Sulfolobus solfataricus*, and *Nanoarcheum equitans*.

In Figure 12, the MIF profiles of the decamer on several Archaea species are illustrated. It is remarkable to observe that this decamer still exhibits conspicuous periodicities. Similar to the MIF profiles observed in primates, in general, the MIF profiles of the decamer in archean species also manifest rugged landscapes with several troughs and peaks.

In *M. jannaschii,* there are several prominent peaks at around 67, 141, 210, and 408 bp whose magnitudes decrease with distance and they are interspersed throughout high-frequency oscillatory dynamics. The spacing of 141 apparently matches that of a nucleosome core sequence length and that of 67 close to half that length. The spacing of 210 could match the distance between two neighboring nucleosomes, and 375 for next-nearest-neighbor nucleosomes. As various linker sequence length may coexist in different regions in the genome, two nucleosome spacings (375/2 = 187.5 and 210) may not necessarily contradict each other.

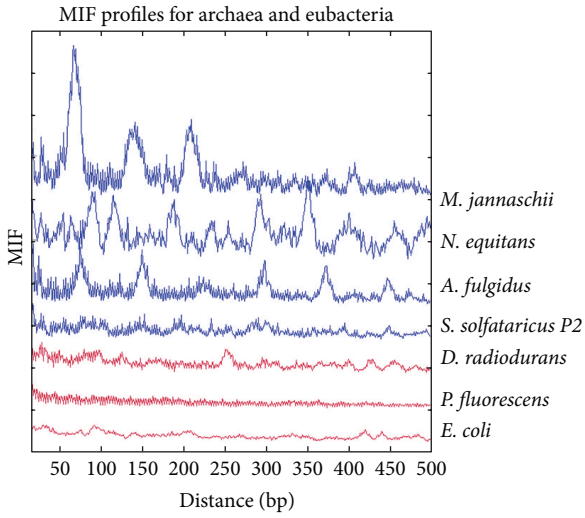


Figure 12: The MIF profiles of the decamer YYYYYRRRRR in some Archaea and Eubacterial genomes.

In *N. equitans,* there are several salient peaks at around 27, 89, 93, 115, 189, 234, 294, 352, 408, 456, and 496 bp with a great variability in their magnitudes, and they are embedded in a high-frequency oscillatory behavior.

Note that in *M. jannaschii* and *N. equitans* there are peaks at ~60 bp and ~85 bp as were found in pLITMUS28 and in *Methanothermus fervidus* [49]. Archaea nucleosomes resemble the structure formed by the $(H3 + H4)_2$ tetramer at the center of the eukaryotic nucleosome. Both structures have a histone tetramer core that recognizes positioning signals, directly contacts ~60 bp, and wraps ~85 bp of DNA alternatively in either a positive or negative toroidal supercoil [49].

In *A. fulgidus,* the MIF profile displays a pattern of high-frequency oscillatory structure that they themselves form jagged bumps, and there are salient peaks at around 75, 150, 300, 375, and 450.

In *S. solfataricus,* the MIF profile is essentially composed by high-frequency oscillatory structure from which jagged bumps are formed with no discernible prominent bumps.

In order to test whether the MIF profiles of Archaea are biologically meaningful, that is, the periodic appearance of the putative nucleosome positioning decamer is due to the repetitive motif within a nucleosome core and the regular spacing of nucleosomes, we also show the MIF of the decamer for several bacteria which are known to be lacking nucleosomes. Three of them (*Escherichia coli, Pseudomonas fluorescens,* and *Deinococcus radiodurans*) are shown in Figure 12. Note that in the corresponding MIF profiles of these three bacteria, the signal is so weak that we cannot ascribe, as expected, that there is a periodicity of the YYYYYRRRRR decamer along the genome. If the decamer is indeed associated with the nucleosome positioning sequence in any species, this is consistent with the absence of nucleosomes in bacterial genomes. We included these bacteria to test whether Archaea cells show evolutionary selection either for or against sequences that favor nucleosome formation. As bacteria do not possess histones, but do show 3 and 10-11 base periodicity due to coding regions, we presumed that *E. coli, P. fluorescence,* and *D. radiodurans* DNA sequences are evolutionarily neutral with respect to nucleosome formation, such that preferred nucleosome forming sequences will occur by chance. These results strongly argue that the Archaean genomes have evolved to favor nucleosome formation.

4. Discussion

In this work, we have found that the proposed nucleosome positioning motif YYYYYRRRRR exhibits expected periodicities in primates and Archaea, thus consistent with the hypothesis that it plays a role in nucleosome positioning. In particular, we placed emphasis on the effect of repetitive sequences on the observed periodicities of the motifs R/Y and Y/R, as well as the S/W motif. We succeeded in the detection of the periodical repetition of the DNA patterns in all chromosomes tested despite weak or previously undetected periodicities with other methods. The extraction of the periodical signals in all chromosomes was due to the fact of using both MIF profiles and the generic decamer R/Y and Y/R to document a comprehensive distribution of nucleosome DNA sequences in primate species and even perhaps in Archaea. The MIF profiles display peaks or bumps in places previously recognized, such as the typical signatures at 31-32, 84, 146, 157, 171 and 200 [25, 41, 46]. New periodicities such as 100, 167, 240, and 320 are reported here. We did find the 10-bp in the histograms (not shown) but not in the MIF profiles because it may be unlikely to detect it. The rationale is as follows: there are 10 million copies of YYYYYRRRRR/RRRRRYYYYY in the human genome and if we assume they do not overlap, this would lead to 90 million bases (when they do overlap, the number would still be smaller). For example, for a segment ...YYYYYRRRRRYYYYY... which contains 2 copies of the motif, it covers only 15 bases instead of covering 20 bases; 90 million bases represent 3% of the human genome (if overlap exists, could be 2%), but at least 20% of the human genome is well positioned with nucleosomes. Therefore, there are not enough 10 mers to cover densely within a nucleosome positioning region. This dense packing is what would lead to the periodicity of 10. On the other hand, we can have longer periodicities. Suppose we have this order: beginning-middle (dyad)-end-linker-beginning-next-nucleosome-.... Assume also that this motif tends to sit at the beginning of a nucleosome, then we do not need 20 copies per nucleosome to cover the whole region, only 1-2 copies per nucleosome at the beginning. This density is more consistent with our observations. Then, the regular spacing of nucleosomes would lead to longer (+200) periodicities, but not the 10-base periodicity within a nucleosome. When repetitive elements were masked in whole chromosomes it became evident that the decamer contributes not only to the presence of the nucleosome structure but it also manifests itself as part of highly repetitive sequences (see S1).

With more than one million copies, *Alu* elements are the most abundant repetitive elements in the human genome; they represent ~10% of the genome mass and belong to

the SINE (short interspersed elements) family of repetitive elements [50]. *Alu* elements emerged ~55 million years ago from a fusion of the 50 and 30 ends of a 7SL RNA gene, which encodes the RNA moiety of the signal recognition particle (SRP). Modern *Alu* elements are ~300 bp in length and are classified into subfamilies according to their relative ages [51]. Dimeric *Alu* elements are unique to primates. *Alu* RNAs, transcribed from *Alu* elements, are present in the cytosol of primate cells. *Alu* elements inherited the internal A and B boxes of the RNA polymerase III (Pol III) promoter from the 7SL RNA gene [52]. The typical *Alu* RNA is a dimer of related but nonequivalent arms that are joined by an A-rich linker and followed by a short poly(A) tail [52].

Not surprisingly, the MIF profiles of the shuffled decamers showed no discernible pattern and no rugged landscape. The MIF profiles of the controls were not similar among chromosomes as the generic decamer was. The MIF profiles of the generic decamer in the three primate species exhibited a uniformity between species for the same chromosome, but heterogeneity within species between different chromosomes. The observed regularity of the patterns allowed us to provide families for the distribution of the generic decamer tested.

We selected the three densest regions in which there were clearly clusters of the decamer which appeared every 80, 160, and 320 bp (multiples of 80) of human chromosome 19 (Figure 10). These clusters of the decamer clearly correspond to the peaks of the MIF profile of human chromosome 19 (Figure 3).

The finding of regular periodic patterns of the decamer along primate chromosomes visualized in distance series of long stretches of the different chromosomes, as well as the patterns reflecting inverted repeats (inverse symmetry), discards the possibility that the generic decamer is biologically meaningless. Periodicities naturally arise if the decamer is tandemly repeated, and/or if the nucleosomes are regularly spaced. Inverted symmetry can be caused by the central role of dyad in the nucleosome cores. We think that the probability of finding such arrangements just by chance would be very low. The patterns of the MIF profiles of the five chromosomes of *C. elegans* are not entirely consistent with the regular reported structure of their nucleosomes [41]. Therefore, most nucleosomes in primate genomes are consistently positioned, either because they are forced into positioned arrays by chromatin remodeling or DNA binding proteins, and/or because they adopt favored sequence positions in genomic regions without active binding. Interestingly enough, the MIF profiles of the generic decamer in all Archaea tested showed prominent peaks in an oscillatory background. We propose that this decamer deserves further studies in order to determine if it has been selected since the origin of nucleosome structure.

It has been noted that the RNA motif SRP9/14 binds primarily to the universally conserved core of the *Alu* RNA 59 domain, which forms a U-turn in the context of a tau-junction [53]. This RNA motif is highly conserved in the SRP RNAs from higher eukaryotes to yeast and from Archaea to some Gram-positive Eubacteria [54]. A dimeric *Alu* RNP complex might be important in the origin or propagation of tandemly arranged *Alu* retroposons, as retropositional success was clearly correlated with the emergence of dimeric *Alu* elements during primate evolution [55]. *Alu* elements play an important role in the regulation of gene expression at various levels, such as in alternative splicing when present in intronic regions of genes [56]. The observed MIF profiles from different chromosomes or different species often differ substantially. Therefore, all these patterns cannot be attributable to the origin of nucleosome structures, or nucleosomes sequence preferences. It is likely that many of the peak features may be ascribed to some species-specific or chromosome-specific DNA sequence features, such as *Alu* repeats, but not necessarily limited to them.

What then accounts for the phenotypic differences between nonhuman primates and humans? It stands to reason to propose that part of the difference might be because of species-specific alternative splicing.

We were able to characterize different classes of MIF profiles within each primate species. The outstanding observation is that the MIF profile of a given chromosome is more similar to the corresponding profile among species than within species. The observed peaks of the MIF profiles using the generic decamer in primates are strongly associated with several highly repetitive sequences. This is in agreement with the recent discovery that the positioning of neighboring nucleosomes seems to be in phase with *Alu* elements as reflected by peaks in the Fourier analysis at 84-bp and 167-bp [46]. In this work, we corroborate this result with both Fourier (not shown) and MIF analyses using the decamer.

We have also found that human repetitive sequence densities are mostly negatively correlated with R/Y-based nucleosome-positioning motifs (NPM) and positively correlated with W/S-based motifs [57]. The positive correlation between YYYYYRRRRR/RRRRRYYYYY and repetitive sequence density is intriguing, as it provides an exception to negative correlation between densities of repetitive sequences and that of R/Y-based NPMs. The scatter plot for YYYYYRRRRR/RRRRRYYYYY is particularly interesting; despite the negative trend followed by the majority of the points, there is a minority trend for high repetitive sequence densities and high NPM densities [57]. We believe that it is in this region in which the generic decamer can be found positively associated with *Alu* elements.

Herein, we focused on MIF profiles of the type R/Y and Y/R with several peaks that overlap with repetitive elements. Amongst the most prominent peaks for most chromosomes in primates are at 84, 100, 167, and 240. In fact, in certain chromosomal segments, a well-defined periodic pattern of the decamer within highly repetitive sequences was observed. Appearance of the CG dinucleotide in the nucleosome positioning pattern is rather surprising, considering its generally low occurrence in eukaryotic sequences. However, recent studies suggest that CG dinucleotides play a special role indeed [36]. First, it displays 10.4-base periodicity almost as often as the AA and TT dinucleotides do, in particular in G+C-rich regions [42, 58]. In the *Alu* sequences, the CG element appears at a distance of 31-32 bases from one another [59], suggesting involvement of the sequences in the nucleosomes. Methylation/demethylation of CpG would

modulate the nucleosome stability, so that the CG-containing nucleosome could be considered as "epigenetic nucleosomes" [59]. Most chromosomes, except 19, 22, X, and Y, show a notorious similarity in regard to the putative positioning of the nucleosomes as obtained with our approach. Even among species within the primate family, the latter still holds. Hence, the conserved MIF profile on primates can reflect the importance of these generic decamer into the architecture of primate genomes. In addition, the conserved and peculiar organization of islands into repetitive elements may allow us to consider that this specific decamer could be implicated in the self-regulation functions inherent in these types of sequences.

The finding of peaks in Archaea and its absence in Bacteria may not be a surprising result since it is known that the former contain histones whereas the latter do not. But it is noteworthy that we have detected for the first time via the MIF profiles putative nucleosome signals in Archaea. In addition, there is a prominent presence of the generic decamer in Archaea as it is shown in their corresponding histograms (not shown). To our knowledge, this is the first description of this generic decamer for the nucleosome in this group and it remains to prove that it may be considered a nucleosome without the subsequent evolutionary refinements conferred by the repetitive elements. Hence, repetitive elements turn out to be basic ingredients of the most fundamental structure of nucleosome positioning in higher Eukaryotes.

In summary, putative nucleosome positioning motifs (NPM) associated to repetitive elements in human, nonhuman primates, and Archaea have been identified by means of mutual information profiles (MIF). Trifonov's group suggested a most recent "finale motif" of the long-searched "chromatin code." The biological significance of this decamer motif and its two degenerate parental motifs is examined in primates and Archaea. Common features in the patterns of the generic decamer R/Y on MIF profiles among primate species are found. The distribution of R/Y motif exhibits previously unidentified periodicities, which are associated to highly repetitive sequences in the genome. *Alu* repetitive elements may contribute to the most fundamental structure of nucleosome positioning in higher Eukaryotes. In some regions of primate chromosomes, the distribution of the R/Y decamer shows symmetrical patterns including inverted repeats. We have detected for the first time via the MIF profiles putative nucleosome signals in Archaea. It is clear that the R/Y motif is relevant in the NPM but it is also certain that there must be other relevant motifs besides the Trifonov "finale." Our findings may contribute to the understanding of the origin of nucleosome structures in Archaea and its remarkable success of *Alu* retroposons in colonizing primate genomes.

Acknowledgments

Marco V. José was financially supported by PAPIIT UNAM, Project IN107112. They thank the Posgrado en Ciencias Biológicas UNAM and the Centro de Ciencias de la Complejidad, UNAM for the server computer support. Tzipe Govezensky offered assistance with S1.

References

[1] G. Felsenfeld and M. Groudine, "Controlling the double helix," *Nature*, vol. 421, no. 6921, pp. 448–453, 2003.

[2] A. Valouev, S. M. Johnson, S. D. Boyd, C. L. Smith, A. Z. Fire, and A. Sidow, "Determinants of nucleosome organization in primary human cells," *Nature*, vol. 474, no. 7352, pp. 516–522, 2011.

[3] D. E. Sterner and S. L. Berger, "Acetylation of histones and transcription-related factors," *Microbiology and Molecular Biology Reviews*, vol. 64, no. 2, pp. 435–459, 2000.

[4] K. Luger, T. J. Rechsteiner, A. J. Flaus, M. M. Y. Waye, and T. J. Richmond, "Characterization of nucleosome core particles containing histone proteins made in bacteria," *Journal of Molecular Biology*, vol. 272, no. 3, pp. 301–311, 1997.

[5] C. A. Davey, D. F. Sargent, K. Luger, A. W. Maeder, and T. J. Richmond, "Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution," *Journal of Molecular Biology*, vol. 319, no. 5, pp. 1097–1113, 2002.

[6] E. I. Campos and D. Reinberg, "Histones: annotating chromatin," *Annual Review of Genetics*, vol. 43, pp. 559–599, 2009.

[7] R. L. Redner, J. Wang, and J. M. Liu, "Chromatin remodeling and leukemia: new therapeutic paradigms," *Blood*, vol. 94, no. 2, pp. 417–428, 1999.

[8] T. J. Richmond and C. A. Davey, "The structure of DNA in the nucleosome core," *Nature*, vol. 423, no. 6936, pp. 145–150, 2003.

[9] K. Sandman, J. A. Krzycki, B. Dobrinski, B. Lurz, and J. N. Reeve, "HMf, a DNA-binding protein isolated from the hyperthermophilic archaeon *Methanothermus fervidus*, is most closely related to histones," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 15, pp. 5788–5791, 1990.

[10] R. L. Fahrner, D. Cascio, J. A. Lake, and A. Slesarev, "An ancestral nuclear protein assembly: crystal structure of the *Methanopyrus kandleri* histone," *Protein Science*, vol. 10, no. 10, pp. 2002–2007, 2001.

[11] I. Ioshikhes, A. Bolshoy, K. Derenshteyn, M. Borodovsky, and E. N. Trifonov, "Nucleosome RNA sequence pattern revealed by multiple alignment of experimentally mapped sequences," *Journal of Molecular Biology*, vol. 262, no. 2, pp. 129–139, 1996.

[12] S. C. Satchwell, H. R. Drew, and A. A. Travers, "Sequence periodicities in chicken nucleosome core DNA," *Journal of Molecular Biology*, vol. 191, no. 4, pp. 659–675, 1986.

[13] A. Valouev, J. Ichikawa, T. Tonthat et al., "A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning," *Genome Research*, vol. 18, no. 7, pp. 1051–1063, 2008.

[14] A. B. Cohanim, Y. Kashi, and E. N. Trifonov, "Yeast nucleosome DNA pattern: deconvolution from genome sequences of *S. cerevisiae*," *Journal of Biomolecular Structure and Dynamics*, vol. 22, no. 6, pp. 687–693, 2005.

[15] I. Ioshikhes, S. Hosid, and B. F. Pugh, "Variety of genomic DNA patterns for nucleosome positioning," *Genome Research*, vol. 21, no. 11, pp. 1863–1871, 2011.

[16] E. Segal, Y. Fondufe-Mittendorf, L. Chen et al., "A genomic code for nucleosome positioning," *Nature*, vol. 442, no. 7104, pp. 772–778, 2006.

[17] Y. Zhang, Z. Moqtaderi, B. P. Rattner et al., "Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions *in vivo*," *Nature Structural and Molecular Biology*, vol. 16, no. 8, pp. 847–852, 2009.

[18] Y. Zhang, Z. Moqtaderi, B. P. Rattner et al., "Evidence against a genomic code for nucleosome positioning," *Nature Structural and Molecular Biology*, vol. 17, no. 8, pp. 920–923, 2010.

[19] N. Kaplan, I. Moore, Y. Fondufe-Mittendorf et al., "Nucleosome sequence preferences influence *in vivo* nucleosome organization," *Nature Structural and Molecular Biology*, vol. 17, no. 8, pp. 918–920, 2010.

[20] A. Travers, "The nature of DNA sequence preferences for nucleosome positioning. Comment on 'Cracking the chromatin code: precise rule of nucleosome positioning' by Trifonov," *Physics of Life Reviews*, vol. 8, no. 1, pp. 53–55, 2011.

[21] E. N. Trifonov, "Cracking the chromatin code: precise rule of nucleosome positioning," *Physics of Life Reviews*, vol. 8, no. 1, pp. 39–50, 2011.

[22] K. Sha, S. G. Gu, L. C. Pantalena-Filho et al., "Distributed probing of chromatin structure *in vivo* reveals pervasive chromatin accessibility for expressed and non-expressed genes during tissue differentiation in *C. elegans*," *BMC Genomics*, vol. 11, no. 1, article 465, 2010.

[23] M. Yaniv and S. C. Elgin, "Chromosomes and expression mechanisms: bringing together the roles of DNA, RNA and proteins," *Current Opinion in Genetics and Development*, vol. 18, no. 2, pp. 107–108, 2008.

[24] L. E. Gracey, Z. Chen, J. M. Maniar et al., "An *in vitro*-identified high-affinity nucleosome-positioning signal is capable of transiently positioning a nucleosome *in vivo*," *Epigenetics and Chromatin*, vol. 3, no. 1, article 13, 2010.

[25] S. Sasaki, C. C. Mello, A. Shimada et al., "Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites," *Science*, vol. 323, no. 5912, pp. 401–404, 2009.

[26] D. Tillo, N. Kaplan, I. K. Moore et al., "High nucleosome occupancy is encoded at human regulatory sequences," *PLoS ONE*, vol. 5, no. 2, Article ID e9129, 2010.

[27] I. Ioshikhes, E. N. Trifonov, and M. Q. Zhang, "Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 6, pp. 2891–2895, 1999.

[28] R. Sadeh and C. D. Allis, "Genome-wide "re"-modeling of nucleosome positions," *Cell*, vol. 147, pp. 263–266, 2011.

[29] S. Henikoff, "Nucleosome destabilization in the epigenetic regulation of gene expression," *Nature Reviews Genetics*, vol. 9, no. 1, pp. 15–26, 2008.

[30] F. Moreno-Herrero, R. Seidel, S. M. Johnson, A. Fire, and N. H. Dekker, "Structural analysis of hyperperiodic DNA from *Caenorhabditis elegans*," *Nucleic Acids Research*, vol. 34, no. 10, pp. 3057–3066, 2006.

[31] H. R. Widlund, H. Cao, S. Simonsson et al., "Identification and characterization of genomic nucleosome-positioning sequences," *Journal of Molecular Biology*, vol. 267, no. 4, pp. 807–817, 1997.

[32] P. Oudet, M. Gross Bellard, and P. Chambon, "Electron microscopic and biochemical evidence that chromatin structure is a repeating unit," *Cell*, vol. 4, no. 4, pp. 281–300, 1975.

[33] W. Linxweiler and W. Hörz, "Reconstitution of mononucleosomes: characterization of distinct particles that differ in the position of the histone core," *Nucleic Acids Research*, vol. 12, no. 24, pp. 9395–9413, 1984.

[34] H. R. Drew and C. R. Calladine, "Sequence-specific positioning of core histones on an 860 base-pair DNA. Experiment and theory," *Journal of Molecular Biology*, vol. 195, no. 1, pp. 143–173, 1987.

[35] N. Kaplan, I. K. Moore, Y. Fondufe-Mittendorf et al., "The DNA-encoded nucleosome organization of a eukaryotic genome," *Nature*, vol. 458, no. 7236, pp. 362–366, 2009.

[36] E. N. Trifonov, "Nucleosome positioning by sequence, state of the art and apparent finale," *Journal of Biomolecular Structure and Dynamics*, vol. 27, no. 6, pp. 741–746, 2010.

[37] I. Gabdank, D. Barash, and E. N. Trifonov, "Open access article nucleosome DNA bendability matrix (*C. elegans*)," *Journal of Biomolecular Structure and Dynamics*, vol. 26, no. 4, pp. 403–412, 2009.

[38] S. M. Johnson, F. J. Tan, H. L. McCullough, D. P. Riordan, and A. Z. Fire, "Flexibility and constraint in the nucleosome core landscape of *Caenorhabditis elegans* chromatin," *Genome Research*, vol. 16, no. 12, pp. 1505–1516, 2006.

[39] S. G. Gu and A. Fire, "Partitioning the *C. elegans* genome by nucleosome modification, occupancy, and positioning," *Chromosoma*, vol. 119, no. 1, pp. 73–87, 2010.

[40] F. Salih, B. Salih, S. Kogan, and E. N. Trifonov, "Epigenetic nucleosomes: *Alu* sequences and CG as nucleosome positioning element," *Journal of Biomolecular Structure and Dynamics*, vol. 26, no. 1, pp. 9–15, 2008.

[41] F. Salih, B. Salih, and E. N. Trifonov, "Sequence structure of hidden 10.4-base repeat in the nucleosomes of *C. elegans*," *Journal of Biomolecular Structure and Dynamics*, vol. 26, no. 3, pp. 273–281, 2008.

[42] T. Bettecken and E. N. Trifonov, "Repertoires of the nucleosome-positioning dinucleotides," *PLoS ONE*, vol. 4, no. 11, Article ID e7654, 2009.

[43] S. L. Pereira and J. N. Reeve, "Histones and nucleosomes in Archaea and Eukarya: a comparative analysis," *Extremophiles*, vol. 2, no. 3, pp. 141–148, 1998.

[44] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[45] W. Li, "Mutual information functions versus correlation functions," *Journal of Statistical Physics*, vol. 60, no. 5-6, pp. 823–837, 1990.

[46] Y. Tanaka, R. Yamashita, Y. Suzuki, and K. Nakai, "Effects of *Alu* elements on global nucleosome positioning in the human genome," *BMC Genomics*, vol. 11, no. 1, article 309, 2010.

[47] D. Holste, I. Grosse, S. Beirer, P. Schieg, and H. Herzel, "Repeats and correlations in human DNA sequences," *Physical Review E*, vol. 67, no. 6, Article ID 061913, pp. 1–7, 2003.

[48] A. Kundaje, S. Kyriazopoulou-Panagiotopoulou, M. Libbrecht et al., "Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements," *Genome Research*, vol. 22, pp. 1735–1747, 2012.

[49] K. A. Bailey, F. Marc, K. Sandman, and J. N. Reeve, "Both DNA and histone fold sequences contribute to archaeal nucleosome stability," *The Journal of Biological Chemistry*, vol. 277, no. 11, pp. 9293–9301, 2002.

[50] E. S. Lander, L. M. Linton, and B. Birren, "International human genome sequencing consortium (2001) Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921.

[51] M. A. Batzer and P. L. Deininger, "Alu repeats and human genomic diversity," *Nature Reviews Genetics*, vol. 3, no. 5, pp. 370–379, 2002.

[52] J. Häsler and K. Strub, "*Alu* RNP and *Alu* RNA regulate translation initiation *in vitro*," *Nucleic Acids Research*, vol. 34, no. 8, pp. 2374–2385, 2006.

[53] O. Weichenrieder, K. Wild, K. Strub, and S. Cusack, "Structure and assembly of the *Alu* domain of the mammalian signal recognition particle," *Nature*, vol. 408, no. 6809, pp. 167–173, 2000.

[54] K. Strub, J. Moss, and P. Walter, "Binding sites of the 9- and 14-kilodalton heterodimeric protein subunit of the signal recognition particle (SRP) are contained exclusively in the Alu domain of SRP RNA and contain a sequence motif that is conserved in evolution," *Molecular and Cellular Biology*, vol. 11, no. 8, pp. 3949–3959, 1991.

[55] A. J. Mighell, A. F. Markham, and P. A. Robinson, "Alu sequences," *FEBS Letters*, vol. 417, no. 1, pp. 1–5, 1997.

[56] J. Kreahling and B. R. Graveley, "The origins and implications of *Aluternative* splicing," *Trends in Genetics*, vol. 20, no. 1, pp. 1–4, 2004.

[57] W. Li, D. Sosa, and M. V. José, "Human repetitive sequence densities are mostly negatively correlated with R/Y-based nucleosome-positioning motifs and positively correlated with W/S-based motifs," *Genomics*, vol. 101, no. 2, pp. 125–133, 2013.

[58] T. Bettecken, Z. M. Frenkel, and E. N. Trifonov, "Human nucleosomes: special role of CG dinucleotides and *Alu*-nucleosomes," *BMC Genomics*, vol. 12, article 273, 2011.

[59] M. S. Ong, T. J. Richmond, and C. A. Davey, "DNA stretching and extreme kinking in the nucleosome core," *Journal of Molecular Biology*, vol. 368, no. 4, pp. 1067–1074, 2007.

Hindawi Publishing Corporation
International Journal of Genomics
Volume 2013, Article ID 257218, 9 pages
http://dx.doi.org/10.1155/2013/257218

Research Article

Genome Microscale Heterogeneity among Wild Potatoes Revealed by Diversity Arrays Technology Marker Sequences

Alessandra Traini,[1] Massimo Iorizzo,[2] Harpartap Mann,[3] James M. Bradeen,[3] Domenico Carputo,[1] Luigi Frusciante,[1] and Maria Luisa Chiusano[1]

[1] *Department of Agricultural Sciences, University of Naples Federico II, Via Università 100, 80055 Portici, Naples, Italy*
[2] *Department of Horticulture, University of Wisconsin-Madison, 1575 Linden Drive, Madison, WI 53706, USA*
[3] *Department of Plant Pathology, University of Minnesota, 495 Borlaug Hall/1991 Upper Buford Circle, St. Paul, MN 55108, USA*

Correspondence should be addressed to Maria Luisa Chiusano; chiusano@unina.it

Received 7 November 2012; Revised 18 March 2013; Accepted 20 March 2013

Academic Editor: Ancha Baranova

Tuber-bearing potato species possess several genes that can be exploited to improve the genetic background of the cultivated potato *Solanum tuberosum*. Among them, *S. bulbocastanum* and *S. commersonii* are well known for their strong resistance to environmental stresses. However, scant information is available for these species in terms of genome organization, gene function, and regulatory networks. Consequently, genomic tools to assist breeding are meager, and efficient exploitation of these species has been limited so far. In this paper, we employed the reference genome sequences from cultivated potato and tomato and a collection of sequences of 1,423 potato Diversity Arrays Technology (DArT) markers that show polymorphic representation across the genomes of *S. bulbocastanum* and/or *S. commersonii* genotypes. Our results highlighted microscale genome sequence heterogeneity that may play a significant role in functional and structural divergence between related species. Our analytical approach provides knowledge of genome structural and sequence variability that could not be detected by transcriptome and proteome approaches.

1. Background

The subgenus *Potatoe* of the Solanaceae family includes approximately 188 tuber-bearing species [1]. They display large ecological adaptation encompassing several traits that are lacking in the commercial potato and useful for breeding [2]. Among wild potato species, *Solanum bulbocastanum* Dun. and *S. commersonii* Dun. ex Poir. have attracted the attention of researchers and breeders. *S. bulbocastanum* is a known source of resistance to late blight disease of potato, and four late blight resistance genes have been cloned from this species to date [3–7]. *S. commersonii* ranks first among *Solanums* in terms of cold tolerance and capacity to cold acclimate, and it is also a source of resistance to pathogens such as *Ralstonia solanacearum* and *Pectobacterium carotovorum* [8, 9]. *S. bulbocastanum* and *S. commersonii* are among approximately 20 diploid potato species classified as superseries *Stellata* by Hawkes [10]. Despite their importance as sources of genes for crop improvement, relatively few genetic and genomic resources are available for these species, and little is known on their genome organization, gene function, and regulatory networks. Recently, a Diversity Arrays Technology (DArT) array was constructed for potato [11]. The array contains markers derived from various *Solanum* species, including *S. bulbocastanum* and *S. commersonii*. DArT arrays offer the potential to simultaneously survey large numbers of anonymous loci distributed throughout the genome. DArT markers are highly transferrable across populations or even across species, since the DArT array comprises a structured marker set that is surveyed in each experiment. Importantly, polymorphic DArT markers correspond to a set of DNA clones that can be sequenced for downstream applications.

The availability of the potato DArT array together with the recent release of the complete genome sequences of cultivated potato [12] and tomato [13] provide an attractive opportunity for comparative genomic studies aimed at understanding genome evolution at the species level. The genomes of potato and tomato are largely syntenic, and molecular

markers and gene content are predominantly conserved [13–16]. This degree of similarity has already enabled cross species comparative genomics approaches for gene mapping and cloning, reviewed by Bradeen [17]. Bioinformatics platforms improve community access to these resources and related *omics* collections, playing an important role for data mining and genome integration [18, 19]. In contrast to this wealth of knowledge and resources for cultivated potato and tomato, very little is known about genome structure and gene content in the wild relatives of potato.

In this paper, we exploited the reference genome sequences of potato and tomato and a collection of sequences of potato DArT array markers that show polymorphic representation across the genomes of *S. bulbocastanum* and/or *S. commersonii* genotypes. Our aim was to define a preliminary collection of marker sequences informative for the two species as a starting point for investigation of genome structure. This collection was also useful to highlight microscale genome sequence heterogeneity that possibly plays a meaningful role in functional and structural divergence between related species.

2. Materials and Methods

2.1. Plant Materials and DArT Marker Analyses. Two genotypes of *Solanum bulbocastanum* and two genotypes of *Solanum commersonii* were analyzed in this study. *S. bulbocastanum* genotypes include PT29 (PI243510), a source of the late blight resistance gene *RB* [3], and G15 (PI255516), a source of the *RB* locus allele *RB-rc* [20]. The *S. commersonii* genotypes include the frost tolerant cmm1T (PI243503) [8] and cmm6-3 (PI590886), a seedling genotype selected based on its crossability with cmm1T [21]. Total genomic DNA of individual plants for molecular marker analysis was isolated from fully expanded leaves from greenhouse-grown plants, following the protocol of Doyle and Doyle [22], with minor modifications. Two grams of leaf tissue were frozen in liquid nitrogen and ground in a mortar and pestle. Ground tissue was suspended in 6 mL lysis buffer (100 mM Tris-HCl pH 8.0, 20 mM EDTA, 2% CTAB, and 1.4 M NaCl) and incubated for 20 min at 65°C with occasional mixing by inversion. One volume of chloroform was added, and the tubes were mixed well and incubated at room temperature for 20 min with occasional inversion. Tubes were then centrifuged for 15 min at 1000 g, and the supernatant was transferred to a separate tube containing 2 volumes of 100% ethanol. Contents were gently mixed by inversion. Precipitated DNA was hooked out using sterile micropipette tips and transferred to 1.5 mL microfuge tubes. The DNA was washed twice with 75% ethanol and resuspended in TE (Tris pH 8.0 + 1 mM EDTA) buffer. DNA was shipped to Diversity Arrays Technology Pty Ltd. (Canberra, Australia) for DArT marker analysis.

Construction of the potato DArT array has been previously described [11]. The potato DArT array contains markers derived from *Solanum* species representative of the secondary and tertiary genepools of potato. Hybridization of genome representations from *S. bulbocastanum* and *S. commersonii* genotypes to the potato array and automatic calling of marker states were performed by Diversity Arrays Technology Pty Ltd. using established protocols [23]. Data that passed quality standards were analyzed for polymorphisms between genotypes within each species, and polymorphic markers were selected for downstream analyses. Clone cultures corresponding to each of these markers were robotically arrayed into a Whatman EasyClone 384 well plate (Whatman plc, Kent, UK) by Diversity Arrays Technology Pty Ltd. following manufacturer's instructions. Briefly, 10 μL of each clone culture was applied to a well followed by air-drying of the plate. The FTA plates were then shipped to the University of Minnesota for PCR amplification and sequencing of clone inserts.

For clone insert PCR, 45 μL of 10 mM Tris pH 8.0 + 0.1 mM EDTA was applied to each FTA plate well for 10 min at room temperature. PCRs were conducted in a 50 μL volume that consisted of 1x PCR buffer (Applied Biosystems, Foster City, CA), 2.5 U of Amplitaq (Applied Biosystems), 200 μM of each dNTP, 1 μL of eluate from the FTA plates (as template), and 50 pmol of each primer (DArT-M13f: GTTTTCCCAGTCACGACGTTG and DArT-M13r: TGAGCGGATAACAATTTCACACAG; Integrated DNA Technologies (Coralville, IA)). Thermocycler (GeneAmp PCR System 2700 (Applied Biosystems)) conditions were 35 cycles of 94°C for 30 sec, 55°C for 30 sec, and 72°C for 30 sec followed by a single cycle of 75°C for 5 min. To each PCR, 5 μL of 3 M NaOAC and 125 μL of ice-cold ethanol were added. The PCR plates were stored at −20°C for at least one hour and then centrifuged at 2,500 g at 4°C for 30 min. The supernatant was gently poured off, and the open plates were centrifuged upside down at 800 g for 30 sec. To each tube, 175 μL of room temperature 70% ethanol was added. The plates were again stored at −20°C and centrifuged as described above. Plates were dried completely at 37°C before adding 20 μL of TE. Amplification was confirmed by agarose gel electrophoresis of 2 μL of each purified PCR, staining with ethidium bromide, and visualization under UV light.

DNA sequencing of inserts was completed at the University of Minnesota BioMedical Genomics Center using BigDye Terminator (Applied Biosystems) cycle sequencing on an Applied Biosystems 3100 or 3700 automatic sequencer. Each sequencing reaction contained 1 μL of purified PCR product and 3.2 pmol of DArT-M13f or DArT-M13r. Each insert was sequenced in both directions in separate reactions. Resulting sequences were trimmed of vector and assembled into consensus sequences using SeqMan, part of the DNASTAR (Madison, WI) Lasergene software package.

Out of 1,423 DArT marker clones sequenced, 756 hybridized in a polymorphic fashion with *S. bulbocastanum* genotypes and 550 hybridized in a polymorphic fashion with *S. commersonii* genotypes. Hereafter, these markers will be referred to as BLB- and CMM-specific markers, respectively. The remaining 117 DArT markers hybridized and were polymorphic in both species (indicated as BLB/CMM).

2.2. Sequence Analysis and Data Interpretation. The genome sequence of *Solanum phureja* [12] served as the reference genome for our analyses. The genome sequence of *Solanum lycopersicum* [13], another reference species among Solanaceae, was also employed. For both genomes, our analyses

included 12 pseudomolecule sequences as well as unanchored scaffolds. We adopted gene annotations reported by the iTAG group (international Tomato Annotation Group) [24], assuring uniform annotation criteria and bioinformatics strategies and allowing coherent comparisons of the two reference genomes herein considered [13].

DArT marker sequences were aligned to the genome sequences using the splicing alignment software GenomeThreader [25] with 70% minimal nucleotide coverage and sequence identity. DArT alignments to genome sequences were grouped into six different categories (Figure 1(a)). A DArT marker sequence that aligned to a genome region independent of other DArT markers (i.e., one that does not overlap with any other marker sequences in the same genomic region) was classified as *solitary*. Each *solitary* marker was further subclassified as (1) *solitary one match*, if it aligned only once to the genome, or (2) *solitary multiple matches*, if it aligned more than once. A DArT marker whose alignment to the genome overlapped that of other DArT marker sequences was classified as an *overlapping* DArT. A DArT marker sequence having multiple matches to the genome, some of which are *solitary* and some of which are *overlapping*, was classified as subcategory (3) *mixed*. Other *overlapping* markers were further classified as *overlapping in uniform groups* when the group was composed of the same set of overlapping DArT marker sequences. This category comprised two subcategories: (4) *overlapping in uniform groups—one match* occurring only once in the genome and (5) *overlapping in uniform groups—multiple matches* appearing in two or more genome locations. DArT marker sequences which show multiple matches to the genome sequence and overlap sets of different DArT markers are defined as (6) *overlapping in heterogeneous groups*.

Fifty-three DArT marker sequences that did not align to either the potato or tomato genome sequences based on the GenomeThreader approach were assembled using CAP3 [26] (parameters: -p 40 –o 80) before a second alignment attempt based on BLASTn [27] (parameters: -e 0.003). These same DArT sequences were also aligned to the GenBank nucleotide collection (nr/nt) using BLASTn and to the nonredundant protein sequences dataset using BLASTp [28]. A BLAST2GO analysis [29, 30] was performed to classify genes associated to DArT marker sequences to show the cellular, biological, and molecular functional information of the subset annotation.

3. Results and Discussion

3.1. Dataset Description. The majority of the 1,423 DArT sequences analyzed have a length ranging between 350 and 850 nucleotides, providing a consistent dataset for subsequent bioinformatics analyses. In particular, 68% of BLB markers and 73% of CMM markers are 450 to 700 nucleotides in length (data not shown).

About 92% and 79% of all DArT sequences could be aligned the potato and tomato genomes, respectively (Table 1). These comprise 93% of BLB, 91% of CMM, and 90% of BLB/CMM DArT markers relative to the potato genome and 78% of BLB, 81% of CMM, and 76% of BLB/CMM DArT markers relative to the tomato genome. The discrepancy between the percentage of alignments to each genome is consistent with the composition of the reference potato DArT array that emphasizes markers from *Solanum* species more closely related to potato [11].

Table 1: Results of DArT alignments to potato and tomato reference genomes. For each collection, the total number of DArT markers and the number (%) of aligned DArT markers are reported.

Collection	Total no. of DArT	No. aligned (%) to	
		Potato	Tomato
BLB	756	703 (92.9)	586 (77.5)
CMM	550	499 (90.7)	446 (81.1)
BLB/CMM	117	105 (89.7)	89 (76.1)
All	1423	1307 (91.8)	1121 (79.0)

Sequence alignments were grouped into six categories, as described in Section 2. In the alignments to both potato and tomato genomes, DArT markers most frequently occurred as group (1) *solitary one match*, with 344 and 321 matches for potato and tomato, respectively, and as group (4) *overlapping in uniform groups-one match*, with 755 matches for potato and 663 matches for tomato (Figure 1(a)). For alignments to the potato genome, these two categories encompass 84% of all sequenced DArT markers: 82% for BLB, 87% for CMM, and 88% for BLB/CMM (Figure 1(b)). For alignments to the tomato genome, these same categories comprise 88% of all DArT marker sequences: 84% for BLB, 91% for CMM, and 91% for BLB/CMM. The remaining four marker alignment categories each represent less than 10% of the total number of aligned DArT marker sequences (Figure 1(b)). Briefly, groups (2) *solitary multiple-matches* and (5) *overlapping in uniform groups-multiple matches* show alignment to more than one genome region; this is probably due to repeated regions in the genome sequence; therefore, we considered these markers to be redundant. Groups (3) *mixed* and (6) *overlapping in heterogeneous groups* comprise DArT sequences with different alignment configurations probably due to intrinsic sequence properties. DArT marker sequences assigned to categories (1) and (4) localize in unique regions in both the potato and tomato genomes. Since these markers are associated unambiguously to specific genome locations, they were considered as nonredundant markers and were subjected to further analyses; DArT markers not assigned to alignment categories (1) and (4) were not considered further.

3.2. Analysis of Nonredundant DArT Markers. In total 1,099 and 984 nonredundant (i.e., group (1) and group (4)) DArT marker sequences align to the potato and tomato genome sequences, respectively. The majority of the marker sequences aligns with a sequence identity exceeding 80% and a coverage greater than 90% (Figure 2). The percentage of alignments in the highest coverage category (between 90 and 100%) is 92% for potato and 75% for tomato. Many of the alignments overlap gene regions in both genomes (Figure 2). This is not unexpected since DArT markers are obtained through digestion by *Pst*I. *Pst*I is a methylation-sensitive enzyme; therefore, it is possible that it acts mainly on hypomethylated

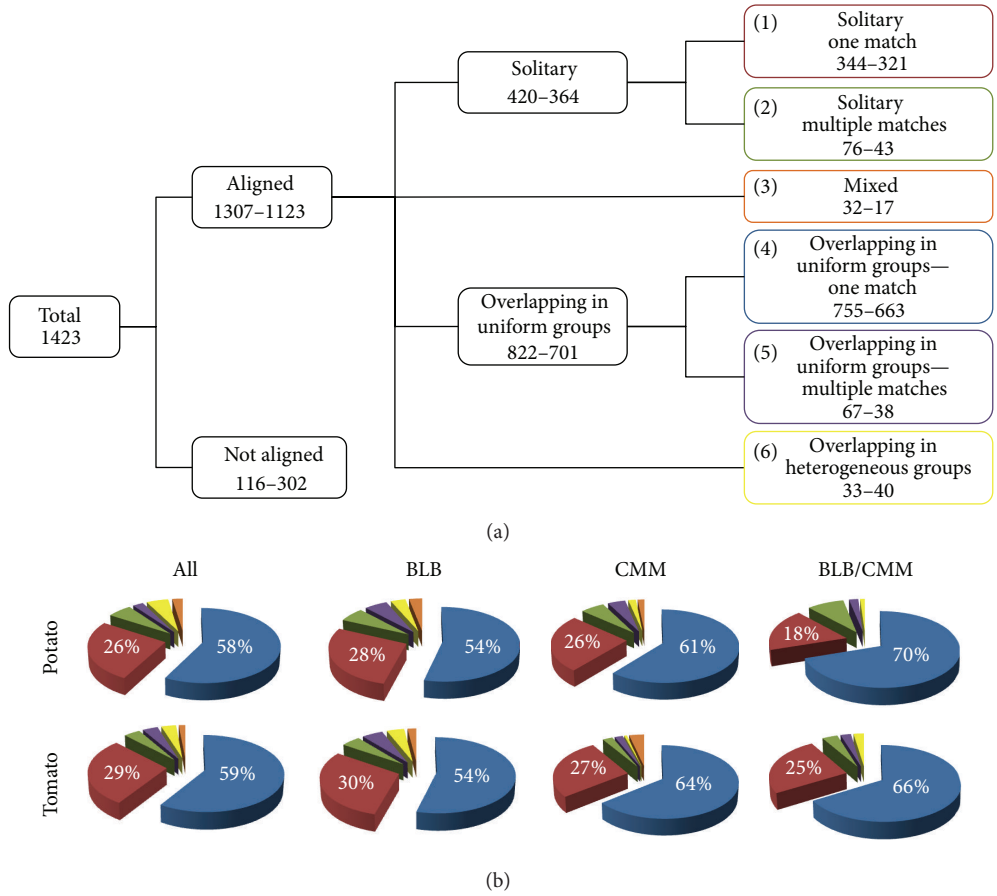


FIGURE 1: Categories of DArT markers alignments. (a) Values represent the number of alignments along the potato and tomato genome, respectively. (b) Pie charts of the percentage of aligned DArT markers, for each collection. The colour code is associated to the coloured rectangles of (a) and percentages are reported only when greater than 10%.

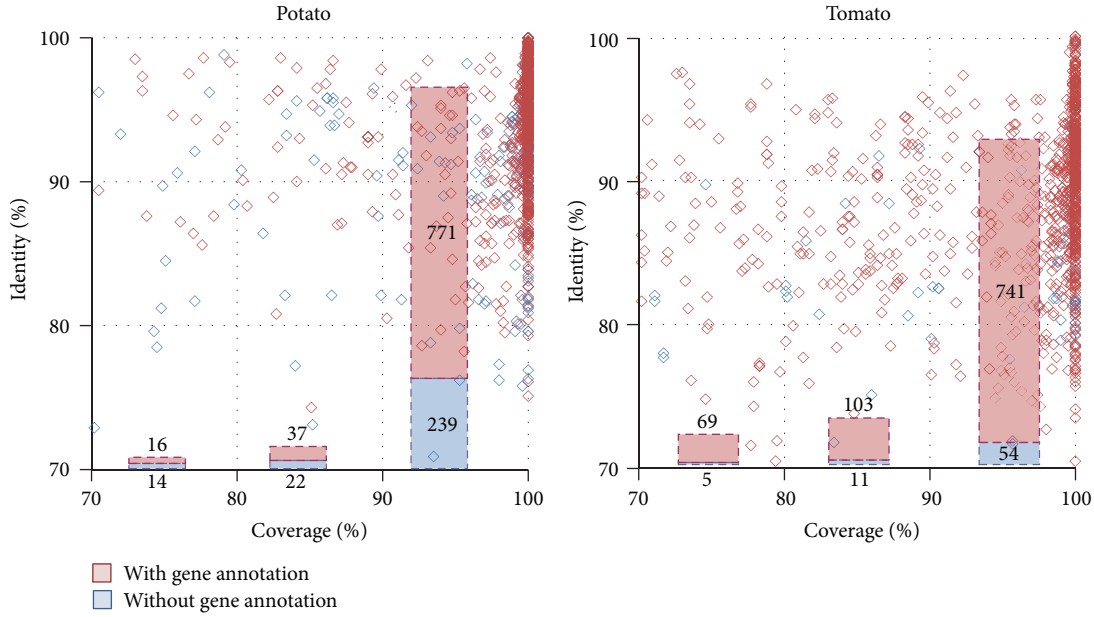


FIGURE 2: DArT marker sequences align predominantly with gene coding regions of the potato and tomato genome. The alignments associated (or not) to a gene locus along the potato and tomato genomes are highlighted in red (or blue). For each group, the number of alignments is also given.

DNA which, in turn, may correspond to gene regions, which are typically hypomethylated [31]. In Figure 3, the BLAST2GO analyses of the genes overlapping DArT marker regions are shown for both potato and tomato annotations. In particular, the figure shows the overrepresentation of genes associated with catalytic and binding activities.

In percentage, the two marker groups (1 and 4) represent 84% and 88% of all markers sequences aligned to the potato and the tomato genomes, respectively. Interestingly, in contrast with average results across all DArT sequences (Table 1) showing more matching DArT sequences to potato than to tomato, a higher proportion of the nonredundant groups align to the tomato genome than to the potato genome. This may be due to the higher contribution of ambiguous alignments (group (2) and (5)) in potato. This in turn suggests a higher sequence repetitiveness in the potato genome or better sequence quality for the tomato genome [12, 13]. Overall, nonredundant DArT marker sequences show very high coverage in potato compared to tomato (Figure 2), confirming higher phylogenetic similarity amongst potato species.

We next examined total coverage of the genome sequences from cultivated potato and tomato represented by alignments with DArT marker sequences (Table 2). Details per chromosomes are reported in the supplementary Table S1 (see Table S1 in Supplementary Material available at http://dx.doi.org/10.1155/2013/257218). In general, BLB DArT markers encompass a greater number of nucleotides in each genome than CMM or BLB/CMM markers. This is not surprising since BLB markers are the largest subset of DArT markers examined in this study. BLB DArT markers represent 208.8 Kbp of the potato genome but only 175.8 Kbp of the tomato genome. In contrast, CMM and BLB/CMM markers represent approximately equivalent regions of the potato and tomato genomes (CMM: 137.9 Kbp for potato versus 139.6 Kbp for tomato; BLB/CMM: 29.4 Kbp for potato versus 24.7 Kbp for tomato). We further divided the nonredundant DArT markers into two subclasses. *Common* markers align with the genome sequences of both potato and tomato; *specific* markers align to only one of the two genomes (Table 2). Within each sub-class, alignments were either *ungapped* (i.e., marker sequences aligned to genome sequences without disruption) or *gapped* (i.e., marker sequences aligned to genome sequences but alignments were interrupted by genome sequence not found in marker sequences). It is noteworthy that the same DArT marker sequence could be *ungapped* when aligned to the potato genome and *gapped* when aligned to the tomato genome or *vice versa*. The relative ratio of *gapped* versus *ungapped* regions of all BLB, CMM, and CMM-BLB DArT marker sequences relative to the potato and tomato genome sequences provides insight into patterns of genome evolution and species relationships. Distinction between *gapped* and *ungapped* alignments is necessary since variability in the length of *gapped* markers can complicate interpretation of the degree of genome coverage by the marker sequences. In potato, for example, the size of most of the gaps (89%) ranges from 20 to ~1000 bps. The remaining ones reach a maximum at ~5000 bps (not shown). For *common* DArT markers, the contribution of *ungapped* regions to total genome representation is higher in potato than in tomato for each marker collection. In contrast, for *common* markers, the contribution of *gapped* regions is generally lower in potato than in tomato. This again reflects higher phylogenetic similarity of the wild species to the cultivated potato. However, it is interesting to note that the relative frequency of *common gapped* regions compared to *common ungapped* ones in potato versus tomato is comparable for both BLB (14.21% in potato and 16.68% in tomato) and BLB/CMM (5.14% potato and 3.16% in tomato) DArT markers. The frequency of CMM *common gapped* and *ungapped* regions differs in potato (7.73%) with respect to tomato (15.64%). This indicates that, in contrast to BLB markers, CMM markers align with fewer gaps to the potato genome sequence than to the tomato genome sequence. This implies that the genomes of *S. commersonii* and potato are more similar at a DNA sequence level than are the genomes of *S. bulbocastanum* and potato, consistent with *S. commersonii* being phylogenetically more closely related to potato than is *S. bulbocastanum*, as the analyses based on plastid genomes previously suggested [32–34].

Considering the contribution of *specific* DArT markers, *ungapped* BLB markers provided the greatest overall genome coverage for both potato and tomato, consistent with higher representation of BLB markers in our dataset (Table 2). Importantly, the relative proportion of *gapped* regions compared to *ungapped* regions for the *specific* alignments indicates a comparable behaviour in the three marker collections in both species.

3.3. Genome Sequence Heterogeneity. We compared marker origins and alignment classifications across the potato and tomato genomes (Table 3). In general, the majority of aligned DArT markers are *ungapped* in both potato and tomato: 328 (77%) for BLB, 297 (83%) for CMM, and 65 (91%) for BLB/CMM. Eight BLB and 16 CMM markers align to both genomes in a *gapped* configuration (Table 3). Interestingly, a high percentage of aligned markers exhibit heterogeneous behaviours across the potato and tomato genomes (i.e., *gapped* versus *ungapped* in potato versus tomato and vice versa). These sequences are a source of marker variability between wild and cultivated species that can be exploited in future studies.

Seven BLB, 16 CMM, and one BLB/CMM markers aligned to the genomes of both potato and tomato in a *gapped* configuration (Table 3). As shown in Table S2, each of the seven BLB DArT markers aligned to gene regions in both species. Among these, five regions corresponded to genes with identical annotations in potato and tomato. On the other hand, among the 16 CMM DArT markers, only 10 and 14 aligned to gene coding regions in potato and tomato, respectively. Of the 10 CMM markers aligning to both potato and tomato gene coding regions, all of the 10 aligned to regions with identical gene annotations in both species (Table S2).

Nine DArT markers, three from BLB and six from CMM, aligned with the same alignment structure (i.e., number and length of *gapped* and *ungapped* regions) to homologous chromosomes in both potato and tomato and to gene loci with the same annotation (Table S2). The remaining two BLB, four

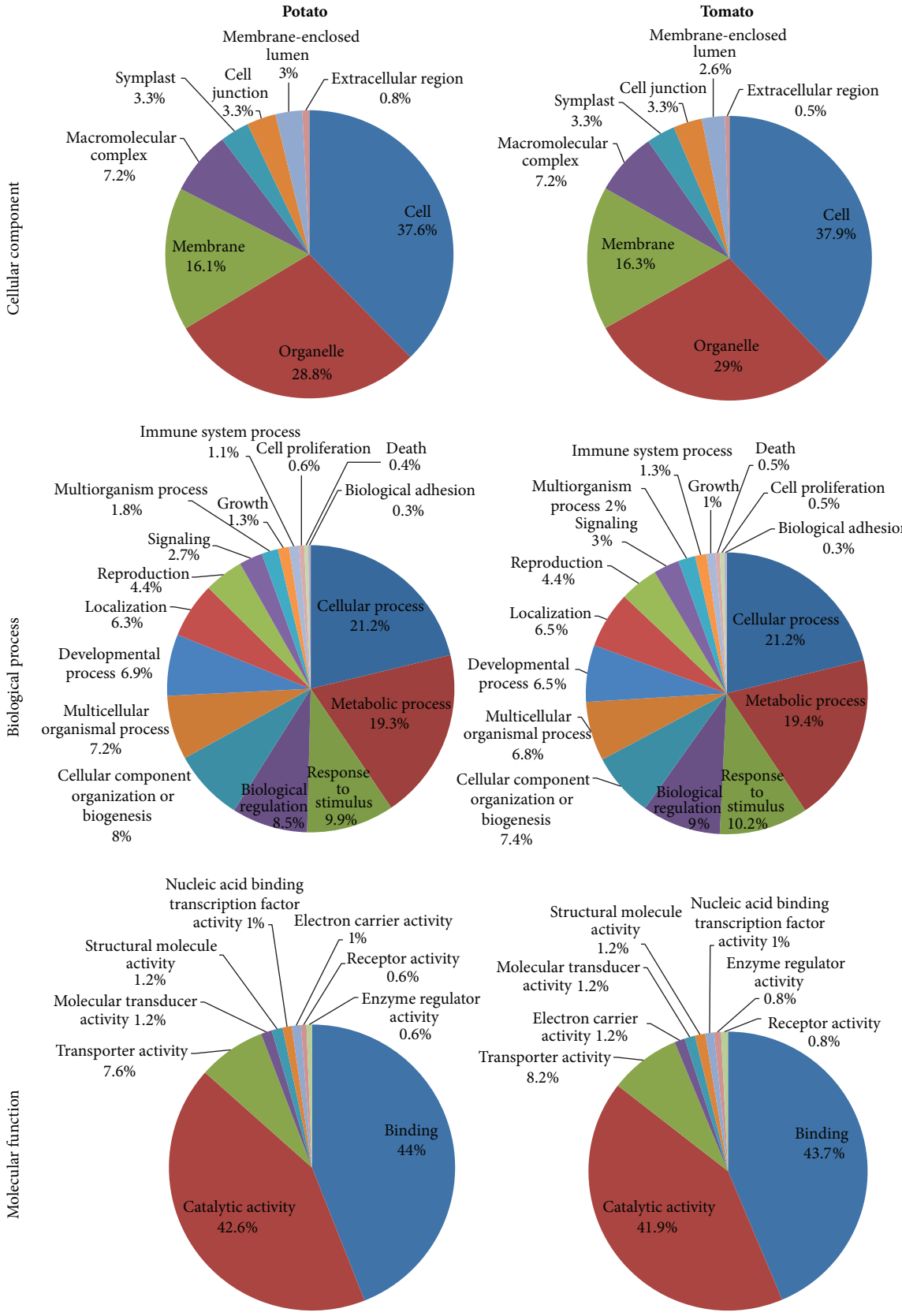


Figure 3: BLAST2GO analyses of the genes overlapping DArT marker regions.

Table 2: Number of nucleotides (in Kbp units) covered by DArT alignments. For details on coverage categories, see Section 2.

Coverage category	Potato			Tomato		
	BLB	CMM	BLB/CMM	BLB	CMM	BLB/CMM
Common						
Ungapped	132.3	97.7	20.5	128.2	95.3	19.1
Gapped	18.8	7.6	1.1	21.4	14.9	0.6
Specific						
Ungapped	46.1	22.3	7.7	22.5	17.3	4.4
Gapped	10.2	10.4	0.2	3.7	12.1	0.7
Total	208.8	137.9	29.4	175.8	139.6	24.7

Table 3: Comparison between DArT alignments to potato and tomato genomes. Number of DArT markers aligned along the potato (horizontal) and tomato (vertical) genomes for each collection, given in parenthesis. Each cell, within each matrix, shows the number of DArT markers per alignment type: ungapped, gapped, or not aligned.

		Potato											
		BLB (573)			CMM (432)			BLB/CMM (94)					
		Ungapped	Gapped	Not aligned	Ungapped	Gapped	Not aligned	Ungapped	Gapped	Not aligned			
Tomato	BLB (494)	Ungapped	328	43	61	CMM (407)	Ungapped	296	14	47	BLB/CMM (82)	Ungapped	65	3	9
		Gapped	50	7	5		Gapped	29	16	5		Gapped	3	1	1
		Not aligned	115	30			Not aligned	63	14			Not aligned	21	1	

CMM, and one BLB/CMM markers, although aligning to homologous chromosomes in genes of the same annotation, showed heterogeneous (i.e., number and length of *gapped* and *ungapped* regions) alignment structure (Table S2). These observations of microscale genome heterogeneity may be relevant to investigation of genome structures, functionalities, and properties of the represented *Solanum* species.

3.4. DArT Marker Sequences Not Aligned to the Reference Genomes. Some DArT markers could not be aligned to one or both genome sequences (Table 1). In particular, 116 marker sequences could not be aligned to the potato genome, 302 marker sequences could not be aligned to the tomato genome, and 51 marker sequences could be aligned to neither to potato nor to tomato. These were selected as putative wild species-specific markers and were assembled using the CAP3 software, yielding seven assembled consensus sequences comprising 20 sequences in total. The remaining 31 DArT marker sequences could not be assembled. Next, we attempted a less stringent alignment of the resulting 38 sequences (31 unassembled sequences plus seven consensus sequences) to the potato and the tomato genome sequences using the BLASTn algorithm (Table S3). Using this approach, 18 DArT marker sequences could be assigned to single locations in both the potato and tomato genomes, and only nine markers aligned to multiple genome locations in one or both species. In these cases, the less stringent alignment search performed by the BLAST software helped to confirm the presence in the potato and tomato genomes of 27 DArT marker sequences, previously unidentified in the more stringent GenomeThreader analysis. Moreover, in some cases, the BLASTn analysis confirmed matches to the same chromosome for both potato and tomato (e.g., DArT markers 472847 (chromosome 1), 537586 (chromosome 8), 473780 (chromosome 2), and 534573 (chromosome 11)). The presence of low level sequence similarity between these markers and the potato or tomato genome sequences revealed distant relationships between the wild and cultivated species and may be exploited in the study of cross-species genome heterogeneity. Twenty-two DArT markers (Table S3) showed extreme repetitive distribution along the potato and tomato chromosomes and were described by ambiguous annotations. Nevertheless, protein-based annotations (BLASTp), when present, generally confirmed homology with *Solanum* proteins or with those from more distantly related plant species. Two DArT marker sequences failed to align to the genomes of either potato or tomato even under more permissive analytical criteria.

4. Conclusions

Potato (*S. tuberosum*) and tomato (*S. lycopersicum*) belong to the subgenus *Potatoe* of the large and diverse genus *Solanum*. Although horticulturally distinct, potato and tomato share a clear evolutionary history that is well supported by molecular

data [35, 36]. The species are thought to have diverged from a common ancestor approximately 6.2 to 7.3 million years ago [37, 38]. Sexual isolation and subsequent divergence of the two species were accompanied by a series of structural genomic changes including chromosome arm inversions and large-scale translocations [14, 15]. Nevertheless, the genomes of potato and tomato are largely syntenic and molecular marker and gene content are predominantly conserved [14–16]. This degree of similarity has enabled cross species comparative genomics approaches for gene mapping and cloning, reviewed by Bradeen [17], efforts that will likely be furthered by the recent release of the complete genome sequences of potato [12] and tomato [13].

In this study, we proposed a suitable methodology to exploit partial genome information from wild species in the presence of reference genomes from related species. This approach, here exploited with DArT marker sequences, can also be employed in partial genome resequencing or similar efforts. Our results also highlighted the presence of divergent sequence relationships and heterogeneous alignment structures, including the presence/absence of gaps, which are detectable thanks to appropriate, less stringent comparative methods. This divergence commonly occurred even in gene pairs with apparent orthologous relationships and presumed functional conservation, and it could often be confirmed both in potato and tomato genomes. Evidence from results supported by two reference-related species partially overcomes possible limits that may be due to the quality of first released genomes and suggests a fine microscale genome structural divergence between wild and cultivated species in the Solanaceae. Our results confirm the utility of suitable analytical approaches that could be applied when partial genome information is available, capable of highlighting genome microscale variability that, although often occurring at the gene level, is not detectable when investigating genome functionality at transcriptome and proteomic levels.

Conflict of Interests

The authors declare no conflict of interests.

Acknowledgments

This research was carried out within the project "Approcci "-omici" integrati per lo studio e l'utilizzazione della biodiversità di patata" funded by the Italian Ministry of Agriculture. It was supported in part by USDA/NIFA through an Agriculture and Food Research Initiative (AFRI) grant to JMB and by the GENHORT project funded by MIUR. The support of the Minnesota Supercomputing Institute at the University of Minnesota is gratefully acknowledged.

References

[1] D. M. Spooner and A. Salas, "Structure, biosystematics, and genetic resources," in *Handbook of Potato Production, Improvement, and Post-Harvest Management*, J. Gopal and S. M. Paul Khurana, Eds., pp. 1–39, Haworth's Press, Binghampton, NY, USA, 2006.

[2] J. E. Bradshaw, "Potato-breeding strategy," in *Potato Biology and Biotechnology: Advances and Perspectives*, D. Vreugdenhil, Ed., pp. 157–177, Elsevier, Oxford, UK, 2007.

[3] J. Song, J. M. Bradeen, S. K. Naess et al., "Gene RB cloned from *Solanum bulbocastanum* confers broad spectrum resistance to potato late blight," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 16, pp. 9128–9133, 2003.

[4] E. Van Der Vossen, A. Sikkema, B. T. L. Hekkert et al., "An ancient R gene from the wild potato species *Solanum bulbocastanum* confers broad-spectrum resistance to Phytophthora infestans in cultivated potato and tomato," *Plant Journal*, vol. 36, no. 6, pp. 867–882, 2003.

[5] E. A. G. Van Der Vossen, J. Gros, A. Sikkema et al., "The Rpi-blb2 gene from *Solanum bulbocastanum* is an Mi-1 gene homolog conferring broad-spectrum late blight resistance in potato," *Plant Journal*, vol. 44, no. 2, pp. 208–222, 2005.

[6] T. H. Park, J. Gros, A. Sikkema et al., "The late blight resistance locus Rpi-blb3 from *Solanum bulbocastanum* belongs to a major late blight R gene cluster on chromosome 4 of potato," *Molecular Plant-Microbe Interactions*, vol. 18, no. 7, pp. 722–729, 2005.

[7] T. Oosumi, D. R. Rockhold, M. M. Maccree, K. L. Deahl, K. F. McCue, and W. R. Belknap, "Gene *Rpi-bt1* from *Solanum bulbocastanum* confers resistance to late blight in transgenic potatoes," *American Journal of Potato Research*, vol. 86, no. 6, pp. 456–465, 2009.

[8] D. Carputo, T. Cardi, J. P. Palta, P. Sirianni, S. Vega, and L. Frusciante, "Tolerance to low temperatures and tuber soft rot in hybrids between *Solanum commersonii* and *Solanum tuberosum* obtained through manipulation of ploidy and endosperm balance number (EBN)," *Plant Breeding*, vol. 119, no. 2, pp. 127–130, 2000.

[9] D. Carputo, R. Aversano, A. Barone et al., "Resistance to ralstonia solanacearum of sexual hybrids between *Solanum commersonii* and *S. tuberosum*," *American Journal of Potato Research*, vol. 86, no. 3, pp. 196–202, 2009.

[10] J. G. Hawkes, *The Potato: Evolution, Biodiversity and Genetic Resources*, Smithsonian Institution Press, Washington, DC, USA, 1990.

[11] J. Sliwka, H. Jakuczun, M. Chmielarz et al., "A resistance gene against potato late blight originating from *Solanum* x *michoacanum* maps to potato chromosome VII," *Theoretical and Applied Genetics*, vol. 124, no. 2, pp. 397–406, 2012.

[12] The Tomato Genome Consortium, "Genome sequence and analysis of the tuber crop potato," *Nature*, vol. 475, no. 7355, pp. 189–195, 2011.

[13] The Tomato Genome Consortium, "The tomato genome sequence provides insights into fleshy fruit evolution," *Nature*, vol. 485, pp. 635–641, 2012.

[14] M. W. Bonierbale, R. L. Plaisted, and S. D. Tanksley, "RFLP maps based on a common set of clones reveal modes of chromosomal evolution in potato and tomato," *Genetics*, vol. 120, no. 4, pp. 1095–1103, 1988.

[15] S. D. Tanksley, M. W. Ganal, J. P. Prince et al., "High density molecular linkage maps of the tomato and potato genomes," *Genetics*, vol. 132, no. 4, pp. 1141–1160, 1992.

[16] R. C. Grube, E. R. Radwanski, and M. Jahn, "Comparative genetics of disease resistance within the solanaceae," *Genetics*, vol. 155, no. 2, pp. 873–887, 2000.

[17] J. M. Bradeen, "Cloning of late blight resistance genes: strategies and progress," in *Genetics, Genomics and Breeding of Potato*, J. M. Bradeen and C. Kole, Eds., pp. 153–183, CRC Press/Science Publishers, Enfield, NH, 2011.

[18] K. Mochida and K. Shinozaki, "Advances in omics and bioinformatics tools for systems analyses of plant functions," *Plant Cell Physiology*, vol. 52, no. 12, pp. 2017–2038, 2011.

[19] M. L. Chiusano, N. D'Agostino, A. Traini et al., "ISOL" An Italian SOLAnaceae genomics resource," *BMC Bioinformatics*, vol. 9, no. 2, article S7, 2008.

[20] M. J. Sanchez, *Allelic mining for late blight resistance in wild* Solanum *species belonging to series Bulbocastana [M.S. thesis]*, Department of Plant Pathology, University of Minnesota, St. Paul, Minn, USA, 2005.

[21] M. Iorizzo, *Uso di strumenti genomici per lo studio di linee di patata ottenute tramite ingegneria genetica e genomica [Ph.D. thesis]*, Department of Soil, Plant and Environmental and Animal Production Science, University of Naples Federico II, 2008.

[22] J. J. Doyle and J. L. Doyle, "Isolation of plant DNA from fresh tissue," *Focus*, vol. 12, pp. 13–15, 1990.

[23] D. Jaccoud, K. Peng, D. Feinstein, and A. Kilian, "Diversity arrays: a solid state technology for sequence information independent genotyping," *Nucleic Acids Research*, vol. 29, no. 4, article E25, 2001.

[24] L. Mueller, S. Tanksley, J. J. Giovannoni et al., "A snapshot of the emerging tomato genome sequence," *The Plant Genome*, vol. 2, no. 1, pp. 78–92, 2009.

[25] G. Gremme, V. Brendel, M. E. Sparks, and S. Kurtz, "Engineering a software tool for gene structure prediction in higher organisms," *Information and Software Technology*, vol. 47, no. 15, pp. 965–978, 2005.

[26] X. Huang and A. Madan, "CAP3: a DNA sequence assembly program," *Genome Research*, vol. 9, no. 9, pp. 868–877, 1999.

[27] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[28] M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, and T. L. Madden, "NCBI BLAST: a better web interface," *Nucleic Acids Research*, vol. 36, pp. W5–W9, 2008.

[29] N. Blüthgen, K. Brand, B. Cajavec, M. Swat, H. Herzel, and D. Beule, "Biological profiling of gene groups utilizing gene ontology," *Genome Informatics*, vol. 16, pp. 106–115, 2005.

[30] A. Conesa, S. Götz, J. M. García-Gómez, J. Terol, M. Talón, and M. Robles, "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research," *Bioinformatics*, vol. 21, no. 18, pp. 3674–3676, 2005.

[31] P. Wenzl, H. Li, J. Carling et al., "A high-density consensus map of barley linking DArT markers to SSR, RFLP and STS loci and agricultural traits," *BMC Genomics*, vol. 7, article 206, 2006.

[32] D. M. Spooner and T. Raul Castillo, "Reexamination of series relationships of South American wild potatoes (Solanaceae: *Solanum* sect. *Petota*): evidence from chloroplast DNA restriction site variation," *American Journal of Botany*, vol. 84, no. 5, pp. 671–685, 1997.

[33] D. M. Spooner, K. McLean, G. Ramsay, R. Waugh, and G. J. Bryan, "A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 41, pp. 14694–14699, 2005.

[34] D. Gargano, N. Scotti, A. Vezzi et al., "Genome-wide analysis of plastome sequence variation and development of plastidial CAPS markers in common potato and related *Solanum* species," *Genetic Resources and Crop Evolution*, pp. 1–12, 2011.

[35] D. M. Spooner, G. J. Anderson, and R. K. Jansen, "Chloroplast DNA evidence for the interrelationships of tomatoes, potatoes, and pepinos (Solanaceae)," *American Journal of Botany*, vol. 80, no. 6, pp. 676–688, 1993.

[36] T. L. Weese and L. Bohs, "A three-gene phylogeny of the genus *Solanum* (Solanaceae)," *Systematic Botany*, vol. 32, no. 2, pp. 445–463, 2007.

[37] Y. Wang, A. Diehl, F. Wu et al., "Sequencing and comparative analysis of a conserved syntenic segment in the solanaceae," *Genetics*, vol. 180, no. 1, pp. 391–408, 2008.

[38] F. Wu and S. D. Tanksley, "Chromosomal evolution in the plant family Solanaceae," *BMC Genomics*, vol. 11, no. 1, article 182, 2010.

Hindawi Publishing Corporation
International Journal of Genomics
Volume 2013, Article ID 105108, 7 pages
http://dx.doi.org/10.1155/2013/105108

Research Article

The Novelty of Human Cancer/Testis Antigen Encoding Genes in Evolution

Pavel Dobrynin,[1,2] Ekaterina Matyunina,[1] S. V. Malov,[2] and A. P. Kozlov[1,2]

[1] *The Biomedical Center, Saint Petersburg 194044, Russia*
[2] *Dobzhansky Center for Genome Bioinformatics, Saint Petersburg State University, Saint Petersburg 190004, Russia*

Correspondence should be addressed to Pavel Dobrynin; pdobrynin@gmail.com

Received 9 October 2012; Revised 16 January 2013; Accepted 13 February 2013

Academic Editor: Ancha Baranova

In order to be inherited in progeny generations, novel genes should originate in germ cells. Here, we suggest that the testes may play a special "catalyst" role in the birth and evolution of new genes. Cancer/testis antigen encoding genes (CT genes) are predominantly expressed both in testes and in a variety of tumors. By the criteria of evolutionary novelty, the CT genes are, indeed, novel genes. We performed homology searches for sequences similar to human CT in various animals and established that most of the CT genes are either found in humans only or are relatively recent in their origin. A majority of all human CT genes originated during or after the origin of Eutheria. These results suggest relatively recent origin of human CT genes and align with the hypothesis of the special role of the testes in the evolution of the gene families.

1. Introduction

In order to be inherited in progeny generations, novel genes should originate in germ cells. Available data suggest that the generation of novel genes in germ cells is ongoing process, for example, the promiscuity of gene expression in spermatogenic cells [1, 2]. Novel genes may originate through different mechanisms (retrogenes, segmental duplicates, chimeric, and *de novo* emerged genes), but all of them are uniformly expressed in the testis ([3–8]; reviewed in [9]). These observations led us to suggest that testes may play a "tissue catalyst" role in the birth and evolution of new genes [9]. Previously, we proposed the expression of evolutionarily novel genes in tumors [10].

Cancer/testis or cancer/germline antigen genes are a class of genes with predominant expression in testis and in a variety of tumors, with a significant exclusion of some CT antigens also expressed in the brain. Here we set forth to test the hypothesis that cancer/testis antigen genes should be composed of evolutionarily new or young gene family. We performed homology searches for sequences similar to human CT in various animals. Additionally, as an extensive traffic of novel genes has been described for mammalian X chromosome [3, 6, 11], we also performed this analysis separately for genes located on this chromosome only.

2. Methods

The list of CT antigens gene was retrieved from CT Database (http://www.cta.lncc.br) and included 265 genes. Among them, there are 105 CT antigens that are encoded by the X chromosome (CT-X genes) and 105 that are located on various autosomes (autosome CT genes, or non-X CT genes). Eight CT antigen encoding genes are located on the Y chromosome.

To assess the evolutionary novelty of the studied group of CT genes by searching orthologues for each of CT genes, the HomoloGene.release 66 (http://www.ncbi.nlm.nih.gov/homologene/) tool from NCBI website was used. HomoloGene is a database of both curated and computed gene orthologs and orthologues and now covers 21 organisms. Curated orthologs include gene pairs from the Mouse Genome Database (MGD) at the Jackson Laboratory, the Zebrafish Information (ZFIN) database at the University of Oregon, and from published reports. Computed orthologs

Table 1: Distribution of all human CT genes according to the origin of their orthologues in different taxa of human lineage.

Taxa	Chromosome names																						
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X	Y
Eukaryote	1	1					1			2			1							1	1		1	
Opisthokonta	1									1														
Bilateria		1				1		1									2						2	
Coelomata						2		2	1												1			
Euteleostomi	4	4	2	1	1	2		2		3	1	2		1	1	1			1				3	
Amniota	3														2		2							
Eutheria	2		3	3	1	2		4	3		3	2			1	1	3	1	6	2		1	21	8
Euarchontoglires		1												1		1		1	1	1		1	4	
Catarrhini																		1		1	1		28	
Homininae								1										1					13	
Homo sapiens															1								33	
Total	11	7	5	4	2	7	1	10	4	6	4	4	1	2	5	3	7	4	8	5	3	2	105	8

and orthologues, which are considered putative, are identified from BLAST nucleotide sequence comparisons between all UniGene clusters for each pair of organisms [12]. As an input, the program uses gene name and/or taxon name, and the output is clusters of orthologues. For this study, the search was performed in several completely sequenced eukaryotic genomes, including *H. sapiens, P. troglodytes, M. mulatta, C. lupus, B. taurus, M. musculus, R. norvegicus, G. gallus, D. rerio, D. melanogaster, A. gambiae, C. elegans, S. cerevisiae, K. lactis, A. gossypii, S. pombe, M. oryzae, N. crassa, A. thaliana, O. sativa,* and *P. falciparum*.

According to the origin of their orthologues in different taxa of human phylogeny, the CT genes and all human genes were distributed into 11 groups. The differences in distribution of CT genes and all human genes were assessed using the chi square test [13]. Sheffe's S method of multiple estimation ([14, 15]; for counts see also [16]) was applied to define the difference and to show stochastically that the origin of human CT genes is substantially more recent than that for all human genes.

3. Results

The results obtained using HomoloGene tool applied to human CT genes are presented in Table 1. The full list of studied CT genes is present in Supplementary material (see Supplementary Material available online at http://dx.doi.org/10.1155/2013/105108. HomoloGene assigned each gene to a certain homology group which includes orthologues from different taxa within human lineage. Of 265 genes represented in CT Database, 47 did not match any homology group, probably because of the differences in the gene names making matches with HomoloGene database difficult. Human CT genes orthologues are widely distributed throughout the human lineage. For example, for one CT-X gene (*FAM133A*), the orthologues were found in all Eukaryota, and for two CT-X genes (*MAGEC1* and *SPANXN4*), the orthologues were first found in Bilateria, and for three CT-X genes (*ARX*, *IL13RA,* and *FAM46D*), the time of origin was placed in Euteleostomi. There were substantially larger numbers of CT-X genes with orthologues emerging in Eutheria, Catarrhini, and Homininae and of CT-X genes that were found exclusively in humans. Interestingly, there was a Eutheria-specific subfamily TSPY1 composed of 8 CT genes and located on chromosome Y.

Similarly searches for the orthologues were performed for all CT-X genes, all autosomal CT genes, all human CT genes, and all annotated protein coding genes in human genome (assembly GRCh37) (Table 2 and Figure 1).

The results show that the proportion of autosomal CT genes that has orthologues originated in Euteleostomi and in Eutheria (24.8% and 36.2%, accordingly) is greater than that on chromosome X. Only a few of autosomal CT genes are exclusive for humans. We found that CT gene *POTEB* (prostate, ovary, testis-expressed protein on chromosome 15, Ensembl: ENSG00000233917) has a poorly characterized homologue (LOC100287399, Ensembl: ENSG00000230031) that is according to HomoloGene criteria is exclusive to *H. sapiens*. This newly described homolog (LOC100287399, Ensembl: ENSG00000230031) has not been previously annotated as a gene of CT family.

Among all annotated human protein coding genes, the proportion of genes specific to humans only is very small (0.85%). The list of these human-specific genes includes 163 entries, 33 of which are CT-X genes.

For CT-X genes, the distribution was different: 31.4% of CT-X genes (five *CT45A* genes, twelve *CT47A* genes, fifteen *GAGE* genes, and four *XAGE* genes) are present in humans only, while 39.1% of CT-X genes have orthologues that emerged in *Catarrhini* or *Homininae*. This means that the majority (70.5%) of CT-X genes present in human genome are either novel or relatively recent. At the same time, distribution of all genes located on X chromosome is similar to that for all human genes (see Supplementary Table IV).

The distribution of all human CT genes shows that 30.73% of CT genes have orthologues that originated in *Eutheria*. This proportion is larger than the proportion of all human genes with pan-*Eutherian* orthologues (16.41%). Importantly, 36.7% of all human CT genes originated in *Catarrhini*, *Homininae,*

Table 2: CT-X genes, autosomal CT genes, all human CT genes, and all annotated human genes with orthologues originated in different taxa of *H. sapiens* lineage.

Taxa	CT-X genes		Autosome CT genes		All CT genes		All human genes	
Eukaryota	1,0%	1	7,6%	8	4,13%	9	15,19%	2900
Opisthokonta			1,9%	2	0,92%	2	3,21%	613
Bilateria	1,9%	2	4,8%	5	3,21%	7	8,12%	1549
Coelomata			5,7%	6	2,75%	6	8,27%	1579
Euteleostomi	2,9%	3	24,8%	26	13,30%	29	32,77%	6256
Amniota			6,7%	7	3,21%	7	8,39%	1601
Eutheria	20,0%	21	36,2%	38	30,73%	67	16,41%	3132
Euarchontoglires	3,8%	4	6,7%	7	5,05%	11	1,75%	334
Catarrhini	26,7%	28	2,9%	3	14,22%	31	2,66%	507
Hominoidea	12,4%	13	1,9%	2	6,88%	15	2,38%	454
Homo sapiens	31,4%	33	1,0%	1	15,60%	34	0,85%	163
Total	100,0%	105	100,0%	105	100,00%	218	100,00%	19088

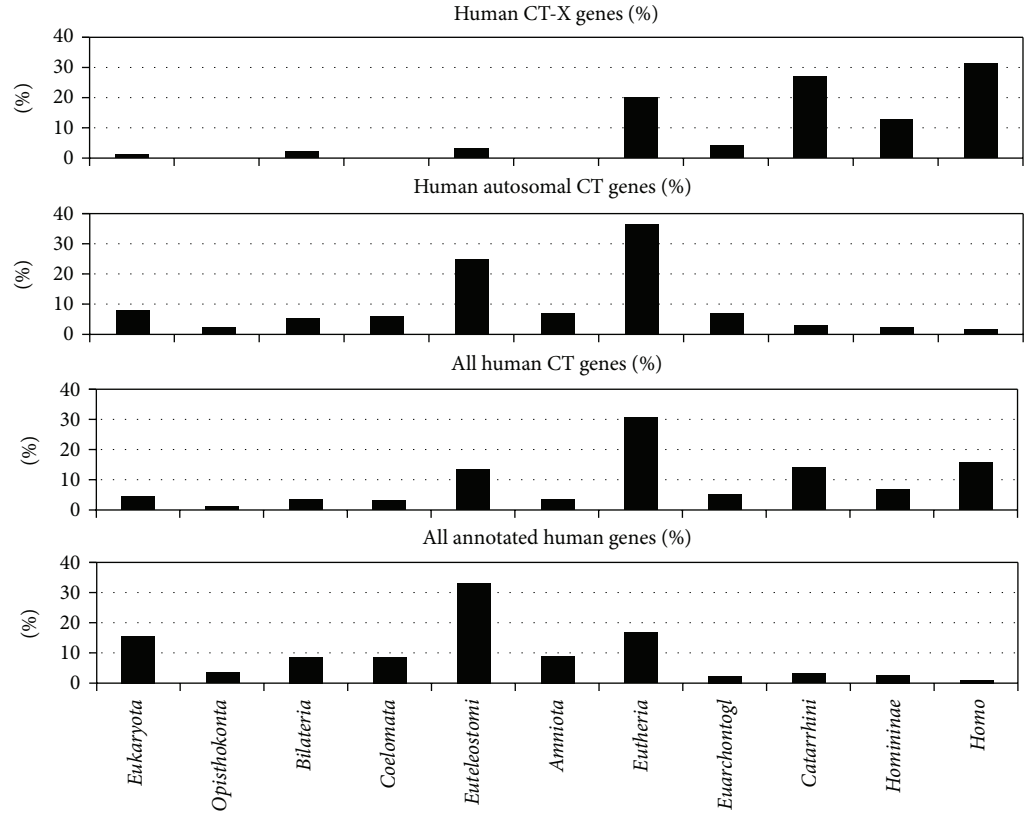


Figure 1: The proportions of CT-X genes, autosomal CT genes, all human CT genes, and all annotated human genes with orthologues originated in different taxa of *H. sapiens* lineage.

or humans. Thus, the majority of human CT genes (72.48%) originated during or after the emergence of Eutheria. On the other side, the majority of annotated human genes (75.95%) were older than *Eutheria*.

A significance of the difference between distribution of all human genes and all human CT genes according to the origin of their orthologues in different taxa was confirmed bychi square test (P value less than 10^{-6}). Moreover, 95% confidence region for the cumulative distribution function of CT human genes displays that CT genes are stochastically younger as compared to all human genes. In other words, the probability that a gene randomly chosen from all human genes is younger than some fixed time T is less than the probability that a randomly chosen CT gene is younger than T. Therefore, there is a significant bias in time of origin for human CT genes as compared to all human genes. If human CT genes would be

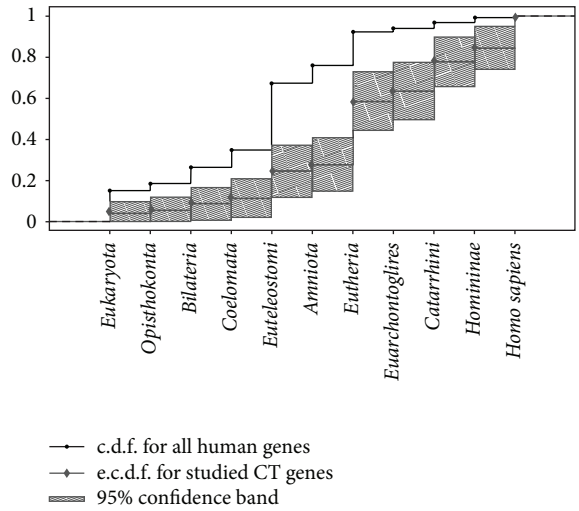


Figure 2: Cumulative distribution function for all human genes and empirical distribution function for all CT human genes, in accordance with the origin of their orthologues in different taxa, with 95% confidence bands. c.d.f.—cumulative distribution function. e.c.d.f.—empirical cumulative distribution function.

obtained as a sample from some probabilistic distribution, the probability that CT human genes originated not earlier than *Catarrhini* or *Eutheria* would be significantly higher than the respective probability for census of all human genes (Figure 2). This statistical trial confirms that the origin of human CT genes is relatively recent.

4. Conclusion

Cancer/testis antigen genes (CTA or CT genes) encode a subgroup of tumor antigens expressed predominantly in testis and various tumors. CT antigens may be also expressed in placenta and in female germ cells [17–20]. In addition, some CT antigens are expressed in the brain [21].

Experimentally, human CT genes were discovered by a variety of immunological screening methods [22], serological identification of antigens by recombinant expression cloning (SEREX) [23], expression database analysis [24, 25], massively parallel signature sequencing [26], and other approaches. The fact that many CT antigens have been identified using SEREX suggests that they are highly antigenic [23, 27].

The first CT gene discovered was *MAGEA1* that encodes for an antigen of human melanoma [22]. This gene belongs to a family of 12 closely related genes clustered at Xq28. A second cluster of *MAGE* genes, *MAGEB*, was discovered at Xp21.3, and the third, encoding *MAGEC* genes, is located at Xq26-27. The expression of *MAGEA-MAGEC* genes (*MAGE-I* subfamily) is restricted to testis and cancer, whereas more distantly related clusters *MAGED-MAGEL* (subfamily *MAGE-II*) are expressed in many normal tissues. *MAGE-I* genes are of relatively recent origin, and *MAGE-II* genes are relatively more ancient. For example, *MAGE-D* genes are conserved between man and mouse. One of these genes corresponds to the founder member of the family, and the other *MAGE* genes are retrogenes derived from the common ancestral gene [19, 28, 29].

To date, CTD atabase (http://www.cta.lncc.br/) includes 265 CT genes. More than half of them are located on X-chromosome (CT-X genes) [21]. The analysis of the DNA sequence of the human X chromosome predicts that approximately 10% of the genes on the X chromosome are of the CT antigen type [30]. Non-X CT genes are distributed throughout the genome and are represented mainly by single-copy genes [19, 27, 31].

In normal testis, CT-X genes are expressed in proliferating germ cells (spermatogonia). Non-X CT genes are expressed during later stages of germ-cell differentiation, that is, spermatocytes [19]. Among human tumors, CT antigens are expressed in melanoma, bladder cancer, lung cancer, breast cancer, prostate cancer, sarcoma, ovarian cancer, hepatocellular carcinoma, hematologic malignancies, and so forth [21, 27, 31, 32]. Genome-wide analysis of 153 cancer/testis genes expression has led to their classification into testis-restricted ($N = 39$), testis/brain-restricted ($N = 14$) and testis-selective ($N = 85$) groups of genes, the latter group showing some expression in nongermline tissues. The majority of testis-restricted genes belong to CT-X group (35 of total 39 testis-restricted groups), while non-X CT genes are expressed in a less restrictive way [21].

Multiple CT antigens are often coexpressed in tumors suggesting that this expression program is coordinated for entire family [19, 23, 33]. CT gene expression is controlled by epigenetic mechanisms which include DNA methylation and histone posttranslational modifications [31]. Other mechanisms of CT gene regulation include sequence-specific transcription factors and signal transduction pathways such as activated tyrosine kinases [34].

The functions of CT-X genes are largely unknown. On the contrary, more is known about functions of non-X CT genes which are associated with meiosis, gametogenesis, and fertilization. Non-X CTs are also more conserved during evolution [21, 27, 31, 32].

CT-X genes tend to form recently expanded gene families, many with nearly identical gene copies [17–20, 26, 32, 35].

The prevalence of large, highly homologous inverted repeats (IRs) containing testes genes on the X- and Y-chromosomes was described in humans and great apes [36, 37]. CT-X gene families are also located in direct or inverted repeats [20].

The study of clusters of homologous genes originated by gene duplication roughly after the divergence of the human and rodent lineages discovered several families of CT genes among recent duplicates [38].

In the other paper, the authors also studied recent duplications in the human genome and found that CT genes were represented in this gene set, including the family of *PRAME* (preferentially expressed antigen of melanoma) genes located on chromosome 1 and expressed in the testis and in a large number of tumors [39]. Duplicated *PRAME* genes are hominid specific, having arisen in human genome since the divergence from chimps. *PRAME* gene family also expanded in other *Eutheria*. Chimp and mouse have

orthologous *PRAME* gene clusters on their chromosomes 1 and 4, respectively [39, 40].

Rapid evolution of cancer/testis genes has been demonstrated on the X chromosome. In particular, the comparison of human: chimp orthologues of these genes has shown that they diverge faster and undergo stronger positive selection than those on the autosomes or than control genes on either X chromosome or autosomes [41].

SPANX-A/D gene subfamily of cancer/testis-specific antigens evolved in the common ancestor of the hominoid lineage after its separation from orangutan. Southern blot and database analyses have detected *SPANX* sequences only in primates [17]. The coding sequences of the *SPANX* genes evolved rapidly, faster than their introns and the 5′ untranslated regions, with accelerated rates of substitutions in both synonymous and nonsynonymous codon positions. The mechanism of *SPANX* genes expansion was segmental DNA duplications, with evidence of positive selection. *SPANX-N* is the ancestral form, from which the *SPANX-A/D* subfamily evolved in the common ancestor of hominoids approximately 7 MYA [35, 42]. *SPANX* genes are expressed in cancer cells and highly metastatic cell lines from melanomas, bladder carcinomas, and myelomas [35].

The *GAGE* cancer/testis antigen gene family contains at least 16 genes which are encoded by an equal number of tandem repeats. All *GAGE* genes are located at Xp11.23. *GAGE* genes are highly identical and evolved under positive selection that supports their recent origin [43, 44].

The *XAGE* family of cancer/testis antigen genes belongs to superfamily of *GAGE*-like CT genes. It is located on chromosome Xp11.21-Xp11.22. Three *XAGE* genes are described, as well as several splice variants of *XAGE-1* [45, 46].

CT45 gene family was discovered by massively parallel signature sequencing. It includes six highly similar (>98%) genes that are cluctered in tandem on chromosome Xq26.3. CT45 antigen is expressed in Hodgkin's lymphoma and in other human tumors [26, 47–49].

CT47 cancer/testis gene family is located on chromosome Xq24. Among normal tissues, it is expressed in the testis and (weakly) in placenta and brain. In tumors, its expression was found in lung cancer and esophageal cancer. The *CT47* family member is characterized by high (>98%) sequence homology. Chimp is the only other species in which a gene homologous to *CT47* was found by other authors [20].

Our work is the first systematic study of the evolutionary novelty of the whole class of CT genes. To assess the evolutionary novelty of CT genes, we applied the HomoloGene tool of NCBI. To construct the clusters of orthologues, the HomoloGene program uses information from blastp, phylogenetic analyses, and syntheny information when it is possible. Cutoffs on bits per position and Ks values are set to prevent unlikely "orthologs" from being grouped together. These cutoffs are calculated based on the respective score distribution for the given groups of organisms [12].

We searched for orthologues of each of CT genes among annotated genes in several completely sequenced eukaryotic genomes and built distributions of all CT-X genes, all autosomal CT genes, all human CT genes, and all annotated protein coding genes from human genome according to the origin of their orthologues in 11 taxa of human lineage.

We have shown that 31.4% of CT-X genes are exclusive for humans and 39.1% of CT-X genes have orthologues originated in *Catarrhini* or *Homininae*. Thereby, the majority of human CT-X genes (70.5%) are novel or recent in its origin. Our data are in good correspondence with evidence obtained by other groups on rapid expansion of certain CT-X gene families and high homology of their members which suggest their recent origin.

Altogether 36.7% of all human CT genes originated in *Catarrhini, Homininae,* and humans. We have also found that 30.73% of all human CT genes originated in *Eutheria*. These CT genes acquired functions in *Eutheria*. This indicates the importance of processes in which tumors and CT antigens were involved during the evolution of *Eutheria*. CT genes originated in *Eutheria* are located mostly on autosomes. CT genes originated in *Catarrhini, Homininae,* and humans are located predominantly on X chromosome. This difference is probably related to evolution of mammalian X chromosome since the origin of *Eutheria* [50], especially to the acquisition of its special role in the origin of novel genes [9].

Thus, the majority of CT-X genes are either novel or young for humans, and the majority of all human CT genes (72.48%) originated during or after the origin of *Eutheria*. These results suggest that the whole class of human CT genes is relatively evolutionarily new.

In its turn, this conclusion confirms our prediction about expression of evolutionary recent and novel genes in tumors [10]. The expression of cancer/testis genes in tumors is then a natural phenomenon, not aberrant process as suggested by many authors (e.g., [19, 27, 32, 34, 40]).

References

[1] E. E. Schmidt, "Transcriptional promiscuity in testes," *Current Biology*, vol. 6, no. 7, pp. 768–769, 1996.

[2] K. C. Kleene, "Sexual selection, genetic conflict, selfish genes, and the atypical patterns of gene expression in spermatogenic cells," *Developmental Biology*, vol. 277, no. 1, pp. 16–26, 2005.

[3] E. Betrán, K. Thornton, and M. Long, "Retroposed new genes out of the X in Drosophila," *Genome Research*, vol. 12, no. 12, pp. 1854–1859, 2002.

[4] C. A. Paulding, M. Ruvolo, and D. A. Haber, "The *Tre2* (USP6) oncogene is a hominoid-specific gene," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 5, pp. 2507–2511, 2003.

[5] X. She, Z. Cheng, S. Zöllner, D. M. Church, and E. E. Eichler, "Mouse segmental duplication and copy number variation," *Nature Genetics*, vol. 40, no. 7, pp. 909–914, 2008.

[6] M. T. Levine, C. D. Jones, A. D. Kern, H. A. Lindfors, and D. J. Begun, "Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 26, pp. 9935–9939, 2006.

[7] T. J. A. J. Heinen, F. Staubach, D. Häming, and D. Tautz, "Emergence of a new gene from an intergenic region," *Current Biology*, vol. 19, no. 18, pp. 1527–1531, 2009.

[8] H. Kaessmann, N. Vinckenbosch, and M. Long, "RNA-based gene duplication: mechanistic and evolutionary insights," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 19–31, 2009.

[9] H. Kaessmann, "Origins, evolution, and phenotypic impact of new genes," *Genome Research*, vol. 20, no. 10, pp. 1313–1326, 2010.

[10] A. P. Kozlov, "The possible evolutionary role of tumors in the origin of new cell types," *Medical Hypotheses*, vol. 74, no. 1, pp. 177–185, 2010.

[11] J. J. Emerson, H. Kaessmann, E. Betrán, and M. Long, "Extensive gene traffic on the mammalian X chromosome," *Science*, vol. 303, no. 5657, pp. 537–540, 2004.

[12] E. W. Sayers, T. Barrett, D. A. Benson et al., "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research*, vol. 40, pp. D13–D25, 2012.

[13] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2011, http://www.R-project.org/.

[14] H. Sheffe, *The Analysis of Variance*, Wiley, New York, NY, USA, 1959.

[15] H. Sheffe, "Multiple testing versus multiple estimation. Improper confidence sets. Estimation of directions and ratios," *The Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 1–29, 1970.

[16] L. A. Goodman, "Simultaneous confidence intervals for contrasts among multinomial populations," *The Annals of Mathematical Statistics*, vol. 35, pp. 716–725, 1964.

[17] A. J. W. Zendman, J. Zschocke, A. A. Van Kraats et al., "The human SPANX multigene family: genomic organization, alignment and expression in male germ cells and tumor cell lines," *Gene*, vol. 309, no. 2, pp. 125–133, 2003.

[18] A. J. W. Zendman, D. J. Ruiter, and G. N. P. Van Muijen, "Cancer/testis-associated genes: identification, expression profile, and putative function," *Journal of Cellular Physiology*, vol. 194, no. 3, pp. 272–288, 2003.

[19] A. J. G. Simpson, O. L. Caballero, A. Jungbluth, Y. T. Chen, and L. J. Old, "Cancer/testis antigens, gametogenesis and cancer," *Nature Reviews Cancer*, vol. 5, no. 8, pp. 615–625, 2005.

[20] Y. T. Chen, C. Iseli, C. A. Yenditti, L. J. Old, A. J. G. Simpson, and C. V. Jongeneel, "Identification of a new cancer/testis gene family, CT47, among expressed multicopy genes on the human X chromosome," *Genes Chromosomes and Cancer*, vol. 45, no. 4, pp. 392–400, 2006.

[21] O. Hofmann, O. L. Caballero, B. J. Stevenson et al., "Genome-wide analysis of cancer/testis gene expression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 51, pp. 20422–20427, 2008.

[22] P. Van Der Bruggen, C. Traversari, P. Chomez et al., "A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma," *Science*, vol. 254, no. 5038, pp. 1643–1647, 1991.

[23] U. Sahin, O. Tereci, H. Schmitt et al., "Human neoplasms elicit multiple specific immune responses in the autologous host," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 25, pp. 11810–11813, 1995.

[24] M. J. Scanlan, C. M. Gordon, B. Williamson et al., "Identification of cancer/testis genes by database mining and mRNA expression analysis," *International Journal of Cancer*, vol. 98, no. 4, pp. 485–492, 2002.

[25] M. J. Scanlan, A. J. Simpson, and L. J. Old, "The cancer/testis genes: review, standardization, and commentary," *Cancer Immunity*, vol. 4, article 1, 2004.

[26] Y. T. Chen, M. J. Scanlan, C. A. Venditti et al., "Identification of cancer/testis-antigen genes by massively parallel signature sequencing," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 22, pp. 7940–7945, 2005.

[27] Y. H. Cheng, E. W. P. Wong, and C. Y. Cheng, "Cancer/testis (CT) antigens, carcinogenesis and spermatogenesis," *Spermatogenesis*, vol. 1, pp. 209–220, 2011.

[28] P. Chomez, O. De Backer, M. Bertrand, E. De Plaen, T. Boon, and S. Lucas, "An overview of the MAGE gene family with the identification of all human members of the family," *Cancer Research*, vol. 61, no. 14, pp. 5544–5551, 2001.

[29] Y. Katsura and Y. Satta, "Evolutionary history of the cancer Immunity antigen MAGE gene family," *PLoS ONE*, vol. 6, no. 6, Article ID e20365, 2011.

[30] M. T. Ross, D. V. Grafham, A. J. Coffey et al., "The DNA sequence of the human X chromosome," *Nature*, vol. 434, pp. 325–337, 2005.

[31] E. Fratta, S. Coral, A. Covre et al., "The biology of cancer testis antigens: putative function, regulation and therapeutic potential," *Molecular Oncology*, vol. 5, no. 2, pp. 164–182, 2011.

[32] O. L. Caballero and Y. T. Chen, "Cancer/testis (CT) antigens: potential targets for immunotherapy," *Cancer Science*, vol. 100, no. 11, pp. 2014–2021, 2009.

[33] U. Sahin, O. Tureci, Y. T. Chen et al., "Expression of multiple cancer/testis antigens in breast cancer and melanoma: basis for polyvalent CT vaccine strategies," *International Journal of Cancer*, vol. 78, pp. 387–389, 1998.

[34] S. N. Akers, K. Odunsi, and A. R. Karpf, "Regulation of cancer germline antigen gene expression: implications for cancer immunotherapy," *Future Oncology*, vol. 6, no. 5, pp. 717–732, 2010.

[35] N. Kouprina, M. Mullokandov, I. B. Rogozin et al., "The SPANX gene family of cancer/testis-specific antigens: rapid evolution and amplification in African great apes and hominids," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 9, pp. 3077–3082, 2004.

[36] H. Skaletsky, T. Kuroda-Kawaguchl, P. J. Minx et al., "The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes," *Nature*, vol. 423, no. 6942, pp. 825–837, 2003.

[37] P. E. Warburton, J. Giordano, F. Cheung, Y. Gelfand, and G. Benson, "Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeated that contain testes genes," *Genome Research A*, vol. 14, no. 10, pp. 1861–1869, 2004.

[38] IHGSC, "International human genome sequencing consortium," *Nature*, vol. 431, pp. 931–945, 2004.

[39] Z. Birtle, L. Goodstadt, and C. Ponting, "Duplication and positive selection among hominin-specific PRAME genes," *BMC Genomics*, vol. 6, article 120, 2005.

[40] T. C. Chang, Y. Yang, H. Yasue, A. K. Bharti, E. F. Retzel, and W. S. Liu, "The expansion of the PRAME gene family in Eutheria," *PLoS ONE*, vol. 6, no. 2, Article ID e16867, 2011.

[41] B. J. Stevenson, C. Iseli, S. Panji et al., "Rapid evolution of cancer/testis genes on the X chromosome," *BMC Genomics*, vol. 8, article 129, 2007.

[42] N. Kouprina, V. N. Noskov, A. Pavlicek et al., "Evolutionary diversification of SPANX-N sperm protein gene structure and expression," *PLoS ONE*, vol. 2, no. 4, article e359, 2007.

[43] Y. Liu, Q. Zhu, and N. Zhu, "Recent duplication and positive selection of the GAGE gene family," *Genetica*, vol. 133, no. 1, pp. 31–35, 2008.

[44] M. F. Gjerstorff and H. J. Ditzel, "An overview of the GAGE cancer/testis antigen family with the inclusion of newly identified members," *Tissue Antigens*, vol. 71, no. 3, pp. 187–192, 2008.

[45] A. J. W. Zendman, A. A. Van Kraats, U. H. Weidle, D. J. Ruiter, and G. N. P. Van Muijen, "The XAGE family of cancer/testis-associated genes: alignment and expression profile in normal tissues, melanoma lesions and Ewing's sarcoma," *International Journal of Cancer*, vol. 99, no. 3, pp. 361–369, 2002.

[46] S. Sato, Y. Noguchi, N. Ohara et al., "Identification of XAGE-1 isoforms: predominant expression of XAGE-1b in testis and tumors," *Cancer Immunity*, vol. 7, article 5, 2007.

[47] Y. T. Chen, M. Hsu, P. Lee et al., "Cancer/testis antigen CT45: analysis of mRNA and protein expression in human cancer," *International Journal of Cancer*, vol. 124, no. 12, pp. 2893–2898, 2009.

[48] Y. T. Chen, A. Chadburn, P. Lee et al., "Expression of cancer testis antigen CT45 in classical Hodgkin lymphoma and other B-cell lymphomas," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 7, pp. 3093–3098, 2010.

[49] H. J. Heidebrecht, A. Claviez, M. L. Kruse et al., "Characterization and expression of CT45 in Hodgkin's lymphoma," *Clinical Cancer Research*, vol. 12, pp. 4804–4811, 2006.

[50] B. T. Lahn and D. C. Page, "Four evolutionary strata on the human X chromosome," *Science*, vol. 286, no. 5441, pp. 964–967, 1999.

Hindawi Publishing Corporation
International Journal of Genomics
Volume 2013, Article ID 671316, 11 pages
http://dx.doi.org/10.1155/2013/671316

Research Article

Effects of Taxon Sampling in Reconstructions of Intron Evolution

Mikhail A. Nikitin and Vladimir V. Aleoshin

Belozersky Institute for Physicochemical Biology, Lomonosov Moscow State University, Moscow 119991, Russia

Correspondence should be addressed to Mikhail A. Nikitin; nikitin.fbb@gmail.com

Received 15 October 2012; Accepted 2 January 2013

Academic Editor: Yuri Panchin

Introns comprise a considerable portion of eukaryotic genomes; however, their evolution is understudied. Numerous works of the last years largely disagree on many aspects of intron evolution. Interpretation of these differences is hindered because different algorithms and taxon sampling strategies were used. Here, we present the first attempt of a systematic evaluation of the effects of taxon sampling on popular intron evolution estimation algorithms. Using the "taxon jackknife" method, we compared the effect of taxon sampling on the behavior of intron evolution inferring algorithms. We show that taxon sampling can dramatically affect the inferences and identify conditions where algorithms are prone to systematic errors. Presence or absence of some key species is often more important than the taxon sampling size alone. Criteria of representativeness of the taxonomic sampling for reliable reconstructions are outlined. Presence of the deep-branching species with relatively high intron density is more important than sheer number of species. According to these criteria, currently available genomic databases are representative enough to provide reliable inferences of the intron evolution in animals, land plants, and fungi, but they underrepresent many groups of unicellular eukaryotes, including the well-studied Alveolata.

1. Introduction

Introns are noncoding sequences inside many eukaryotic genes. Their abundance may vary several orders of magnitude, from hundreds of thousands in mammalian genomes to less than 100 in the genome of *Saccharomyces cerevisiae*. The origin and evolution of introns is a highly controversial topic despite 25 years of research. The variation of intron content between different lineages suggests a high variation in the rate of intron gain and loss, which may relate to differences in population size, absence or presence of the sexual process, activity of transposable elements, properties of the splicing mechanism, and many other characteristics of genomes, organisms, and populations.

Intron sequences evolve at a high rate, and the negative selection typically stabilizes only few nucleotide positions in splicing sites and the branch point. Interestingly, a significant portion of introns occupy the same positions in the same genes in species that diverged billions years ago (20% of common introns in mammals and *Arabidopsis thaliana* for the set of 684 conservative genes [1]). These introns were interpreted either as ancestral, originating before the divergence of major eukaryotic lineages, or as convergently inserted in the same positions due to sequence properties (the "protosplice sites"). Both explanations can be true for different subsets of introns in same genome; however, the proportion of ancestral and convergent intron positions in shared introns is controversial. Its estimates vary from 2% to 18% [2] and even more than 50% [3]. The latter estimate was obtained using ad hoc algorithm and datasets and is not directly comparable to others. Functional explanations of the extremely low rate of intron loss are not known. Many introns were found to contain functional elements, such as transcriptional and splicing regulators [4], small regulatory RNAs which produced during the subsequent cleavage of the excised intron, nonsense-mediated decay signals, signals of nuclear export, and others. However, these processes are not evolutionarily conserved, which therefore does not explain the survival of introns for billions of years.

With the influx of new genomic data, our view of intron evolution changes. For example, the high intron content in vertebrate genomes was initially interpreted as a derived feature of this lineage. However, genomic analysis of the polychaete *Platynereis dumerili* [5] suggested that most vertebrate introns were already present in ancestral Bilateria and subsequently lost in insect and nematode lineages. In next years, analyses of genomic sequences of cnidarians [6], Placozoa [7], sponges [8], choanoflagellate [9], and early-diverging fungi [10] pushed the origin of abundant vertebrate introns back to the ancestral metazoans and, for some, even earlier, to the unicellular common ancestor of animals and fungi. A recent study of the intron evolution in Alveolata and stramenopiles with data on 23 species infers a highly intron-rich ancestors of Alveolata and Alveolata+stramenopiles, with latter containing more introns per gene as humans [11]. This is unexpected, because all extant members of these groups exhibit a low or at best moderate (*Thalassiosira pseudonana*) intron density [12]. These examples raise the following questions.

(i) How does the available taxon sampling affect our studies of the evolution of introns?

(ii) How can the taxon sampling be tested to provide accurate reconstructions?

(iii) Which species should be added to compensate for an incomplete taxon sampling?

2. Materials and Methods

To address these questions, we compiled dataset of intron-exon structures of two ribosomal protein genes (rpS5 and rpL12) for 80 species representing three major eukaryotic groups, Opisthokonta, Plantae, and SAR (Stramenopiles-Alveolata-Rhizaria), using data from publicly available databases of completed and ongoing genome projects. Phylogenetic relations of the analyzed species according to recent studies [13–20] are depicted on Figure 1. For unannotated data, putative rpS5 and rpL12 cDNA and genomic sequences were found with BLAST, and intron-exon boundaries were established using Genscan [21]. We generated 660 random subsamplings ranging from 15 to 75 species from the initial 80 species set using custom Python scripts (100 subsamplings with 15 and 20 species, 80 with 25, 60 with 30 and 35 each, 40 with 40, 45, 50 and 55 each, 30 with 60 and 65, and 20 with 70 and 75 species). The Csuros [22] and NYK [2] algorithms of inferring intron evolution were run on each of these subsamplings.

Results were imported in STATISTICA 8 for statistical analysis and scatterplot generation. If no members of a taxon were present in a subsampling, this subsampling was discarded from calculations and scatterplots for this taxon.

3. Results

3.1. Overview. We reconstructed intron phylogenies for the full set of 80 species and for 660 random subsamplings using the algorithms by Csuros and NYK. As depicted in Figure 2, different taxon samplings produce different results. In many smaller subsets, there were no members of a particular taxa. These subsets were excluded from calculations of average ancestral intron densities for these taxa. For example, in 105 out of 660 subsets, there were no nematodes, and they were excluded from calculations of average intron density in the ancestor of Nematoda.

For the NYK algorithm, the most striking difference is observed in the internal nodes of the bikont half of the eukaryotic tree. Using subsets of 20 species, one can see a more or less constant intron density among the internal branches in different bikont groups such as Alveolata, stramenopiles, and Viridiplantae. The analysis of original set of 80 species inferred almost intronless ancestors for these groups and recent episodes of intron gain along terminal branches. A similar, but less pronounced, pattern is also observed for the Ascomycota, Basidiomycota, and the animal-fungal ancestor. These internal nodes also appear more intron rich when sparse taxon coverage is used. Among the Metazoa and their closest relatives, Choanoflagellata, the results do not change significantly varying the taxon coverage.

For the Csuros algorithm, significant differences were also observed between broader and narrower taxon samplings. Again, these differences are most prominent on internal branches of the Bikonta and Fungi. For small species sets, the output of Csuros algorithm is similar to that of NYK. Analysis of the complete taxon set of 80 species returns very high intron densities for the ancestors of Sporozoa, Apicomplexa, Alveolata, and Ascomycota, far exceeding the observations in recent organisms. Particularly, in the ancestor of Apicomplexa, the estimated intron density in analysis of 80 taxa equals 22/kb, which is three times higher than in mammals (7/kb).

3.2. Specific Effects of Taxon Sampling on Different Nodes. As can be noticed, varying the taxon sampling size affects particular nodes (such as Alveolata and Viridiplantae) more than others (e.g., Metazoa). Figure 3 reproduces this pattern in more detail. For the Bilateria, one can see that average statistics are the same for both algorithms and do not correlate with sampling size. The only observed effect of sampling is a significant dispersion of estimates between smaller taxon sets and uniform patterns when the sampling size is 40 or more species. At internal nodes of Bilateria, such as Nematoda, Insecta, and Spiralia, a significant positive correlation is observed between the intron number and taxon sampling size. Dispersion between samplings of same size is large for smaller samplings and decreases on larger taxon sets. For the Fungi, Basidiomycota, Opisthokonta, and Viridiplantae, a high dispersion is observed for all sampling sizes, albeit less on the larger ones. The average intron density in Fungi and Opisthokonta shows a negative correlation with the sampling size for the both NYK and Csuros algorithms (Table 1), while for the Basidiomycota and Viridiplantae, these correlations are insignificant. The Alveolata exhibit a high correlation of the inferred intron density with the sampling size, however correlation patters are different for

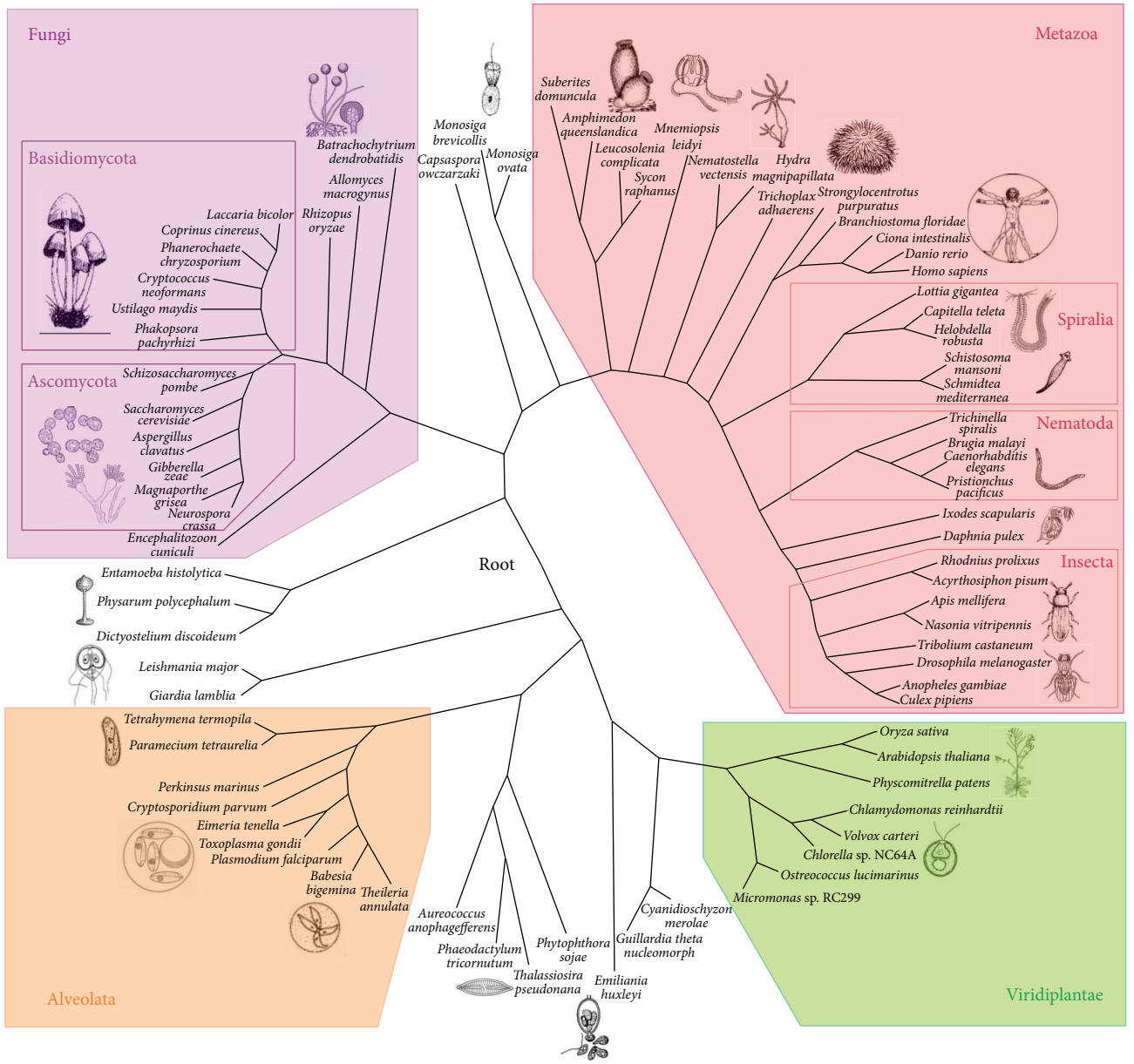


Figure 1: The phylogenetic tree for the initial species set according to [13–20].

the Csuros and NYK algorithms. The two algorithms behave similarly with smaller samplings; however, with larger sets, Csuros infers extremely high intron numbers, and NYK infers their complete absence. Similar patterns were also observed for the common ancestors of Apicomplexa, Sporozoa, and Ascomycota.

3.3. Factors Affecting the Reconstruction. One may discuss three factors that influence ancestral reconstructions under varying the taxon sampling. First, broader sampling usually produces more descendants of a given internal node in analyses. Second, the number of outgroup taxa also depends on the sampling size. Third, particular key taxa may strongly affect reconstruction when present in the dataset, which are more likely be found in larger samplings. These three factors may contribute differently and produce a mixed effect. To evaluate their contributions separately, we performed a multiple regression analysis using the numbers of descendants and outgroup species and the presence/absence of particular descendants as independent variables. The results are presented in Table 1.

The regression analysis shows that for the Insecta and Spiralia, the main affecting factor is the presence of particular species in the dataset for both Csuros and NYK algorithms. These species are *Rhodnius prolixus* for the Insecta and *Lottia gigantea* for the Spiralia (partial correlation coefficients beta equal 0.65 and 0.67). Correlations with the number of descendants for these nodes range within 0.35–0.37 for Csuros and

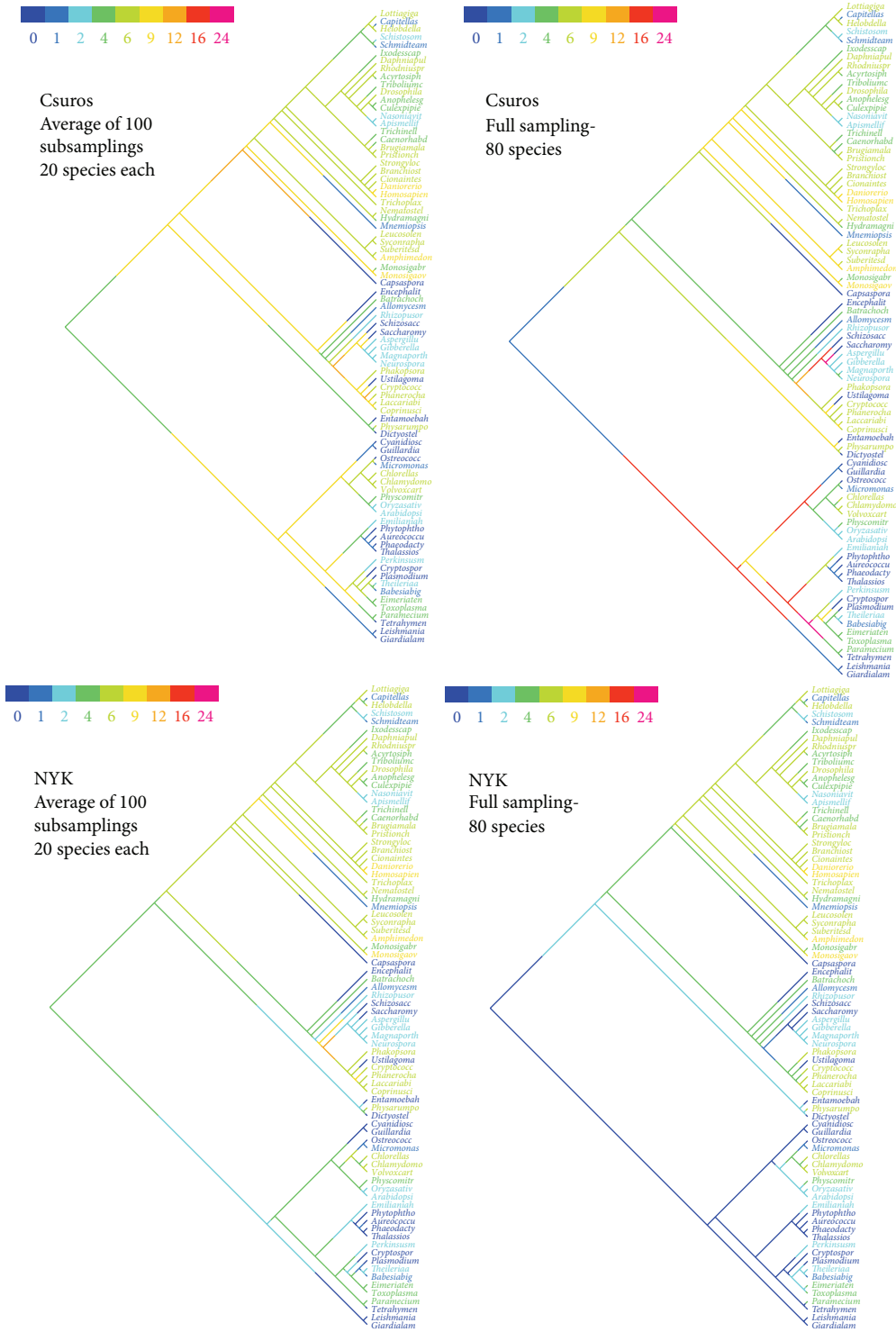


Figure 2: Examples of the intron evolution reconstructed with different taxon samplig size. The intron density on each branch (in introns/kb) is color-coded according to scale in the upper left corner. Upper row: Csuros algorithm, lower row: NYK algorithm. Left column: averaged intron densities using 100 subsets of 20 species each. Right column: full set of 80 species.

0.12–0.30 for NYK. Correlations with the sampling size are less pronounced. For Insecta, we also found that correlation of results with the sampling size is significant only when *Rhodnius prolixus* is not sampled (Figures 4 and 5).

For Nematoda, presence of each of the four of its descendants shows a significant effect, with beta positive, ranging within 0.33–0.39 for *Trichinella spiralis*, *Brugia malayi,* and *Pristionchus pacificus,* and negative −0.20 for *Caenorhabditis*

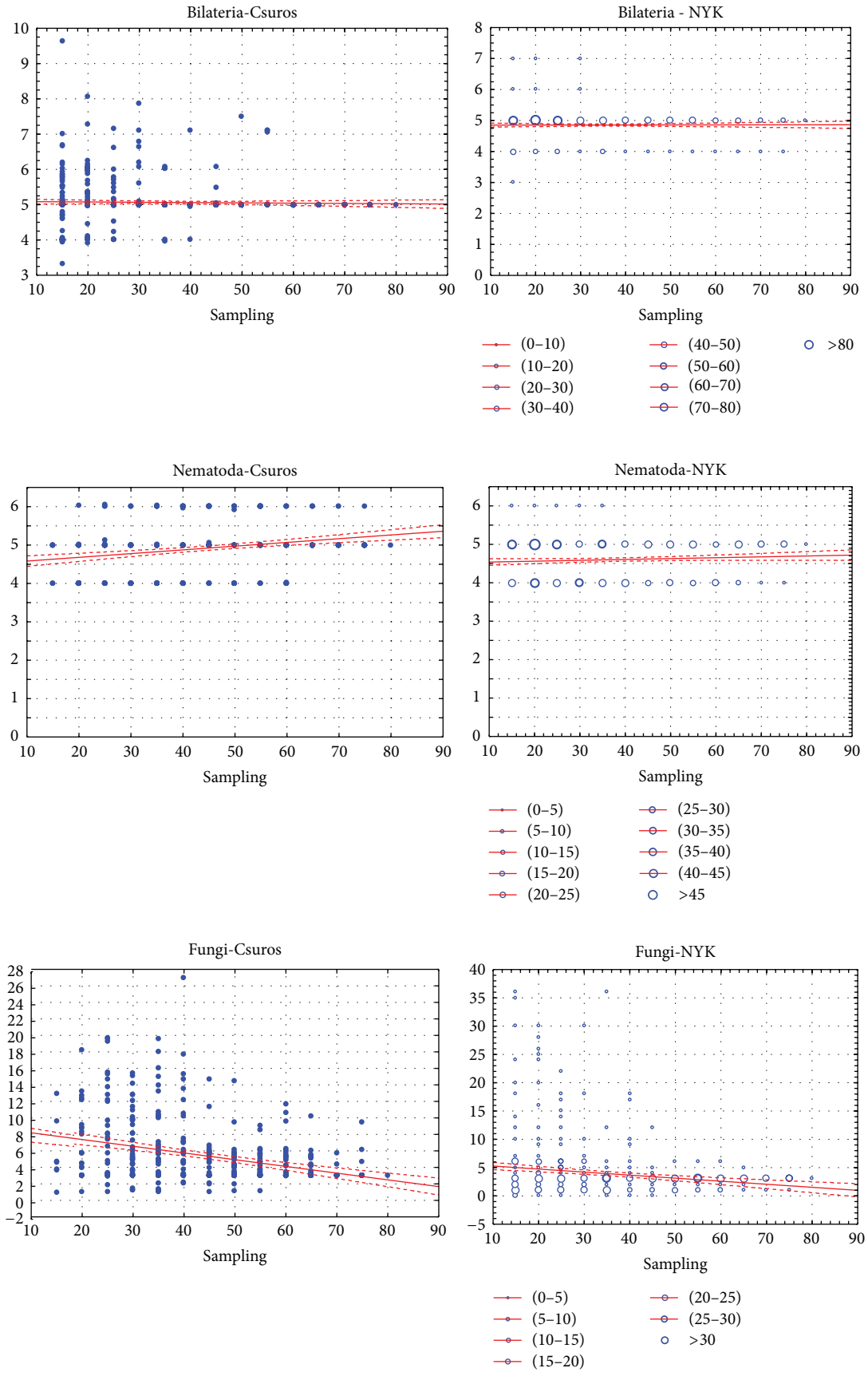


Figure 3: Continued.

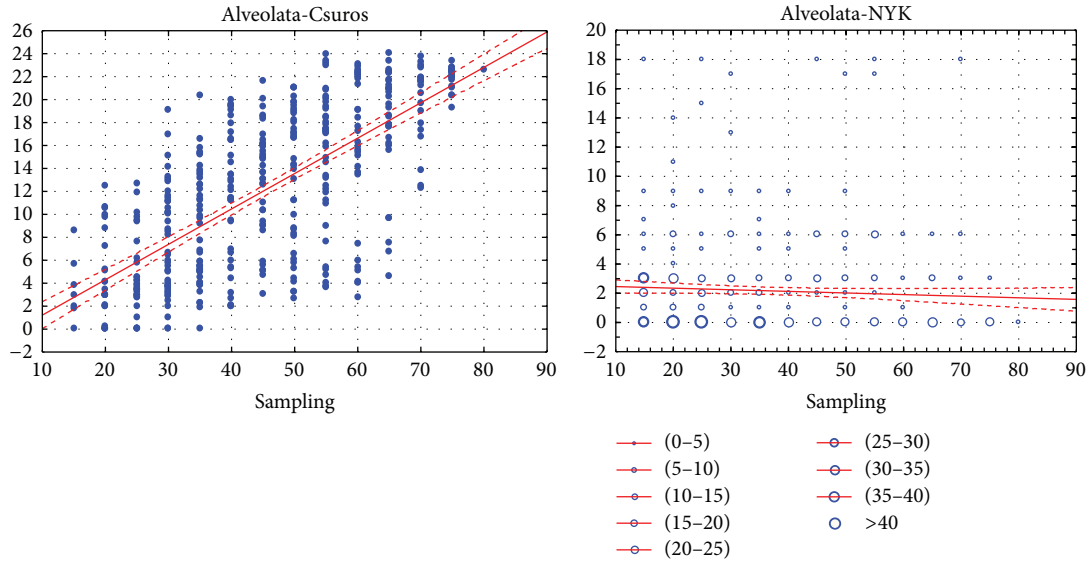


Figure 3: Examples of sampling size-intron density correlation patterns inferred by the Csuros (left) and NYK (right) algorithms. Horizontal axis: number of species in the sampling. Vertical axis: inferred intron density at the node (introns/kb). Note that NYK returns estimations of ancestral intron count rounded down to integers, therefore bubble diagram was used. Bubble size represents number of data points at the same scatterplot coordinates.

elegans. These correlations are almost the same for Csuros and NYK algorithms.

For other nodes, significance of different factors depends on the algorithm choice.

When the Csuros algorithm is used, for the nodes of Fungi, Basidiomycota, Apicomplexa, and Viridiplantae, the presence of some species affects inferences of intron evolution more than other factors. These species are *Allomyces macrogynus* and *Batrachochytrium dendrobatidis* for the Fungi (beta = −0.48 and −0.53), *Phakopsora pachyrhizi* for the Basidiomycota (beta = −0.32), *Chlorella variabilis* for the Viridiplantae (beta = −0.23), and *Perkinsus marinus* for the Apicomplexa (beta = −0.26). Furthermore, a significant correlation of the ancestral intron density with the sampling size and the number of descendants is observed only when the critical species are absent (both species absent in case of Fungi).

For the Ascomycota, Alveolata, and Sporozoa, critical species are not easily identified. Presence of every descendant of these nodes shows a significant correlation with the inferred ancestral intron density. Still, some species are more important than others. Among Ascomycota, these are yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* (beta = 0.46 and 0.41, while for other species is lower than 0.20). For the Sporozoa, most variation is due to *Cryptosporidium parvum* (beta = 0.45, for others less then 0.23), and in the Alveolata, these are two ciliates *Paramecium tetraurelia* and *Tetrahymena thermophila* (beta = 0.35 and 0.61, while lower than 0.15 in other cases). Unlike the aforementioned nodes, for the Ascomycota, Sporozoa, and Alveolata, most significant correlation of the ancestral intron density with sampling size and descendants number is observed only when critical species are present in subsets.

For NYK algorithm, certain species usually show the highest impact on reconstructions of ancestral introns, but the overall picture is often more complicated. For the Fungi, similarly to the Csuros algorithm, the critical species are *Allomyces macrogynus*, *Batrachochytrium dendrobatidis,* and *Encephalitozoon cuniculi.* Presence of any of them greatly reduces the dispersion between subsamplings and prevents very high or very low estimates. For the Alveolata, the critical species are again ciliates, but presence of *Paramecium tetraurelia* positively correlated with the ancestral intron density, and that of *Tetrahymena thermophila*—negatively. In the Sporozoa, the highest correlation is observed for *Cryptosporidium parvum* and *Eimeria tenella*, again with opposite signs. In the Apicomplexa, there are four species that exhibit significant effects—*Cryptosporidium parvum, Eimeria tenella*, *Paramecuim tetraurelia,* and *Tetrahymena thermophila.* Among the Viridiplantae, there are two important descendants, *Ostreococcus tauri* and *Oryza sativa*, and two important outgroup species, *Paramecuim tetraurelia* and *Tetrahymena thermophila.* A significant effect on the variations of ancestral intron count for the Basidiomycota was found for six species: four descendants (*Cryptococcus neoformans, Phakopsora pachyrhizi, Ustilago maydis,* and *Coprinus cinereus*) and two outgroup species (*Allomyces macrogynus* and *Batrachochytrium dendrobatidis*). Analyses for the Ascomycota robustly produce the estimation of 2 introns/kb, with only 13 out of 660 subsamplings exhibiting much higher estimates.

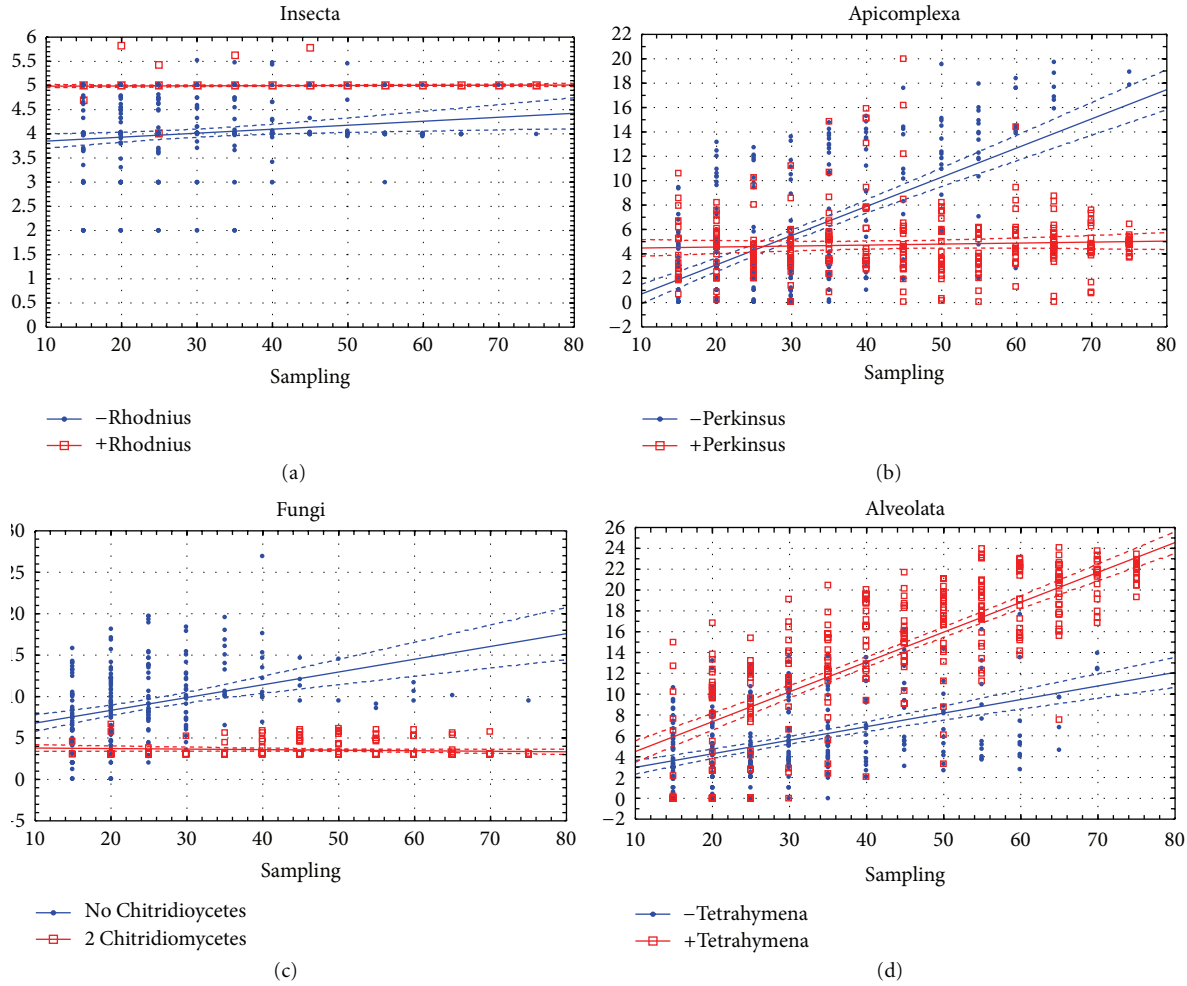


FIGURE 4: Effects of critical species sampling on the behavior of the Csuros algorithm. Horizontal axis: number of species in the sampling. Vertical axis: inferred intron density at the node (introns/kb).

One can see that many key species are the same for both algorithms—for Fungi or ciliates for Alveolata. However, there are significant differences in the importance of outgroup species. They are often important for the NYK algorithms but show only minor effects for the Csuros.

4. Discussion

4.1. Cases of Overestimation of the Ancestral Intron Count. The Csuros algorithm in many cases outputs unrealistic, very high intron densities of 15–20/kb, which is three times higher than the observed values in any recent organism and seems unlikely if we consider the spliceosome positioning on pre-mRNA. Such overestimation is commonly found for the Alveolata, Sporozoa, and Ascomycota and also occurs in a portion of subsamplings for the Apicomplexa, Fungi, and Basidiomycota. Our analysis of these anomalies shows that they occur when very intron-poor taxa occupy the basal position among descendants of a node. Yeasts, ciliates, and *Cryptosporidium parvum* are intron-poor and basal for the Ascomycota, Alveolata, and Sporozoa, respectively, in our full set of 80 species. The chance that these species are present in the analysis increases with the subset size, leading to a high positive correlation of the inferred intron density with sampling size. With the Fungi, Basidiomycota, and Apicomplexa, the full set contains relatively intron-rich basal species (*Allomyces macrogynus*, *Batrachochytrium dendrobatidis*, *Phakopsora pachyrhizi,* and *Perkinsus marinus*), followed by intron-poor branches (yeasts, *Ustilago maydis*, and *Cryptosporidium parvum*). For these nodes, the inferred ancestral intron densities show a bimodal distribution, depending on which species happens to be basal in subsamplings. Interestingly, even for the Metazoa, there are several cases when the Csuros algorithm overestimates the ancestral intron density to exceed 10/kb. In all such cases, we found that the extremely intron-poor ctenophore *Mnemiopsis leydi* in these subsamplings falls in the basal position within Metazoa, while all intron-rich poriferan species are absent.

The NYK algorithm is also prone to overestimation of the ancestral intron count. This is often observed with the Fungi and sometimes with the Basidiomycota, Ascomycota, and Alveolata. As we have found, the prerequisite for such an overestimation is the absence of *Allomyces macrogynus* and

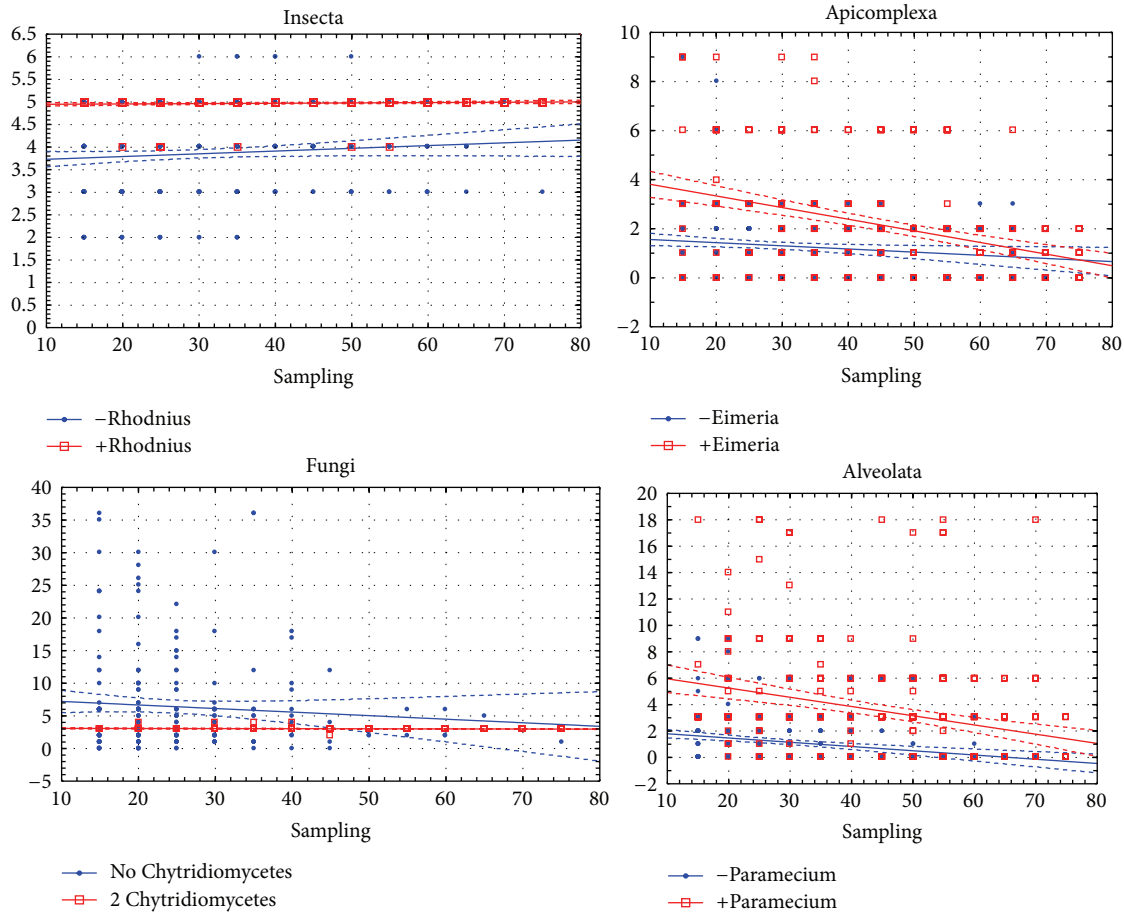


Figure 5: Effects of critical species sampling on the behavior of the NYK algorithm. Horizontal axis: number of species in the sampling. Vertical axis: inferred intron density at the node (introns/kb).

Batrachochytrium dendrobatidis (for Fungi, Basidiomycota, and Ascomycota), presence of *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* (for Ascomycota only), and presence of *Paramecuim tetraurelia* and *Tetrahymena thermophila* (for Alveolata).

These conditions are similar for those for the Csuros algorithm, but NYK shows a much lesser degree of systematic overestimation. It is especially shown with the example of Ascomycota; the overestimation by Csuros was found in more than half of subsamplings, while by NYK—only in 13 out of 660 subsamplings. The factor analysis also shows that the set of outgroup taxa does not affect the reconstructions with the Csuros algorithm but is important in the case of NYK.

4.2. How Many Taxa Are Enough? Using the nodes where different algorithms produce similar results, we could evaluate the number of descendants required for accurate reconstructions of intron evolution. In the Spiralia and Insecta, basal intron-rich species (*Lottia gigantea* and *Rhodnius prolixus*) strongly affect the results, while in the case of 8-species sets, the results with and without *Rhodnius prolixus* are very similar (Figure 6). For the Spiralia, a similar trend exists, however, less pronounced due to only 5 available descendants. For the Bilateria and Metazoa, reconstructions are the same with the both algorithms and almost do not depend on sampling, possibly due to a high number of descendants in the sampling (24 and 32, resp.). The average intron count for these nodes does not correlate with the sampling size even if all subsamplings with more than 10 descendants of these nodes are discarded. So, the Metazoa and Bilateria are not very useful for estimating the sampling adequacy. With the results for Insecta, we conclude that 8–10 species should be enough given no catastrophic intron loss among the descendants of the analyzed node. The results obtained for the Bilateria and Metazoa do not contradict with this conclusion.

4.3. Comparison with Earlier Intron Evolution Studies. The recent work by Csuros et al. [23] uses an MCMC-based algorithm for the reconstruction of the intron evolution and a broad sampling of 99 species. It also shows that the reliability of the reconstruction of the intron evolution differs between nodes. Their algorithm produced not only inferred estimates of the ancestral intron density, but also its Bayesian posterior distributions. Similarly to our results, the estimates for the Metazoa and Bilateria are robust, the Alveolata exhibit a significant uncertainty, and the stramenopiles-Alveolata (SAR) group or Amoebozoa posterior distribution

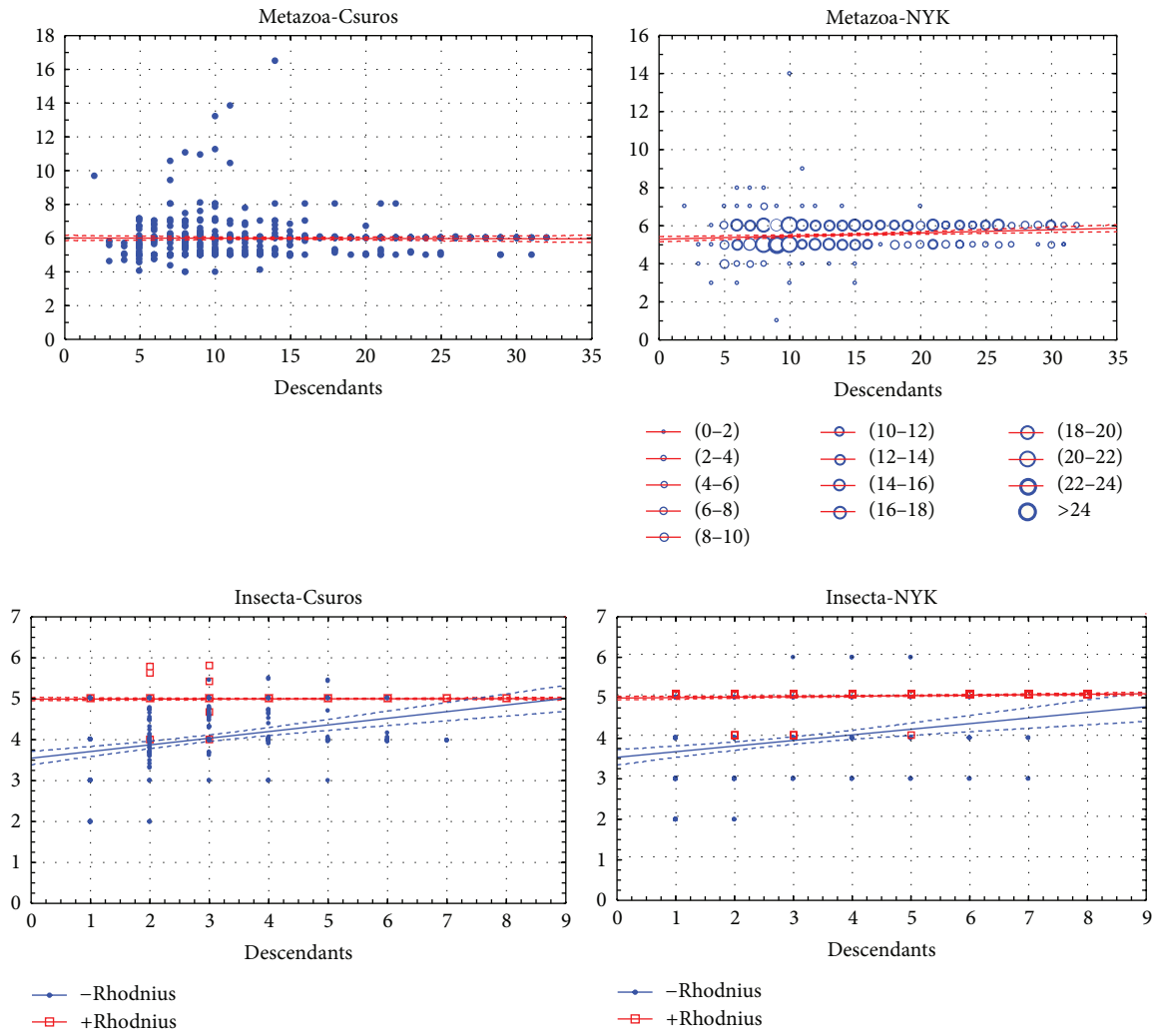


Figure 6: Effects of the number of descendants and sufficiency of taxon sampling. Horizontal axis: number of species in the sampling. Vertical axis: inferred intron density at the node (introns/kb).

shows that the estimations are unreliable. Despite the broad taxon sampling, the authors do not use data on such deep-branching intron-rich species as *Perkinsus marinus* in the Alveolata and *Physarum polycephalum* in the Amoebozoa. We predict that adding these species to the authors' 99-species set will stabilize the results for the Alveolata and Amoebozoa, respectively.

The Csuros's algorithm tested in our work was used, for example, in the study [11] of intron evolution in the Alveolata and stramenopiles. The authors report unusually high estimates of the ancestral intron densities for many nodes. The highest was 7.5 introns/kb in ancestor of Alveolata, which is 20% higher than in the most intron-rich modern organisms. Our observations suggest that this is likely a systematic bias of the Csuros algorithm. This view is supported by the results obtained with the MCMC algorithm from [14], where the inferred intron density in the ancestral Alveolata is more conservatively estimated at 5.0 introns/kb.

It is of interest to compare the results of our study with [10]. Stajich et al. studied the intron evolution in Fungi, using a sampling of 25 species and four algorithms, NYK, Csuros, Roy-Gilbert, and EREM. The results of all algorithms were in good agreement for most nodes, including the Ascomycota. No systematic overestimations by any algorithm were detected, unlike our work and [11]. There might be two reasonable explanations: (1) a broad gene sampling (1161) allowed to correctly estimate the rate of intron loss even under the low intron density in the basal ascomycete *Schizosaccharomyces pombe*; (2) a broad species sampling (5) of extremely intron-poor hemiascomycete yeasts and a differential intron loss in them allowed for the conservation of a considerable subset of the ancestral introns. Unfortunately, a broader gene sampling is not always available for groups with an extensive gene loss. For example, for 23 species in [11] (11 stramenopiles and Alveolata and 12 outgroup species), only 394 orthologous genes were present in at least 18 out of

Table 1: Partial correlations of the sampling size, number of descendants, outgroups, and presence/absence of key species with intron densities in ancestral nodes. Italics are correlations significant at P level > 0.05. In **bold** are highest correlations for the node.

NYK	Sampling		Descendants		Outgroup-partial		1st descendant			2nd descendant		
	Beta	P level	Beta	P level	Beta	P level	Beta	P level	Name	Beta	P level	Name
Fungi	*−0.24*	*0.00*	*−0.09*	*0.02*	−0.01	0.72	**−0.27**	**0.00**	**Allomyces**	*−0.18*	*0.00*	*Encephalitozoon*
Bilateria	*0.16*	*0.00*	0.05	0.21	−0.04	0.26	**−0.25**	**0.00**	**Trichinella**	*0.22*	*0.00*	*Daphnia*
Spiralia	*0.14*	*0.01*	*0.30*	*0.00*	−0.08	0.05	**0.68**	**0.00**	**Lottia**	*0.30*	*0.00*	*Helobdella*
Nematoda	*0.20*	*0.00*	*0.21*	*0.00*	−0.08	0.05	**−0.29**	**0.00**	**Caenorhabditis**	*0.25*	*0.00*	*Brugia*
Insecta	*0.33*	*0.00*	*0.35*	*0.00*	−0.08	0.05	**0.66**	**0.00**	**Rhodnius**	*0.24*	*0.00*	*Drosophila*
Ascomycota	−0.09	0.07	0.07	0.09	*−0.09*	*0.03*	**−0.13**	**0.01**	**Schizosacch**	*0.10*	*0.03*	*Magnaporthe*
Basidiomycota	**−0.31**	**0.00**	*−0.13*	*0.00*	*−0.09*	*0.02*	*−0.22*	*0.00*	*Cryptococcus*	*−0.15*	*0.00*	*Ustilago*
Viridiplantae	−0.06	0.24	0.04	0.29	*−0.09*	*0.03*	**−0.24**	**0.00**	**Ostreococcus**	−0.17	*0.00*	*Oryza*
Alveolata	−0.09	0.09	*0.13*	*0.00*	*−0.15*	*0.00*	**0.45**	**0.00**	**Paramecium**	−0.23	*0.00*	*Tetrahymena*
Apicomplexa	*−0.19*	*0.00*	*0.10*	*0.02*	*−0.17*	*0.00*	**0.35**	**0.00**	**Eimeria**	−0.24	*0.00*	*Cryptosporidium*
Sporozoa	*−0.12*	*0.02*	0.02	0.60	−0.07	0.08	**0.30**	**0.00**	**Eimeria**	**−0.30**	**0.00**	**Cryptosporidium**
Csuros	Sampling		Descendants-partial		Outgroup		1st descendant			2nd descendant		
	Beta	P level	Beta	P level	Beta	P level	Beta	P level	Name	Beta	P level	Name
Fungi	*−0.34*	*0.00*	*−0.24*	*0.00*	0.02	0.63	**−0.53**	**0.00**	**Batrachochytrium**	*−0.48*	*0.00*	*Allomyces*
Bilateria	−0.05	0.30	−0.09	0.09	0.02	0.65	**−0.15**	**0.00**	**Nematostella**	*−0.13*	*0.00*	*Branchiostoma*
Spiralia	*0.13*	*0.01*	*0.36*	*0.00*	*−0.10*	*0.01*	**0.71**	**0.00**	**Lottia**	*0.48*	*0.00*	*Helobdella*
Nematoda	*0.26*	*0.00*	**0.36**	**0.00**	−0.06	0.19	*0.39*	*0.00*	*Brugia*	*0.34*	*0.00*	*Trichinella*
Insecta	*0.34*	*0.00*	*0.36*	*0.00*	*−0.08*	*0.04*	**0.64**	**0.00**	**Rhodnius**	*0.30*	*0.00*	*Drosophila*
Ascomycota	**0.61**	**0.00**	*0.54*	*0.00*	*0.23*	*0.00*	*0.46*	*0.00*	*Saccharomyces*	*0.41*	*0.00*	*Schizosaccharomyces*
Basidiomycota	0.09	0.06	0.07	0.09	0.07	0.08	**−0.32**	**0.00**	**Phakopsora**	*0.19*	*0.00*	*Cryptococcus*
Viridiplantae	−0.03	0.51	0.00	0.92	0.06	0.17	**−0.23**	**0.00**	**Chlorella**	*0.12*	*0.00*	*Ostreococcus*
Alveolata	**0.71**	**0.00**	*0.52*	*0.00*	*0.18*	*0.00*	*0.61*	*0.00*	*Tetrahymena*	*0.35*	*0.00*	*Paramecium*
Apicomplexa	*0.20*	*0.00*	*0.12*	*0.00*	*0.09*	*0.02*	**−0.26**	**0.00**	**Perkinsus**	*0.14*	*0.00*	*Cryptosporidium*
Sporozoa	*0.45*	*0.00*	*0.47*	*0.00*	0.00	0.92	**0.49**	**0.00**	**Cryptosporidium**	*0.28*	*0.00*	*Toxoplasma*

23 taxa. With methods that do not allow for missing data, like NYK, gene sampling would be even poorer.

5. Conclusions

We observe that the number and composition of taxa often have a strong impact in reconstructions of the intron evolution. While insignificant for some nodes, such as the Bilateria and Metazoa in our analyses, it can be significant for many others. A stronger influence of taxon sampling is observed in nodes with descendants possessing an intensive intron loss. If such a descendant occupies the basal position, the ancestral intron reconstructions are often unreliable. In the indicated cases, the Csuros algorithm exhibits a trend to systematically overestimate the ancestral intron count. Overestimations also occur with the NYK algorithm, however, under a more complex set of conditions and in our analyses were frequently observed only in the node of Fungi. If a group suffers from massive intron loss, a recommended strategy to improve the accuracy of inferring the intron evolution is to identify and add to dataset a deep-branching member of this group with a high intron density, such as *Perkinsus marinus* in the Apicomplexa.

Acknowledgments

This work was supported by grants of the Russian Foundation for Basic Research (12-04-01716-a, 12-04-91331-NNIO, and 12-04-00154), Deutsche Forschungsgemeinschaft GRK 1563 (RECESS), and the Ministry for Education and Science of the Russian Federation (8089, 8334, and 8820). Special thanks to Dr. Leonid Rusin for helpful discussions.

References

[1] S. W. Roy and W. Gilbert, "Complex early genes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 6, pp. 1986–1991, 2005.

[2] H. D. Nguyen, M. Yoshihama, and N. Kenmochi, "New maximum likelihood estimators for eukaryotic intron evolution," *PLoS Computational Biology*, vol. 1, no. 7, article e79, 2005.

[3] W. G. Qiu, N. Schisler, and A. Stoltzfus, "The evolutionary gain of spliceosomal introns: sequence and phase preferences," *Molecular Biology and Evolution*, vol. 21, no. 7, pp. 1252–1263, 2004.

[4] H. Le Hir, A. Nott, and M. J. Moore, "How introns influence and enhance eukaryotic gene expression," *Trends in Biochemical Sciences*, vol. 28, no. 4, pp. 215–220, 2003.

[5] F. Raible, K. Tessmar-Raible, K. Osoegawa et al., "Evolution: vertebrate-type intron-rich genes in the marine annelid *Platynereis dumerilii*," *Science*, vol. 310, no. 5752, pp. 1325–1326, 2005.

[6] J. C. Sullivan, A. M. Reitzel, and J. R. Finnerty, "A high percentage of introns in human genes were present early in animal evolution: evidence from the basal metazoan Nematostella vectensis," *Genome Informatics*, vol. 17, no. 1, pp. 219–229, 2006.

[7] M. Srivastava, E. Begovic, J. Chapman et al., "The Trichoplax genome and the nature of placozoans," *Nature*, vol. 454, no. 7207, pp. 955–960, 2008.

[8] M. Srivastava, O. Simakov, J. Chapman et al., "The *Amphimedon queenslandica* genome and the evolution of animal complexity," *Nature*, vol. 466, no. 7307, pp. 720–726, 2010.

[9] N. King, M. J. Westbrook, S. L. Young et al., "The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans," *Nature*, vol. 451, no. 7180, pp. 783–788, 2008.

[10] J. E. Stajich, F. S. Dietrich, and S. W. Roy, "Comparative genomic analysis of fungal genomes reveals intron-rich ancestors," *Genome Biology*, vol. 8, no. 10, article R223, 2007.

[11] M. Csurös, I. B. Rogozin, and E. V. Koonin, "Extremely intron-rich genes in the alveolate ancestors inferred with a flexible maximum-likelihood approach," *Molecular Biology and Evolution*, vol. 25, no. 5, pp. 903–911, 2008.

[12] S. W. Roy and D. Penny, "A very high fraction of unique intron positions in the intron-rich diatom *Thalassiosira pseudonana* indicates widespread intron gain," *Molecular Biology and Evolution*, vol. 24, no. 7, pp. 1447–1457, 2007.

[13] T. Y. James, F. Kauff, C. L. Schoch et al., "Reconstructing the early evolution of Fungi using a six-gene phylogeny," *Nature*, vol. 443, no. 7113, pp. 818–822, 2006.

[14] C. W. Dunn, A. Hejnol, D. Q. Matus et al., "Broad phylogenomic sampling improves resolution of the animal tree of life," *Nature*, vol. 452, no. 7188, pp. 745–749, 2008.

[15] H. Philippe, N. Lartillot, and H. Brinkmann, "Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and protostomia," *Molecular Biology and Evolution*, vol. 22, no. 5, pp. 1246–1253, 2005.

[16] K. S. Pick, H. Philippe, F. Schreiber et al., "Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships," *Molecular Biology and Evolution*, vol. 27, no. 9, pp. 1983–1987, 2010.

[17] K. Meusemann, B. M. Von Reumont, S. Simon et al., "A phylogenomic approach to resolve the arthropod tree of life," *Molecular Biology and Evolution*, vol. 27, no. 11, pp. 2451–2464, 2010.

[18] K. M. Kjer, "Aligned 18S and insect phylogeny," *Systematic Biology*, vol. 53, no. 3, pp. 506–514, 2004.

[19] D. Bhattacharya and L. Medlin, "Algal phylogeny and the origin of land plants," *Plant Physiology*, vol. 116, no. 1, pp. 9–15, 1998.

[20] T. R. Bachvaroff, S. M. Handy, A. R. Place, and C. F. Delwiche, "Alveolate phylogeny inferred using concatenated ribosomal proteins," *Journal of Eukaryotic Microbiology*, vol. 58, no. 3, pp. 223–233, 2011.

[21] C. Burge and S. Karlin, "Prediction of complete gene structures in human genomic DNA," *Journal of Molecular Biology*, vol. 268, no. 1, pp. 78–94, 1997.

[22] M. Csuros, J. A. Holey, and I. B. Rogozin, "In search of lost introns," *Bioinformatics*, vol. 23, no. 13, pp. i87–i96, 2007.

[23] M. Csuros, I. B. Rogozin, and E. V. Koonin, "A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes," *PLoS Computational Biology*, vol. 7, no. 9, 2011.