Evolution and function of compositional patterns in mammalian genomes

Ashwin Prakash

The University of Toledo

Follow this and additional works at: http://utdr.utoledo.edu/theses-dissertations

Recommended Citation
http://utdr.utoledo.edu/theses-dissertations/683

This Dissertation is brought to you for free and open access by The University of Toledo Digital Repository. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of The University of Toledo Digital Repository. For more information, please see the repository's About page.
A Dissertation
Entitled
Evolution and Function of Compositional Patterns in Mammalian Genomes
By
Ashwin Prakash

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the Doctor of Philosophy Degree in Biomedical Sciences

_______________________________________
Dr. Alexei Fedorov, Committee Chair

_______________________________________
Dr. Joseph Shapiro, Committee Member

_______________________________________
Dr. Deepak Malhotra, Committee Member

_______________________________________
Dr. Sadik Khuder, Committee Member

_______________________________________
Dr. Robert Trumbly, Committee Member

_______________________________________
Dr. Patricia Komuniecki, Dean
College of Graduate Studies

The University of Toledo
December 2011
An Abstract of
Evolution and Function of Compositional Patterns in Mammalian Genomes
By
Ashwin Prakash

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the
Doctor of Philosophy Degree in Biomedical Sciences
The University of Toledo
December 2011

The protein coding sequences of humans and of most mammals represent less than 2% of their genomes. The remaining 98% is made up of 5'- and 3'-untranslated regions of mRNAs (<2%), introns (~30%), and intergenic regions (~66%). These vast non-protein coding genomic areas, previously frequently referred to as "junk" DNA, contain numerous functional signals of various origin and purpose. The presence of functional information within these vast non-protein coding regions creates function dependent compositional variations. The traditional techniques for mining functional regions have focused mainly on comparative methods where strong sequence similarity of genomic regions between evolutionarily distant species, was deemed to be indicative of a fixation bias in order to maintain functional integrity. In this study we present a novel method of mining functional regions which range in length from 30 nucleotides to several thousands of nucleotides, which we call Mid-Range scale or Mid-Range Inhomogeneity (MRI / genomic MRI). The focus of this study is to demonstrate that at these mid-range scales, genomes of complex eukaryotes consist of a number of different nucleotide compositional patterns which are associated with unusual DNA conformations, RNA secondary structures and non-coding RNA. Some of these patterns are scarcely investigated and still await thorough exploration and recognition.
I would like to dedicate this work to my Father S. V Prakash and Mother, Rekha Prakash, who have sacrificed much to provide me with plenty.
Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 0643542.

I am very grateful to my research advisor, Dr. Alexei Fedorov, for his patient and invaluable guidance and for providing me the opportunity to take a leading role in multiple research projects that broadened my research horizon. Furthermore, the experience and advice from my committee members (Dr. Joseph Shapiro, Dr. Deepak Malhotra, Dr. Sadik Khuder, and Dr. Robert Trumbly) has been most valuable and is deeply appreciated.

I am also indebted to the co-authors of all my manuscripts without whose diligent efforts it would be impossible to get through the stringent review processes of journals. And lastly without my lab-mates and co-workers—Andrew McSweeny, Dave Rearick, Maryam Nabiyouni, Jason Bechtel and especially Samuel Shepard —I could not have endured the stress and strain of graduate school.
### CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td>Contents</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>ix</td>
</tr>
<tr>
<td>List of Figures</td>
<td>x</td>
</tr>
</tbody>
</table>

1. **Chapter 1:** Introduction; Understanding patterns of non-random nucleotide composition
   ........................................................................................................ 1

2. **Chapter 2:** Genomic MRI - a Public Resource for Studying Sequence Patterns within Genomic DNA. (Journ. Of Visualized experiments (JOVE), May 2011)............. 9

3. **Chapter 3:** Evolution of genomic sequence inhomogeneity at mid-range scales.
   (BMC Genomics, November 2009)........................................................................ 22
   3.1. Abstract.................................................................................................. 23
   3.2. Background............................................................................................ 24
   3.3. Results
      3.3.1. Substitution and polymorphism inside MRI regions...................... 25
      3.3.2. Insertions and deletions inside MRI regions............................... 33
   3.4. Discussion............................................................................................. 35
   3.5. Conclusion............................................................................................ 38
   3.6. Methods
3.6.1. Genomic samples and computation of recent human mutations ("fixed
substitutions") .................................................................................. 39
3.6.2. Processing of SNP data ............................................................ 40
3.6.3. Calculation of the substitution ratios in MRI and control regions ...... 43
3.6.4. Calculation of base composition equilibrium for the observed substitution
rates .......................................................................................................... 44
3.6.5. Supplementary materials ............................................................ 46

4. Chapter 4: Database of Mammalian Orthologous Introns of Five Species
(DOMINO5) .......................................................................................... 52
4.1. Abstract .......................................................................................... 53
4.2. Methods
   4.2.1. Characterizing Homologous Proteins ........................................ 54
   4.2.2. Orthologous introns from homologous proteins ....................... 55
   4.2.3. Alignment of orthologous introns ............................................. 56

5. Chapter 5: Association of genomic Mid Range Inhomogeneity with DNA
conformations ....................................................................................... 61
5.1. Introduction .................................................................................... 62
5.2. Results
   5.2.1. Mid range inhomogeneous structures in introns of DOMINO5 ...... 65
   5.2.2. Putative Z-DNA structures within DOMINO5 introns ............... 66
   5.2.3. Putative H-DNA structures within DOMINO5 introns ............... 71
   5.2.4. GC-Rich MRI features within DOMINO5 introns ...................... 77
   5.2.5. GT-Rich MRI features within DOMINO5 introns ...................... 82
   5.2.6. AT-Rich MRI features within DOMINO5 introns ...................... 87
5.3. Methods
   5.3.1. Database of Orthologous Mammalian Introns of 5 species (DOMINO5). 93
   5.3.2. Characterizing Z-DNA features in DOMINO5 ............................ 93
   5.3.3. Characterizing H-DNA features in DOMINO5 ............................ 94
   5.3.4. Characterizing MRI features in DOMINO5 ............................... 94
5.3.5. Calculation of sequence and positional conservation ........................................ 95
5.3.6. Calculating distribution of features within genes ........................................... 95
5.3.7. Graphical representation of the features ........................................................... 96
5.4. Discussion ............................................................................................................ 96

6. Chapter 6: Critical association of ncRNA with introns. (Nucl. Acids Res. 2010) ... 104
6.1. Introduction: Association of Mid Range Inhomogeneity with non coding RNA ................................................................. 105
6.2. Manuscript: Critical association of ncRNA with introns .................................. 110
   6.2.1. Abstract ........................................................................................................ 111
   6.2.2. Introduction ................................................................................................ 111
   6.2.3. Materials and methods
         ● Databases ........................................................................................................ 112
         ● Sequence processing ...................................................................................... 113
         ● Analyzing the distribution of ‘random’ ncRNAs within genomic regions ........ 116
         ● Long intronic ncRNA ..................................................................................... 117
         ● Statistics ........................................................................................................... 118
         ● Programs ........................................................................................................... 118
   6.2.4. Results
         6.2.4.1. snoRNAs, a byproduct of intron splicing in animals ......................... 118
         6.2.4.2. miRNAs are significantly enriched in the transcribed strands of human introns ................................................................. 120
         6.2.4.3. piRNAs are twice as abundant in the transcribed strand of introns as the complementary strand ........................................... 123
         6.2.4.4. Putative endogenous siRNAs within introns ........................................ 125
         6.2.4.5. Long ncRNAs inside introns ................................................................. 127
   6.2.5. Discussion .................................................................................................... 130
   6.2.6. Supplementary data ..................................................................................... 135

7. Source code for relevant programs ................................................................. 145
List of Tables

1. **Table 3.1**: The calculated equilibria percentages for X-bases in X-rich MRI and control regions with average X-composition......................................................... 31

2. **Table 3.2**: Impact of Indels on X-rich MRI Regions, with X Representing Any Single Base.......................................................................................................................... 32

3. **Table 3.3**: Impact of Indels on MRI Regions, with X Representing Combinations of Any Two Bases........................................................................................................ 33

4. **Table 6.1**: Distribution of human pre-miRNAs within exons, introns and intergenic regions.......................................................................................................................... 119

5. **Table 6.2**: The distribution of pre-miRNAs inside introns................................. 121

6. **Table 6.3**: Distribution of human piRNAs within human genome..................... 122

7. **Table 6.4**: The distribution of piRNAs inside introns......................................... 122
List of Figures

1. Figure 2-1. An example of the MRI analyzer graphical output from step 2.5.7..... 15
2. Figure 2-2. MRI analyzer output for the random sequence "userfile.rand1_4"..... 18
3. Figure 2-3. An example of the beginning of a textual output file from MRI
   analyzer........................................................................................................... 19
4. Figure 3-1 Substitution Rates in MRI Regions for a Combination of Nucleotides
   Substitution Rates in MRI Regions for a Combination of Nucleotides............. 27
5. Figure 3-2 Substitution Rates in MRI Regions for Single Nucleotides Substitution
   Rates in MRI Regions for Single Nucleotides............................................. 28
6. Figure 4-1: Identification of Isoform association between two species............. 55
7. Fig 4-2.1: Identification of Orthologous introns between species................... 56
8. Fig 4-2.2: Processing Blast output for homologous proteins......................... 57
9. Figure 5-1: Putative Z-DNA sequences in multiple alignments...................... 66
10. Figure 5-2: Z-DNA structures in introns less than 60000 nucleotides long within
    quintiles......................................................................................................... 67
11. Figure 5-3: Z-DNA structures in all human introns within quintiles.............. 68
12. Figure 5-4: Z-DNA structures by sequence conservation.............................. 69
13. Figure 5-5 Z-DNA structures by preservation of position............................ 70
14. Figure 5-6: Z-DNA structures by preservation of position............................ 70
15. Figure 5-7: Putative HDNA sequences in multiple alignments.................... 72
16. Figure 5-8: HDNA structures in introns less than 60000 nucleotides long within
    quintiles......................................................................................................... 73
17. Figure 5-9: HDNA structures in all human introns within quintiles............. 74
18. Figure 5-10: HDNA structures by sequence conservation............................. 75
19. Figure 5-11: HDNA structures by preservation of position.......................... 76
20. Figure 5.12: HDNA structures by preservation of position......................... 76
21. Figure 5-13: GC-MRI features in multiple alignments................................. 77
22. Figure 5-14: GC-MRI features in introns less than 60000 nucleotides long within quintiles

23. Figure 5-15: GC-MRI features in all human introns within quintiles

24. Figure 5-16: GC-MRI features by sequence conservation

25. Figure 5-17: GC-MRI features by preservation of position

26. Figure 5.18: GC-MRI features by preservation of position

27. Figure 5-19: GT-MRI features in multiple alignments

28. Figure 5-20: GT-MRI features in introns less than 60000 nucleotides long within quintiles

29. Figure 5-21: GT-MRI features in all human introns within quintiles

30. Figure 5-22: GT-MRI features by sequence conservation

31. Figure 5-23: GT-MRI features by preservation of position

32. Figure 5.24: GT-MRI features by preservation of position

33. Figure 5-25: AT-MRI features in multiple alignments

34. Figure 5-26: AT-MRI features in introns less than 60000 nucleotides long within quintiles

35. Figure 5-27: AT-MRI features in all human introns within quintiles

36. Figure 5-28: AT-MRI features by sequence conservation

37. Figure 5-29: AT-MRI features by preservation of position

38. Figure 5-30: AT-MRI features by preservation of position

39. Figure 6-1: Distribution of piRNA along mRNA and introns

40. Figure 6-2: Evolutionarily conserved regions within the third intron of the Heparanase 2 (HPSE2) gene
Chapter One

Understanding Patterns of non random nucleotide composition.
The protein coding sequences of humans and of most mammals represent less than 2% of their genomes. The remaining 98% is made up of 5'- and 3'-untranslated regions of mRNAs (<2%), introns (~30%), and intergenic regions (~67%) [1]. These vast non-protein coding genomic areas, previously frequently referred to as "junk" DNA, contain numerous functional signals of various origin and purpose. They include thousands of non-protein coding RNAs [2], numerous gene expression regulatory signals around genes, chromatin folding structures which include nucleosome positioning sites and scaffold/matrix attached regions [3,4]. These functional DNA regions are non-random in their genomic sequence, the non-randomness or inhomogeneity of base composition exist in varying layers of complexity and sequence length. Prior studies of the informational content of genomes have focused mostly on the extremes: the relative frequency of occurrence of small oligonucleotides in any genome called short-range inhomogeneity (SRI) and on the other hand the isochore structure of most amniote chromosomes which is referred to as long-range inhomogeneity (LRI) [8,9,10]. Compositional inhomogeneity is known to exist in all kinds of species from bacteria to humans. However, the particular arrangement of such sequence patterns is often species-specific [5].

Genomic patterns at short-range scales represent various “words” composed from nucleotide “letters.” Each of these words is repeated several times within the genome. The longest words referred to as “pyknons,” are up to 17-nucleotide-long sequences and are over-abundant in the exons and introns of humans and other mammals [6,7]. Over a trillion different words which are 20 nucleotides long can be formed from the various combinations of the four nucleotides. More than 99% of
these 20-mer oligonucleotides never occur in the entire human genome which only spans about three billion base pairs in length. There have been considerable scientific efforts to study the nature of these short words (<20 nucleotides long) and has led to the discovery of several functional “words” within the genome such as codons, transcription factor binding sites, intron splice signals, gene promoters etc.

Long-range inhomogeneity on the other hand has been described as the "mosaic" structure of eukaryotic chromosomes, referring to the variegated arrangement of the genome in large (on the order of $10^5$-$10^6$ base pairs) sections of relatively uniform nucleotide compositions termed as isochores [11]. The compositional nature of isochores are usually large regions of DNA with a high degree of uniformity in GpC and CpG (collectively G+C content). Isochores have been related to gene density, intron and UTR size, rate of transcription (chromatin structure), rate of recombination, replication timing, G+C content heterogeneity etc [11,12]. It is also known that the genes located in different isochores have different patterns of codon usage biases and since codon usage bias affects the rate of nucleotide substitution, they may evolve at different rates [14].

In this study we consider genomic patterns longer than 30 nucleotides and up to several thousands of nucleotides to be called Mid-Range scale or Mid-Range Inhomogeneity (MRI / genomic MRI). At this mid-range, most of the sequences are unique, i.e. occur only once in the entire genome, hence, it is more appropriate to characterize or group them not by their exact sequence of nucleotides but rather by their overall nucleotide composition, such as G+T-richness, purine-richness, etc. The main focus of this study is to show that at mid-range scales, genomes of complex
eukaryotes consist of a number of different patterns and are associated with unusual DNA conformations, RNA secondary structures and non-coding RNA. Some of these patterns are scarcely investigated and still await thorough exploration and recognition.

An accurate representation would be to view the genome as a “self administered operating system” that calls upon various subroutines embedded within the genome, to perform several biomolecular functions, called upon by internal or external environmental signals. The operating systems become ever so elaborate and complex with the increase in interplay of subroutines, as the organism or its environment becomes more complex. Some of the well-known examples of such a system are seen in neural networks for which mathematical models describing their behavior have been developed. According to Pellionisz [13], “recursive genome function is a process when, at every step of development already-built proteins iteratively access sets of primary and ensuing auxiliary information packets of DNA to build constantly developing hierarchies of protein structures.” In other words, there is a crucial flow of information from DNA, RNA then to proteins and back to the genomic DNA.

This perspective elucidates the importance of studying MRI regions which could be specific sites of genomic functional information. MRI regions are shown here to be intricately associated with unusual DNA conformations, which in turn are binding sites for a number of proteins, which could in turn affect gene expression. These proteins could stabilize and/or initiate DNA conformation transformation and propagate the signal along neighboring DNA segments. For instance, Z-DNA binding proteins could initiate this transformation from right-handed B-DNA to the left-
handed Z-form. This structural transition changes the DNA supercoiling for the regional DNA landscape and additionally creates specific B-Z-boundaries with flipped-over bases. Such transformation could modify, open, and/or hide, some information on the genomic DNA not only at the protein binding site but within neighboring regions. This conformational change could impact the accessibility of transcription factor binding sites, promoter regions, splice sites etc.

Apart from the impact on DNA conformations, MRI regions within spliceosomal introns, could impact RNA secondary structures, alternative splicing of mRNA. The presence of MRI regions can affect the stability of RNA secondary structures which might be associated with several functional non coding RNA (ncRNA) whose function is closely associated with their secondary structure formation. It is also important to note that most ncRNA are functional elements in the genome whose lengths are within the definition of MRI. Thus the study of inhomogeneity at these scales within spliceosomal introns could lead to the discovery of novel ncRNA families within the genome.

A challenge that faces molecular biology is to understand the wealth of information which lies within the non protein coding regions of the genome. Within these vast areas of previously thought “junk DNA,” represented by introns and intergenic sequences, functional regions would exist as an intricate mosaic of MRI regions with extreme base compositions. This study shows that various genomic MRI regions are tightly associated with unusual DNA conformations; RNA secondary structures etc, and thus must be of crucial importance for the functioning of multicellular eukaryotes. Understanding of genomic MRI functions is critical for the newly
emerged field of personal genomics and also for drug discovery.
References


Chapter Two

Genomic MRI - a Public Resource for Studying Sequence Patterns within Genomic DNA

Authors:

Ashwin Prakash, Jason Bechtel, Alexei Fedorov

Department of Medicine, University of Toledo Health Science Campus.
Abstract

Non-coding genomic regions in complex eukaryotes, including intergenic areas, introns, and untranslated segments of exons, are profoundly non-random in their nucleotide composition and consist of a complex mosaic of sequence patterns. These patterns include so-called Mid-Range Inhomogeneity (MRI) regions -- sequences 30 to 10,000 nucleotides in length that are enriched by a particular base or combination of bases (e.g. (G+T)-rich, purine-rich, etc.). MRI regions are associated with unusual (non-B-form) DNA structures that are often involved in regulation of gene expression, recombination, and other genetic processes (Fedorova & Fedorov 2010). The existence of a strong fixation bias within MRI regions against mutations that tend to reduce their sequence inhomogeneity additionally supports the functionality and importance of these genomic sequences (Prakash et al. 2009). Here we demonstrate a freely available Internet resource -- the Genomic MRI program package – designed for computational analysis of genomic sequences in order to find and characterize various MRI patterns within them (Bechtel et al. 2008). This package also allows generation of randomized sequences with various properties and level of correspondence to the natural input DNA sequences. The main goal of this resource is to facilitate examination of vast regions of non-coding DNA that are still scarcely investigated and await thorough exploration and recognition.

Protocol

All the used programs in the paper have been written using perl, and all the web pages have been created using PHP.

Starting Point:
2.1. **Open the home page of the online Genomic MRI package at** 
http://mco321125.meduoio.edu/~jbechtel/gmri/. The web resource also provides instructions/explanations on the programs in the "Help (How-to/README)" link, while all published materials on *Genomic MRI* and similar algorithms are listed in the "Links to relevant resources" link.

2.2. **Preparation and Uploading of Input Sequence(s).** Create a file with FASTA-formatted sequence(s) to start a GMRI analysis session. Each nucleotide sequence in this format should be preceded with a single line starting with the ">" character that represents an identifier, followed on the same line by a short description of this sequence. Nucleotide sequences for GMRI analysis also permits characters like R, Y, N, X, etc. However, non-A, T, C, G characters will not be processed by the program and will be skipped. Sequences in which repetitive elements have been "masked" (replaced by "N"s) can be used as input. Note that sequence characters are case insensitive.

2.2.1. Begin a GMRI session by clicking on the "Start or Resume" button on the *Genomic MRI* home page. This takes the user to a page where nucleotide sequences can be uploaded.

2.2.2. Copy-and-paste your FASTA-formatted sequences or upload a file containing the sequences from your local computer using the "choose file" button.

2.2.3. Click on the "start new session with this file" button. A confirmation message should appear above the input window stating that "Your sequence has been successfully uploaded" and you should also get an alphanumeric "GMRI identifier" [the site calls it a "session label"] for your session (e.g.
b16yMj), which can be used to retrieve and continue a session for up to two weeks after first use.

NOTE: Henceforth the input sequences are referred to as "userfile".

2.3. Get an Oligonucleotide Frequency Distribution of the Input Sequences (optional).

Click on the "SRI Analyzer" tab (top row) in order to get a distribution of oligonucleotide frequencies for the entire set of input sequences. The acronym SRI stands for short-range inhomogeneity. At this juncture, the user may specify the highest length of oligonucleotides (from 2 up to 9 nucleotides, default 6 nts) for which frequencies will be calculated. This selection is made by clicking on the desired option within the "Maximum oligomer size" list box. Then press the "Analyze File" button to initiate computation. A rough representation of the input sequence composition will immediately appear as a short table in the middle of this web page and downloadable as "userfile.comp.tbl". This table represents only the most and the least abundant oligonucleotides within the input sequences. The entire frequency table for all possible oligonucleotides is generated as a file named "userfile.comp", which can be obtained via the "Download composition file" link.

NOTE: SRI analyzer counts the entire set of all overlapping oligonucleotides.

2.4. Generate Random Sequences with the Same Oligonucleotide Composition As in the Input Sequences (optional).

(Completion of step 3 of the protocol is required for this task).

2.4.1. Click on the "SRI Generator" tab (top row) to open up a new web page
that creates random sequences. Choose the number of samples of random sequences to be generated using the list box on this web page. Each of these sample files will contain random sequences of the same number and length as the input sequences in "userfile". Moreover, if an input sequence contains non-A, T, C, or G characters, the random sequence will have "N"s at exactly the same positions as in the input sequence.

2.4.2. Choose the longest length of oligonucleotides for which frequencies will be approximated in the random sequences. This can be chosen by checking the radio button for the desired oligomer level (e.g. "4-mers" for four-base oligonucleotides) in the table at the center of the screen. It is to be noted here that random sequences will consist of not only the approximate frequencies at the chosen oligomer level, but also the corresponding frequencies of shorter oligomer levels, as in the input sequences. Small fluctuations in the oligonucleotide frequencies of input and random sequences are possible due to the Markov Model procedure applied for the generation of random sequences.

2.4.3. Start the program by clicking the "Generate File" button. If the input sequences are large it could take a couple of minutes to generate random sequences. Thus, a user should wait until blue "Download" links appear at the bottom of this page. The random sets are placed in files with names such as "userfile.randX_Y" where X is the number of the random set and Y is the chosen oligomer level (e.g. "userfile_rand2_4").

2.5. Analysis of Mid-Range Inhomogeneity (MRI) of Input and Random Sequences.
2.5.1. Click on the "MRI Analyzer" tab (top row), which opens up a new web page that analyzes the mid-range inhomogeneity of the nucleotide composition of sequences.

2.5.2. Select a sequence to be analyzed from the "File to analyze" list box (a choice between the input sequence and generated sets of random sequences can be made here).

2.5.3. Choose the content type of MRI to be analyzed via the provided list box. (Seven content options are available: G+C; G+A; G+T; A; G; C; or T.)

2.5.4. Choose the length of the window for which content-rich and content-poor sequences will be examined via the "Window size" list box (default is 50 nucleotides; the valid range is from 30 to 1000).

2.5.5. Choose the upper threshold and lower threshold for content-rich and content-poor regions, respectively. These thresholds can be defined by the exact number of particular nucleotides in the current window (using the by number option in the list box) or by percentage of these nucleotides in the window (using the by percentage option).

2.5.6. After all five choices have been made (for example: Sequence = "userfile"; Content = GC; Window size = 50; Upper threshold = 35; Lower threshold = 15), invoke the program by pressing the Analyze File button. The program scans through all sequences from the selected input consecutively. At each step it obtains a segment of the current sequence with length equal to the specified window size and computes whether the number or percentage of nucleotides of the chosen content is above the upper threshold or below the lower threshold. If the window does not match either criteria, the next overlapping window (shifted by one nucleotide) is selected for the same
analysis. When a window is found where the sequence meets one of the threshold requirements for content-rich or -poor composition, the program saves the sequence of this window in the output file and generates a spike on the graphical output. After this, the program jumps to the next non-overlapping adjacent window and resumes the scanning process until the end of the sequence is reached.

2.5.7. After completion of the program, a link to the output file (with name "userfile_GC_50_35..15" for the example above) appears and a graphical representation of the results is displayed in the middle of the web page (see Figure 2-1). On this graphical display all input sequences from the userfile are concatenated into a single string and presented as a horizontal black line on the X axis, with length in kilobases (kb) shown below. All content-rich regions along input sequences are marked as blue "upward" spikes, and content-poor regions as red "downward" spikes. The total numbers of content-rich and content-poor windows are shown in parentheses in the legend at the bottom of this figure (32 and 19, respectively). The figure serves to illustrate the relative abundance and the arrangement of MRI regions. Meanwhile specific details are presented in the output file (see Figure 2-3). In this file, all nucleotide sequence segments that match content-rich or -poor criteria and their coordinates are available to a user as a list according to their consecutive positions along the input file.
The results have been obtained on a sample of 44 human introns. Blue bars represent positions of GC-rich regions along these introns. Red bars represent GC-poor (or AT-rich) MRI regions. The y-axis contains upper and lower thresholds for the given content type.

After completion of MRI analysis for the chosen sequence a user can start a new process in the same web page by making changes to parameters and/or input files. For example, in order to examine the previously generated random sample #1 with the same MRI parameters, the user only needs to change the File to analyze option and select the "userfile_rand1_4" file, and then press the Analyze File button again. A new file and graphical display will replace the old one. The results and figures of ALL examinations under each "session lable" (GMRI identifier) will be saved and be available for two weeks from the last activity. In order to save the results/figures permanently, the user should select the "Download Files" tab (top row) and download the entire session or individual files, as needed.

With this MRI Analyzer web page a user can study

- (G+C)-rich and (A+T)-rich regions
- Purine (A+G)-rich and Pyrimidine (C+T)-rich regions
- Keto (G+T)-rich and aMino (A+C)-rich regions
- A-rich and A-poor regions
- G-rich and G-poor regions
- T-rich and T-poor regions
- C-rich and C-poor regions

2.5.10. The latest release of *Genomic MRI* has a new option for studying regions rich with Purine(R)/Pyrimidine(Y) alternation patterns that might form Z-DNA conformations. Currently, this option is available from the link "Z-DNA" and it works on the same basis as other aforementioned MRI regions. A user should select upper and lower thresholds for the number of (RY+YR) overlapping dinucleotides in the scanning window. The program produces a similar graphical output and a file of DNA segments enriched and depleted by alternating purines and pyrimidines. The putative Z-DNA regions must be highly enriched by alternating R/Y bases (see review F&F 2011).

2.6. **Additional Programs within the Genomic MRI Package (optional).**

The *Genomic MRI* resource also has two advanced options for generation of very specific random sequences. They are available through the "MRI Generator" and "CDS Generator" tabs in the top row.

2.6.1. *MRI generator* creates randomized sequences with the same oligonucleotide composition as the input file (similar to *SRI generator*). However, in addition, randomized sequences mimic a particular MRI pattern specified by the user. Within this web page a user should specify from a list box a particular MRI pattern to be mimicked. The list box contains all patterns that have been examined in this session by *MRI analyzer* (e.g. "userfile_GC_50_35..15"). A random sequence generated with this option will have the same oligonucleotide composition as the selected input file and also the same GC-rich and -poor patterns as seen in "userfile_GC_50_35..15".
2.6.2. \textit{CDS generator} is used for randomization of protein coding sequences. It preserves the same amino acid sequence as the one coded by the user-specified input. In addition the program retains the same codon and di-codon biases as specified in the user-chosen input table. The online version of the \textit{CDS generator} also accepts a protein sequence as an input. All other options for the program are offered only via stand-alone Perl scripts available for download from the main Genomic MRI web page.

2.7. \textbf{Representative Results}

This protocol allows a user to study compositional inhomogeneity of nucleotide sequences. Importantly, it also supports the generation of a variety of randomized sequences with an oligonucleotide composition approximating that of the input sequences. Usually, genomic sequences of complex eukaryotes are not homogeneous in composition, but rather represent a complex mosaic of sequence segments enriched by particular nucleotides (for example, purine-rich, (G+T)-rich, (A+T)-rich, etc.). These patterns at mid-range scale (30-1000 bp) are visualized by the graphical output of \textit{MRI analyzer} that shows selected content-rich segments as upper blue spikes and content-poor segments as lower red spikes (see Figures 2-1 and 2-2). Typically, the number of any content-rich and content-poor regions in a natural sequence (Figure 2-1) is on the order of times higher than the number of the same types of regions in corresponding randomized sequences (Figure 2-2) having the same oligonucleotide composition. These sequence segments with mid-range inhomogeneity in nucleotide composition may be of interest to the user. They are available from the \textit{Genomic MRI} output files for further investigation.
Figure 2-2. MRI analyzer output for the random sequence "userfile.rand1_4". The graphical representation of MRI within a randomly generated sequence using the SRI generator program.

Figure 2-3. An example of the beginning of a textual output file from MRI analyzer. All content-rich and content-poor sequences detected by the program are presented in the last (fourth) column. Their relative positions, measured in the number of windows, are shown in the first column. The second and third columns are indicators for content-rich and content-poor regions, respectively.

Discussion
Regions with inhomogeneous nucleotide composition at mid-range scales (30-1000 nucleotides) are overabundant in the genomes of complex eukaryotes and can be found anywhere (intergenic regions, introns, untranslated regions of exons, repetitive elements). These regions are frequently associated with unusual DNA conformations. For instance, purine-/pyrimidine-rich sequences tend to form DNA triplexes (H-DNA); sequences with alternating purine/pyrimidine bases are associated with Z-DNA conformations; (G+C)-rich regions exhibit structural abnormalities in B-DNA and could be prone to backbone cleavage; (A+T)-rich regions might form an unusual structure - a DNA unwinding element; etc. (reviewed by Fedorov & Fedorova 2010). Some of these mid-range patterns (e.g. (G+T)-rich regions) are scarcely investigated and still await thorough exploration and recognition. The main aim of our Genomic MRI web resource is to help users in the identification of these MRI regions for their further experimental analysis and for exploration of their possible functions. Knowledge of MRI regions could be incorporated into and improve the new generation of gene predictor programs (Shepard 2010) and advance our understanding of genome functions and properties.

Acknowledgements

We are thankful to Samuel Shepard, Peter Bazeley, and John David Bell for the administration of the Genomic MRI web pages. This work was supported by National Science Foundation Career award "Investigation of intron cellular roles" [grant number MCB-0643542].

Materials

- Computer with Internet
- Files with nucleotide sequences for examination.
References


Chapter Three

Evolution of genomic sequence inhomogeneity at mid-range scales

Authors
Ashwin Prakash,# Samuel S Shepard,# Jie He, Benjamin Hart, Miao Chen, Surya P Amarachintha, Olga Mileyeva-Biebesheimer, Jason Bechtel, and Alexei Fedorov

#Contributed equally.

∥Corresponding author.
Abstract

Background

Mid-range inhomogeneity or MRI is the significant enrichment of particular nucleotides in genomic sequences extending from 30 up to several thousands of nucleotides. The best-known manifestation of MRI is CpG islands representing CG rich regions. Recently it was demonstrated that MRI could be observed not only for G+C content but also for all other nucleotide pairings (e.g. A+G and G+T) as well as for individual bases. Various types of MRI regions are 4-20 times enriched in mammalian genomes compared to their occurrences in random models.

Results

This paper explores how different types of mutations change MRI regions. Human, chimpanzee and *Macaca mulatta* genomes were aligned to study the projected effects of substitutions and indels on human sequence evolution within both MRI regions and control regions of average nucleotide composition. Over 18.8 million fixed point substitutions, 3.9 million SNPs, and indels spanning 6.9 Mb were procured and evaluated in human. They include 1.8 Mb substitutions and 1.9 Mb indels within MRI regions. Ancestral and mutant (derived) alleles for substitutions have been determined. Substitutions were grouped according to their fixation within human populations: fixed substitutions (from the human-chimp-macaca alignment), major SNPs (> 80% mutant allele frequency within humans), medium SNPs (20% - 80% mutant allele frequency), minor SNPs (3% - 20%), and rare SNPs (<3%). Data on short (< 3 bp) and medium-length (3 - 50 bp) insertions and deletions within MRI regions and appropriate control regions were analyzed for the effect of indels on the
expansion or diminution of such regions as well as on changing nucleotide composition.

**Conclusion**

MRI regions have comparable levels of de novo mutations to the control genomic sequences with average base composition. De novo substitutions rapidly erode MRI regions, bringing their nucleotide composition toward genome-average levels. However, those substitutions that favor the maintenance of MRI properties have a higher chance to spread through the entire population. Indels have a clear tendency to maintain MRI features yet they have a smaller impact than substitutions. All in all, the observed fixation bias for mutations helps to preserve MRI regions during evolution.

**Background**

The protein coding sequences of humans and of most other mammals represent less than 2% of their genomes. The remaining 98% is made up of 5'- and 3'-untranslated regions of mRNAs (<2%), introns (~37%), and intergenic regions (~60%) [1]. These vast non-protein coding genomic areas, previously frequently referred to as "junk" DNA, contain numerous functional signals of various origin and purpose. They include thousands of non-protein coding RNAs [2], numerous gene expression regulatory signals that surround each gene, chromatin folding structures which include nucleosome positioning sites and scaffold/matrix attached regions [3,4]. These functional DNA regions are non-random in their genomic sequence. The non-randomness or inhomogeneity of base composition has been described at different levels of complexity and sequence length. Starting on the short scale,
inhomogeneity occurs in the non-random associations of neighboring bases with each other [5], through the over and under-abundance of particular "words" (usually 5-10 base long oligonucleotides) [6] or longer stretches of DNA, also known as "pyknons" (~18 bases long) [7,8], and up to large regions that cover hundreds of thousands of nucleotides [9]. Compositional inhomogeneity is known to exist in all kinds of species from bacteria to human. However, the particular arrangement of such sequence patterns is often species-specific [10].

It has been the focus of our research to elucidate the genomic sequence non-randomness that we call Mid-Range Inhomogeneity or MRI [11]. We define MRI to be genomic regions from 30 bp to several thousand nucleotides with particular nucleotide enrichments. For large mammalian genomes, there is a high probability that a random sequence of length 20 nucleotides will be unique. Thus, for examining mid-range genomic signals we do not look at particular "words" but only the overall compositional content of particular base(s) that we refer to as $X$ ($X$ could be a single nucleotide A, G, C, or T or any of their combinations like A+C, or G+T+C). We created a public Internet resource, "Genomic MRI" to study the distribution of $X$-rich regions in any sequence of interest. It was demonstrated that $X$-rich MRI regions are highly overrepresented in mammalian genomes for all kinds $X$-contexts. Particular properties of MRI have also been investigated previously by Mrazek and Kypr [12] and also by Nikolaou and Almirantis [13]. This paper studies the effect of mutations on the evolution of MRI regions in primates.

Results

Substitution and polymorphism inside MRI regions

From the whole-genome human-chimp-macaque alignment we extracted all
the aligned sequences with inhomogeneous nucleotide compositions that satisfy the criteria for MRI (so-called X-rich MRI regions; see the Materials and Methods section) and also control regions with nucleotide compositions equal to the average values for the entire human genome. We used the default MRI region length of 100 nucleotides for all computations. Only SNPs located within these MRI and control regions were studied. We particularly focused on the single nucleotide substitutions that maintain or erode MRI features. For example, in GT-rich MRI regions we counted the total number of novel polymorphisms that erode the feature, i.e. G or T → C or A substitutions, denoted as \(N_{GT\rightarrow CA}\) and also the total number of those that maintain the MRI features, i.e. C or A → G or T substitutions, denoted as \(N_{CA\rightarrow GT}\). In addition, the entire set of recent human substitutions; that is, those nucleotides that differed in human but were the same in chimp and macaque, were processed for the MRI and control regions and presented as "fixed substitutions". The substitution ratio, \(S_X\) (recall that: \(S_X = \frac{N_{X\rightarrow nonX}}{N_{nonX\rightarrow X}}\)) for the numbers of substitutions that maintain and erode X-rich MRI features was calculated for each substitution subtype (rare, minor, medium, and major SNPs and 'fixed'--refer to the Methods section for a detailed explanation) and presented in Figures Figures 3-1 and 3-2. With respect to X-rich or poor MRI regions, the \(X\) in Figure Figure 3-1 represents a two base combination such as GC, AG, GT, etc. while in Figure 3-2 can be any single nucleotide, e.g., A, T, C, and G. If the \(S_X\)-ratio is equal to 1 the X-rich region does not tend towards a change in its X-base composition. When \(S_X > 1\), the substitutions reduce the X-richness of the examined regions, whereas when \(S_X < 1\), substitution rates elevate the X-richness of the regions. Figures Figures 3-1 and 3-2 demonstrate clear linear trends for \(S_X\)-ratios with respect to increasing fixation of substitutions within human populations.
**Figure 3-1 Substitution Rates in MRI Regions for a Combination of Nucleotides**

Substitution Rates in MRI Regions for a Combination of Nucleotides. For each X MRI region—where X is for GC-, GT-, or GA-rich or poor regions—the X-base composition rate of change is given for all substitutions at different levels of fixation within the human population. The rate of change ($S_X$) is the ratio of X to nonX substitutions over nonX to X substitutions in those particular X-rich regions. Thus, a ratio of 1 means no change in the X-richer of the region whereas a ratio greater than 1 implies degradation of the X-rich region and less than 1 implies enrichment of the X-rich MRI region. Note that in the control X-average regions the $S_X$-ratio is always inverse to $S_{nonX}$-ratio ($S_X = 1/S_{nonX}$). Therefore, only one graph for each $S_X$ and $S_{nonX}$ pair is presented. Since there are significant variations in $S_X$-ratios for different X compositions, the graphs are presented in two different scales. The white
background presents changes of $S_X$-ratios in the 0.8 to 2 range, while the gray background presents changes in the 0 to 7 range. Vertical bars show the standard error of the means (see Methods section).

Figure 3-2 Substitution Rates in MRI Regions for Single Nucleotides Substitution

Rates in MRI Regions for Single Nucleotides. For each X MRI region—where X is for A-, T-, G-, or C-rich or poor regions—the X-base composition rate of change is
given for all substitutions at different levels of fixation within the human population. The rate of change ($S_X$) is the ratio of $X$ to non$X$ substitutions over non$X$ to $X$ substitutions in those particular $X$-rich regions. Thus, a ratio of 1 means no change in the $X$-richness of the region whereas a ratio greater than 1 implies degradation of the $X$-rich region and less than 1 implies enrichment of the $X$-rich MRI region. Note that in the control $X$-average regions the $S_X$-ratio is always inverse to $S_{\text{non}X}$-ratio ($S_X = 1/S_{\text{non}X}$). Therefore, only one graph for each $S_X$ and $S_{\text{non}X}$ pair is presented. Since there are significant variations in $S_X$-ratios for different $X$ compositions, the graphs are presented in two different scales. The white background presents changes of $S_X$-ratios in the 0.8 to 2 range, while the gray background presents changes in the 0 to 7 range. Vertical bars show the standard error of the means (see Methods section).

For the cases of GT-, AC-, AG-, and TC-rich MRI regions (Figure 3-1), all $S$-ratios for rare SNPs are close to 1.8 (showing erosion of the MRI features). For major SNPs and fixed mutations the $S_{\text{GT}}$ and $S_{\text{AC}}$-ratios reach 1.0 (which means no change in the corresponding base composition) and $S_{\text{GA}}$ and $S_{\text{CT}}$-ratios reach 1.2 respectively. As for the corresponding control GT-, AC-, AG-, and TC-average regions (all having 50% of corresponding base composition) these lines are flat with all $S$-ratios equal to 1. The latter result is highly expected because of the symmetry of (+) and (-) chromosomal strands for these particular base compositions. Figure 3-1 also demonstrates that in GC-rich MRI regions the $S_{\text{GC}}$-ratio change has the highest slope from 7.0 for rare SNPs to 1.6 for fixed substitutions. In AT-rich MRI regions (also referred to as nonGC-rich in the tables) the change of $S_{\text{GT}}$-ratio has the lowest slope starting from 1.7 (rare SNPs) and ending at 1.3 (fixed substitutions). The control regions with the average GC/AT compositions (40-42% GC and 58-60% AT) also
demonstrate a clear change of S-ratios during substitution fixation. In the control GC-average regions, rare SNPs favor increasing AT-richness ($SGC$-ratio of 1.3) whereas fixed mutations demonstrate the opposite effect ($SGC$-ratio of 0.8). The data for the S-ratios for single nucleotides (Figure 3-2) are very similar to the trends seen in GC- and AT-rich regions. As expected from (+/-) strand symmetry, $S_{G}$-ratios are equal to $S_{C}$-ratios and represent about a half of the GC trend. The minor differences between G- and C-rich regions are within the errors of measurement. In the same way the $S_{T}$-ratios are seen to be the same as the $S_{T}$-ratios and they comprise approximately half of the effect seen for AT-rich regions.

Based on the observed $S_{X}$-ratios and the current percentage of $X$ bases in the genomic regions under investigation, we calculated the projected equilibrium composition representing the future $X$-composition toward which the examined substitution rates drive these regions. In other words, the equilibrium $X$-composition shows the future level of $X$-richness that would be approached if the $S_{X}$-ratio as it is observed now were maintained indefinitely. The computed equilibria for each subgroup of substitutions are presented in Table 3.1. For instance, in the GA-rich regions (G+A composition of 70%), rare SNPs drive GA-richness of these MRI regions down to an equilibrium of 56.6%, while nearly fixed or fixed substitutions drive the GA-composition only to the 65.8% level. For each type of $X$-rich MRI, there is a trend toward minimizing the damage of mutations and preserving the MRI feature as the fixation of the observed substitutions increases. The highest preservation effect is seen for GT- and AC-rich regions (with an observed $X$-base composition of 70%), where the equilibria for fixed substitutions reach about the same level of 70%. For the rest of the types of MRI regions, their equilibria composition is a little below the currently observed base composition.
Table 3.1: The calculated equilibria percentages (see Equation 3) for X-bases in X-rich MRI and control regions with average X-composition. Projected equilibria are given based on the substitution rates of rare, minor, medium, and major SNPs as well as for the fixed substitution rates (chimp-macaque to human).

In order to estimate mutation rates for MRI regions versus their respective control regions, we counted the occurrence rates for rare SNPs. The frequency ratio of rare SNPs in MRI rich regions to those in the control regions was calculated. The smallest ratio observed was for A+C content (0.464). This means the frequency of rare SNPs within MRI AC-rich regions is approximately half that of control regions. The highest occurrence ratio for rare SNPs was observed in G- and C-rich MRI regions (1.16 and 1.17 respectively). Thus, the occurrence rates of rare SNPs is slightly lower in MRI regions than in the corresponding control regions with the exception of G- and C-rich MRI regions. The entire dataset for the SNPs occurrences
in MRI and control regions is presented in Additional file 1. The prevalence of rare and minor SNPs over major SNPs was also observed, their proportion over every MRI and control regions being 5.79.

<table>
<thead>
<tr>
<th></th>
<th>A-rich</th>
<th>nonA-rich</th>
<th>A-average</th>
</tr>
</thead>
<tbody>
<tr>
<td>total length</td>
<td>66.9 Mb</td>
<td>72.4 Mb</td>
<td>800.4 Mb</td>
</tr>
<tr>
<td>content of A</td>
<td>49.6%</td>
<td>12.9%</td>
<td>30.3%</td>
</tr>
<tr>
<td>ΔA</td>
<td>-16850</td>
<td>-7390</td>
<td>-44182</td>
</tr>
<tr>
<td>ΔnonA</td>
<td>-24748</td>
<td>-29257</td>
<td>-98769</td>
</tr>
<tr>
<td>net A% change INDEL</td>
<td>0.006%</td>
<td>-0.004%</td>
<td>-0.0001%</td>
</tr>
<tr>
<td>net A% change SUBST</td>
<td>-0.0177%</td>
<td>-0.0022%</td>
<td>-0.0014%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>T-rich</th>
<th>nonT-rich</th>
<th>T-average</th>
</tr>
</thead>
<tbody>
<tr>
<td>total length</td>
<td>67.8 Mb</td>
<td>71.1 Mb</td>
<td>800.4 Mb</td>
</tr>
<tr>
<td>content of T</td>
<td>49.5%</td>
<td>13.1%</td>
<td>30.3%</td>
</tr>
<tr>
<td>ΔT</td>
<td>-21849</td>
<td>-7078</td>
<td>-47238</td>
</tr>
<tr>
<td>ΔnonT</td>
<td>-24084</td>
<td>-22716</td>
<td>-97057</td>
</tr>
<tr>
<td>net T% change INDEL</td>
<td>0.001%</td>
<td>-0.004%</td>
<td>-0.0004%</td>
</tr>
<tr>
<td>net T% change SUBST</td>
<td>-0.024%</td>
<td>-0.0022%</td>
<td>-0.0013%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>G-rich</th>
<th>nonG-rich</th>
<th>G-average</th>
</tr>
</thead>
<tbody>
<tr>
<td>total length</td>
<td>52.0 Mb</td>
<td>60.4 Mb</td>
<td>884.7 Mb</td>
</tr>
<tr>
<td>content of G</td>
<td>40.10%</td>
<td>7.20%</td>
<td>20.40%</td>
</tr>
<tr>
<td>ΔG</td>
<td>-1185</td>
<td>-7080</td>
<td>-31780</td>
</tr>
<tr>
<td>ΔnonG</td>
<td>-12064</td>
<td>-37512</td>
<td>-139126</td>
</tr>
<tr>
<td>net G% change INDEL</td>
<td>0.009%</td>
<td>-0.006%</td>
<td>0.003%</td>
</tr>
<tr>
<td>net G% change SUBST</td>
<td>-0.052%</td>
<td>0.009%</td>
<td>0.016%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>C-rich</th>
<th>nonC-rich</th>
<th>C-average</th>
</tr>
</thead>
<tbody>
<tr>
<td>total length</td>
<td>52.0 Mb</td>
<td>60.4 Mb</td>
<td>883.9 Mb</td>
</tr>
<tr>
<td>content of C</td>
<td>40.10%</td>
<td>7.20%</td>
<td>20.50%</td>
</tr>
<tr>
<td>ΔC</td>
<td>-829</td>
<td>-6700</td>
<td>-33823</td>
</tr>
<tr>
<td>ΔnonC</td>
<td>-12418</td>
<td>-35777</td>
<td>-140331</td>
</tr>
<tr>
<td>net C% change INDEL</td>
<td>0.009%</td>
<td>-0.006%</td>
<td>0.003%</td>
</tr>
<tr>
<td>net C% change SUBST</td>
<td>-0.049%</td>
<td>0.009%</td>
<td>0.015%</td>
</tr>
</tbody>
</table>

Table 3.2: Impact of Indels on X-rich MRI Regions, with X Representing Any Single Base. The impact of indels on X-rich MRI regions and on X-average regions, where X is for A-, T-, C-, or G-rich or poor. For each particular region we give the total length of examined regions in mega-bases, the percentage composition or content of X, the number of changes in X due to insertions and deletions (ΔX = Nins(X) - Ndel(X)), and the change in X composition due to both indels and
substitutions.

<table>
<thead>
<tr>
<th></th>
<th>GC-rich</th>
<th>nonGC-rich</th>
<th>GC-average</th>
</tr>
</thead>
<tbody>
<tr>
<td>total length</td>
<td>17.8 Mb</td>
<td>54.8 Mb</td>
<td>780.6 Mb</td>
</tr>
<tr>
<td>content of GC</td>
<td>71.00%</td>
<td>20.30%</td>
<td>40.90%</td>
</tr>
<tr>
<td>ΔGC</td>
<td>1405</td>
<td>-9100</td>
<td>-31622</td>
</tr>
<tr>
<td>ΔnonGC</td>
<td>-765</td>
<td>-5951</td>
<td>-56278</td>
</tr>
<tr>
<td>net GC% change INDEL</td>
<td>0.005%</td>
<td>-0.011%</td>
<td>0.001%</td>
</tr>
<tr>
<td>net GC% change SUBST</td>
<td>-0.094%</td>
<td>0.042%</td>
<td>0.034%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>GT-rich</th>
<th>nonGT-rich</th>
<th>GT-average</th>
</tr>
</thead>
<tbody>
<tr>
<td>total length</td>
<td>34.9 Mb</td>
<td>34.6 Mb</td>
<td>1192.1 Mb</td>
</tr>
<tr>
<td>content of GT</td>
<td>69.10%</td>
<td>30.90%</td>
<td>50.00%</td>
</tr>
<tr>
<td>ΔGT</td>
<td>-8278</td>
<td>-6817</td>
<td>-121644</td>
</tr>
<tr>
<td>ΔnonGT</td>
<td>-4518</td>
<td>-3502</td>
<td>-120128</td>
</tr>
<tr>
<td>net GT% change INDEL</td>
<td>0.002%</td>
<td>-0.006%</td>
<td>-0.0001%</td>
</tr>
<tr>
<td>net GT% change SUBST</td>
<td>0.004%</td>
<td>0.001%</td>
<td>-0.0003%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>GA-rich</th>
<th>nonGA-rich</th>
<th>GA-average</th>
</tr>
</thead>
<tbody>
<tr>
<td>total length</td>
<td>69.2 Mb</td>
<td>70.0 Mb</td>
<td>978.3 Mb</td>
</tr>
<tr>
<td>content of GA</td>
<td>69.75%</td>
<td>30.22%</td>
<td>49.99%</td>
</tr>
<tr>
<td>ΔGA</td>
<td>-23641</td>
<td>-13935</td>
<td>-96617</td>
</tr>
<tr>
<td>ΔnonGA</td>
<td>-14185</td>
<td>-28480</td>
<td>-100013</td>
</tr>
<tr>
<td>net GA% change INDEL</td>
<td>0.004%</td>
<td>-0.002%</td>
<td>0.0002%</td>
</tr>
<tr>
<td>net GA% change SUBST</td>
<td>-0.014%</td>
<td>0.014%</td>
<td>0.0002%</td>
</tr>
</tbody>
</table>

Table 3.3: Impact of Indels on MRI Regions, with X Representing Combinations of Any Two Bases. The impact of indels on X-rich MRI regions and on X-average regions, where X is for GC-, GT-, or GA-rich or poor. For each particular region we give the total length of examined regions in mega-bases, the percentage composition or content of X, the number of changes in X due to insertions and deletions (ΔX = Nins(X) - Ndel(X)), and the net change in X composition due to both indels and substitutions.

Insertions and deletions inside MRI regions

Using the same computational approach as for substitutions, we analyzed human-chimp-macaque triple alignments for the characterization of *indels* (insertions & deletions) that occurred in the human genome during the last 10 million years after the divergence of *H. sapiens* and *P. troglodytes* species. We particularly investigated
how indels change the nucleotide composition of MRI regions and control regions with average nucleotide composition. The complete set of data representing short indels (whose sizes are less than three nucleotides) and medium indels (whose sizes are from three to fifty nucleotides) is presented in Additional file 2 (S2A--contains MRI for combinations of nts; S2B--MRI for individual nts). Large indels with sizes over 50 bp were not examined since they are comparable with the sizes of MRI regions and, thus, compromise proper characterization of MRI. The summary data on the influence of both short and medium indels on the composition of MRI and control regions are presented in the Tables 3.2 and 3.3 (Table 3.2 shows MRI regions where $X$ represents any single nucleotide; Table 3.3 is for when $X$ represents any combination of 2 nucleotides). For each type of $X$-rich and $X$-control regions the total number of inserted and deleted $X$ and non-$X$ nucleotides have been computed: $N_{ins}(X), N_{ins}(\text{non}X), N_{del}(X), N_{del}(\text{non}X)$. Finally, the net change in $X$ and non-$X$ compositions due to indels have been calculated using the following formulas:

$$\Delta X = N_{ins}(X) - N_{del}(X)$$

Tables 3.2 and 3.3 demonstrate that in the human genome there is a prevalence of deletions over insertions (i.e. negative values of $\Delta X$ and $\Delta \text{non}X$) for every type of nucleotide content studied and for every type of MRI and control region with the exception of GC-indels in GC-rich MRI regions. In the last case $\Delta GC$ is positive and equal to 1405 added nucleotides (over a total set of 1.8 million nucleotides). For all other cases of $X$ except $X = \text{GC}$, short and medium indels cause gradual contraction of genomic regions in humans. This means that there is no nucleotide composition equilibrium to which the indels drive the genome in the indefinite future and, therefore, these equilibria have not been calculated. Table Table22 shows that, for every $X$-rich region, indels result in the increasing the richness of corresponding MRI
regions (positive net \(X\)% change for \(X\)-rich region and negative net \(X\)% change for non\(X\)-rich region). In all \(X\)-control regions the net \(X\)% change is several times less than in the corresponding \(X\)-rich and non\(X\)-rich regions.

Finally, we calculated the percentage of nucleotide composition changes in case of both substitutions and indels separately, that occurred in the human genome during last ten million years after the divergence of human and chimpanzee. These results are presented in Tables 3.2 and 3.3 and serve to measure the relative importance of substitutions versus indels to the nucleotide composition of MRI regions.

**Discussion**

Consistent with Chargaff's second parity rule [14], both the G or C base content of the human genome are equal to 21.1%, while A or T comprise 28.9% each. However, in thousands and thousands of genomic regions of various lengths, the composition of A, T, C, or G content (or different combinations of these bases) exist at extremes quite different from the aforementioned averages. De novo mutations constantly occur in populations and could dramatically change the base composition of a genomic region during the course of evolution. A good choice for a large-scale computational analysis of these novel mutations is in the examination of 'rare' single-nucleotide polymorphisms (SNPs, or mutations that are present only in a small group of individuals and absent in a majority of the population). Rare SNPs are mutations that have recently occurred. However, even among rare SNPs there exists a minor subgroup of "older" mutations that have diminished their frequency to rare events. The relative size of this subgroup is in reverse proportion to the effective size of the population [15], and hence, it represents only a minor fraction of the recent mutations
Here we show that rare SNPs in genomic regions with average nucleotide composition are enriched by G or C → T or A substitutions that drive the genomic composition of those regions to a level of 35% for G+C and 65% for A+T. On the other hand, examining the same regions for mutations that have substantially propagated into human populations (i.e. medium and high frequency SNPs as well as "fixed" recent mutations) demonstrates that these fixed or nearly fixed substitutions are much less prone to G or C → T or A changes. Instead, high frequency SNPs as well as fixed substitutions tend to drive genomic regions with average base composition to 45% G+C composition.

Here we have focused particularly on the influence of mutations on the evolution of specific genomic regions with strongly inhomogeneous base compositions that are far from the average distribution of nucleotides (so-called MRI regions where G+C, G+A, C+T, G+T, or A+C composition is at least 70%, A+T composition is above 80%, or single base frequency reaches nearly 50%). For all types of MRI regions, we found that novel substitutions (rare SNPs) tend to more strongly erode the compositional extremes (X-richness) of the region. At the same time, these mutations undergo a strong fixation bias during their propagation into populations in such a way that fixed substitutions tend to preserve MRI regions. For example, rare SNPs inside GC-rich MRI regions drive the nucleotide composition of those regions to the 26% GC level. However, fixed substitutions in the same GC-rich MRI regions drive GC composition only to 61%. The highest fixation was seen for GT- and AC-rich MRI regions, which preserves the current GT- and AC-composition of 70%.

This trend of preserving nucleotide composition of MRI regions with respect to the increasing fixation of substitutions could be explained by at least two different
mechanisms. First, one could observe that there are some important functional roles for MRI regions. For instance, GC-rich MRI regions include well-known CG-islands, prominent regulators for gene expression [16,17]. Thus, these regions should be under the constraint of purifying selection, preserving their important features. Other MRI regions may be under similar selective pressure due to association with functional genomic elements and/or, as yet unknown, sequence signals. Second, fixation bias inside MRI regions might be due to some non-symmetry in cellular molecular machinery involving DNA repair, replication, and/or recombination processes. For example, the Biased Gene Conversion (BGC)-theory engages this particular scenario in order to explain the maintenance of CG-rich regions [18,19]. (It must be observed, however, that this theory operates on much larger genomic scales and refers to isochores that cover from hundreds of thousands to millions of bases.) Thus far it is inconclusive as to which of these two scenarios, or a combination thereof, best fits the observed trends. For the case of GC-rich sequences, we conjecture that both scenarios could be taking place to some extent to preserve MRI.

Interestingly, the highest level of MRI erosion for rare SNPs is observed in GC-rich MRI regions. Novel substitutions in these particular regions try to drive GC-content to the lowest level of 26% (see Table 1). We explain this phenomenon via uneven distribution of CpG dinucleotides, which are most abundant in GC-rich MRI regions. It is well known that CpG dinucleotides are extreme hot spots for the C → T and G → A mutations, which cause CpG to be the most underrepresented dinucleotide in vertebrate genomes. Therefore, CG-rich MRI regions, which are known to have the highest concentration of CpG dinucleotides, should have the highest rate of de novo mutations in the direction C or G → T or A. Human SNPs having C/T alleles in the CpG/TpG context with the orthologous chimp allele in the
TpG context have an increased error rate of 9.8% for ancestral misidentification (see the Methods section) due to the probability of a coinciding chimp SNP at the same locus [20]. However, since the strength of the mutational erosion in the GC-rich MRI regions is so high, even an error rate of 9.8% will not change the observed trend. So far we have discussed only the effect of substitutions on the nucleotide composition of mid-range genomic regions. Insertions and deletions are the other types of mutations that change genomic sequences and, therefore, should also be considered. In mammals, short and medium indels are several times less frequent than substitutions. Currently, there is not enough data on human indel SNPs to perform the same analysis of their fixation process as we did for substitutions. For this reason we studied only fixed indels in humans (indels present in human but differing in chimp and macaque). Our examination demonstrated that indels weakly influence the nucleotide content of MRI regions toward preserving their inhomogeneous composition, in the same manner as the fixation bias of fixed substitutions (see Tables 22 and 33).

**Conclusion**

The fixation bias on both fixed substitutions and indels tend to protect MRI regions from degradation of their compositional extremes amid the constant flow of random mutations, thus suggesting their contribution in the preservation of functional and structural complexities of the human genome. Future research on these genomic elements as well as refinement of our approach should help determine the extent of maintenance of MRI by natural selection.

**Methods**
Genomic samples and computation of recent human mutations ("fixed substitutions")

Taking human-chimp (human build 36.1 and chimp build 2 version 1) and human-macaque (macaque build v1 edit4) whole-genome pairwise alignments from the UCSC Genome Browser [21] http://hgdownload.cse.ucsc.edu/downloads.html as input, we generated a Perl script for the identification of the common genomic regions for these three species. The process involved the usage of the human genomic sequence as the reference for the location with the chimp and macaque sequences being extracted only in areas where the sequences of all three species were represented. We then invoked the ClustalW (v1.83) program with default parameters to obtain a whole-genome human-chimpanzee-macaque triple alignment. The obtained alignment is available at our website http://bpg.utoledo.edu/human_chimp_macaque.html. This triple alignment was used to calculate the dataset of recent mutations in humans. We considered a recent substitution at a particular position (for example T → C at position 23456719 on chromosome 7) to be valid if the human genome has a C base while both chimp and macaque have a T base in the corresponding aligned positions. In addition, we required that the quality of the alignment in the vicinity of the mutation be reliable (more than 70% similarity between human and macaque in the 20 bp flanking region [-10, +10]). The frequency table of all inferred recent human mutations is presented in the Additional file 3. We analyzed these recent substitutions together with the SNP datasets and call the former mutations "fixed substitutions," assuming that the majority of them occurred less than 10 million years ago and were already fixed across all human populations. In the same manner we processed indels in the triple
alignments and computed all unambiguous cases of human insertions and deletions with sizes from 1 to 49 nucleotides.

**Processing of SNP data**

Over 4.62 million human SNPs from all chromosomes were obtained (dbSNP build 128 [22], ftp://ftp.ncbi.nih.gov/snp/), filtered for completeness and correctness annotations (676499 records discarded total), and mapped onto the whole-genome human-chimpanzee alignment. SNP allele frequencies were averaged from the frequency data of all populations of that allele. However, only those SNPs that were successfully located within the alignment were processed further. For each SNP site we verified the existence of the particular polymorphic bases in the specified position of the human genome reference sequence and also in the corresponding aligned position on the chimp genomic sequence. If any of these two species had different bases than the SNP alleles, the SNP was discarded (20469 SNPs discarded total). Otherwise, we defined the origin of the polymorphism based on the chimpanzee nucleotide. Consider the following example to illustrate this process: suppose one has an A/G polymorphism located at position 34567812 of chromosome 5 with an average A allele frequency of 0.6 and a G allele frequency of 0.4. Then at position 34567812 of chromosome 5 of the human genome reference sequence (Genbank build 36.1), we would first examine if the A or G allele is present at that position and discard the SNP if not. Next, using the flanking region of that SNP we could align the chimp genomic sequence. If the chimp nucleotide were T or C then the SNP would also be discarded because those alleles are not a part of the human haplotype at that position. However supposing that the chimp nucleotide were G, then the polymorphism would be declared as a G → A polymorphism with G being declared the ancestral allele that at
some point in human evolution mutated into an A allele within some human population(s). From the frequency data we may finally characterize this example SNP more precisely as a 0.4G → 0.6A polymorphism.

Using this approach we successfully characterized 3.93 million human SNPs. This last group of SNPs was divided into four subgroups based on the abundance of the mutant allele in the given human populations:

i. Rare polymorphisms with the frequency of the mutated allele being less than 3%;

ii. Minor polymorphisms with frequencies ranging from 3% to 20%;

iii. Medium polymorphisms with frequencies going from 20% to 80%;

iv. Major polymorphisms with the frequency being above 80%.

For our method, misidentification of the ancestral allele might arise when the site for the human SNP is also polymorphic in chimp populations (e.g. A/G polymorphism) or for the possible case that this site had a recent substitution in chimps (A → G) after their divergence from humans. Human and chimpanzee genomes only differ by 1.23% due to single nucleotide substitutions with 1.06% being due to fixed substitutions and the rest (0.17%) being due to polymorphisms in human and chimp [20]. Moreover, according to the Chimpanzee Sequencing and Analysis Consortium the average estimated error rate of human alleles being misidentified due to chimp polymorphisms is only ~1.6% across all typical SNPs, which is acceptably low. It is also observed, however, that in the mutational hotspot of the CpG dinucleotide, there is an increased error rate for ancestral misidentification. If the human alleles are C/T in the CpG and TpG context and the chimp allele is T (in the TpG context) then the estimated error rate is actually 9.8% [20]. Thus, in the context
of studying our MRI regions, any substitution (especially in GC-rich MRI regions since they contain an overabundance of G and C) going from TpG → CpG could have the ancestral allele misidentified, which would mean that the substitution would actually be CpG → TpG, although in the case of GC-rich MRI regions where such dinucleotides are more likely, an error rate of 9.8% is not sufficient to change the trend or conclusion of our results.

**X-rich MRI genomic regions and control regions with average base composition**

Any base or combination of bases can be described by a parameter \(X\). For example, \(X\) could be G-base; C+T-bases; or A+T+G bases, et cetera. It is also useful to refer to \(\text{non}X\) base(s) as all bases not \(X\). Thus, \(X + \text{non}X\) must represent all four nucleotides A, G, T, and C. For the examples above, \(\text{non}X\) are A+T+C-bases; G+A-bases, and C-base, respectively. MRI is characterized by a specific base composition within a region under analysis. We characterize \(X\)-rich MRI regions based on an overabundance of the \(X\) base(s) within a region of a certain length (the so-called window), where the percentage of \(X\) should be above a certain threshold (Bechtel et al 2008). We calculated MRI regions in the human genome for single nucleotides and various nucleotide combinations using a stretchy window of 100+ nucleotides with the following threshold parameters: for A or T the threshold was 49%; for G or C we used 40%; for G+C it was 70%; for A+T the threshold was at 80%; for G+T, C+A, G+A, and C+T were at 70%; nonA or nonT was 87%; and non G or non C the threshold was 93%. These thresholds were chosen experimentally in such a way that MRI regions should represent about 2% of the whole human genome. A stretchy window of \(N+\) nucleotides means that we scan genomic sequence with an \(N\)-size window to find a genomic MRI region that fits the threshold criterion, then we extend the window above the detected region by 10 nt steps until the criterion is no longer
met. After registering the full MRI region we jump beyond the current MRI region and continue with the default N-size window. Using this approach we characterized all MRI regions in the triple human-chimp-macaque alignments using the human sequence for calculating nucleotide composition and MRI features. We also discarded those MRI regions in the alignments where the indel composition exceeded 50%. For the collection of control regions with average base compositions we used the same stretchy window approach with the nucleotide composition corresponding to the following average genomic frequencies: for A, T between 30 and 31% thresholds; for G, C between 20-21%; for G+C between 40-42%; A+T at 58-60%; G+T, C+A, G+A, or C+T were at 49-51%.

Note that control regions with genome-average X-composition also have genome-averaged nonX-composition. Therefore, their substitution ratios are in inverse proportion to each other: $S_X = 1/S_{nonX}$. Due to this only one ratio for X and nonX pair is shown in Figures Figures11 and and22.

**Calculation of the substitution ratios in MRI and control regions**

Studying SNPs and fixed substitutions in X-rich MRI regions we measured the number of changes from X to nonX (denoted as $NX \rightarrow nonX$) and also the number of changes from nonX to X (denoted as $N_{nonX} \rightarrow X$). The fluctuations in the observed distribution of $NX \rightarrow nonX$ and $N_{nonX} \rightarrow X$ are well-known as Poisson noise. Thus, the standard deviation for the true values for $NX \rightarrow nonX$ and $N_{nonX} \rightarrow X$ is calculated according to the Poisson distribution, that is: $\sigma N = \sqrt{N}$. For each X-rich MRI region we measured the substitution $S_X$-ratios with $S_X = NX \rightarrow nonX/N_{nonX} \rightarrow X$ shown in Figures Figures11 and and22. The propagation of uncertainty for a ratio $f = A/B$ can be calculated using the formula $(\sigma f/f)^2 = (\sigma A/A)^2 + (\sigma B/B)^2 - 2(\sigma A \cdot \sigma B)/(A \cdot B) \cdot \rho_{AB}$,
where $\rho_{AB}$ is the correlation coefficient for $A$ and $B$ variables. Because the observed frequency of having a SNP at a genomic site in humans is less than 1%, it is correct to assume that the correlation between $NX\rightarrow nonX$ and $NnonX\rightarrow X$ is negligible. Therefore, the standard deviation for the $SX$ ratio was calculated by the following formula:

$$\sigma = \sqrt{\frac{N_X}{N_{nonX}} \sqrt{\frac{1}{N_X} + \frac{1}{N_{nonX}}}}$$

**Calculation of base composition equilibrium for the observed substitution rates**

As described in the previous paragraphs, for studying SNPs and fixed substitutions in $X$-rich MRI regions of the human genome we measured the number of changes from $X$ to $nonX$ (denoted as $NX\rightarrow nonX$) and also the number of changes from $nonX$ to $X$ (denoted as $NnonX\rightarrow X$). These $NX\rightarrow nonX$ and $NnonX\rightarrow X$ helped us to estimate the frequencies of these two types of mutations per $X$ or $nonX$ site, named here as $FX\rightarrow nonX$ and $FnonX\rightarrow X$, correspondingly. Suppose one has a sample of MRI regions with a total nucleotide sequence length of $L$ and a composition of $X$ with the region richness given as $PX$ being measured in numbers from 0 to 1. Then, the total number of $X$ sites in this sample will be $L \cdot PX$, and the total number of $nonX$ sites will be $L \cdot (1 - PX)$. During a certain time interval called $\Delta T$ there will be $\Delta NX\rightarrow nonX$ and $\Delta NnonX\rightarrow X$ substitutions. Therefore the frequency of substitutions per site is $FX\rightarrow nonX = \Delta NX\rightarrow nonX/(\Delta T \cdot L \cdot PX)$ and $FnonX\rightarrow X = \Delta NnonX\rightarrow X/(\Delta T \cdot L \cdot (1 - PX))$. It is impossible to measure directly these $\Delta N$ values for a specific time interval of $\Delta T$. However, with a good approximation we can assume that the frequencies are proportional to the observed numbers $NX\rightarrow nonX$ and $NnonX\rightarrow X$ and can be represented by the simple formula: $FX\rightarrow nonX = A \cdot NX\rightarrow nonX/PX$ and $FnonX\rightarrow X = A \cdot NnonX\rightarrow X/(1 - PX)$, where $A$ is a scaling factor having the same value for
$FX\to\text{non}X$ and $F\text{non}X\to X$, since $NX\to\text{non}X$ and $N\text{non}X\to X$ are counted from the same sample. In a gedanken experiment, let's assume that the current $FX\to\text{non}X$ and $F\text{non}X\to X$ values will stay unchangeable forever for our MRI sample. Then, in time, mutations should alter the base composition of our sample until it reaches an equilibrium composition with a new percentage for $X$-bases denoted here as $Q_X$. This equilibrium composition $Q_X$ can be computed using the observed parameters of $PX$, $NX\to\text{non}X$, and $N\text{non}X\to X$. Indeed, under the equilibrium, the number of changes from $X$ to non$X$ must be equal to the number of reverse changes from non$X$ to $X$, or:

$$\Delta N_{X\to \text{non}X} = \Delta N_{\text{non}X\to X}$$

(1)

We can compute these $\Delta N_{X\to \text{non}X}$ and $\Delta N_{\text{non}X\to X}$ values from frequencies in such a way:

$$\Delta N_{X\to \text{non}X} = F_{X\to \text{non}X} \cdot \Delta T \cdot L \cdot Q_X = A \cdot \frac{N_{X\to \text{non}X}}{P_X} \cdot \Delta T \cdot L \cdot Q_X$$

also in a similar way

$$\Delta N_{\text{non}X\to X} = F_{\text{non}X\to X} \cdot \Delta T \cdot L \cdot (1 - Q_X) = A \cdot \frac{N_{\text{non}X\to X}}{1-P_X} \cdot \Delta T \cdot L \cdot (1 - Q_X)$$

By putting these transformations into Equation 1 we get:

$$A \cdot \Delta T \cdot L \cdot Q_X \cdot \frac{N_{X\to \text{non}X}}{P_X} - A \cdot \Delta T \cdot L \cdot (1 - Q_X) \cdot \frac{N_{\text{non}X\to X}}{(1-P_X)}$$

or

$$Q_X \cdot \frac{N_{X\to \text{non}X}}{P_X} = (1 - Q_X) \cdot \frac{N_{\text{non}X\to X}}{(1-P_X)}$$

(2)

Finally, simple transformation of Equation 2 gives us the final Equation 3 for calculation of equilibrium percentage:
In the Results section, Formula 3 is used to compute the equilibrium percentage for $X$-bases in the studied MRI regions.

**Abbreviations**

MRI: mid-range inhomogeneity; SNP: Single nucleotide polymorphisms; Mb: Megabase(s); indels: insertions and deletions; nt: nucleotide(s).

**Authors' contributions**

AP, BRH, SPA, MC, JH, OMB were responsible for computational processing of the human-chimp-macaque datasets and creating the described programs. SS was responsible for the procuring and processing of the SNP data from dbSNP. JMB was responsible for the quantification of SNP data in the alignment. AP was also responsible for the processing of fixed substitutions and indel data from the three way alignment. AF supervised the project, provided guidance and wrote the draft. SS and AP also contributed to editing, typesetting, and writing the draft. All authors have read and approved the final manuscript.

**Supplementary Material**

**Additional file 1**

S1 - SNP dataset. Complete set of data representing the number of occurrences of major, middle, minor and rare SNPs in all $X$ MRI regions (where $X$ could represent either a two-base combination [GC, AG, GT, etc.] or a single base) and their
corresponding control regions, which have an average nucleotide composition for the respective $X$.

**Additional file 2**

**S2A - Indel data for combination of nts, S2B - Indel data for individual nts.** • **2A** - Complete set of data representing the total number of $X$ nucleotides (where $X$ represents a two base combination such as GC, AG, GT, etc.) inserted or deleted due to short indels, whose sizes are less than three nucleotides in length, or medium indels, whose sizes range from three to fifty nucleotides in length, for all $X$ MRI regions, and for the control regions, which have an average nucleotide composition for the respective $X$. • **2B** - Complete set of data representing the total number of $X$ nucleotides (where $X$ represents a single base) inserted or deleted due to short indels, whose sizes are less than three nucleotides in length, or medium indels, whose sizes range from three to fifty nucleotides in length, for all $X$ MRI regions, and for the control regions, which have an average nucleotide composition for the respective $X$.

**Additional file 3**

**S3 - Dataset of fixed substitutions.** Complete set of data representing the number of occurrences of fixed substitutions on all $X$ MRI regions (where $X$ could represent either a two-base combination [GC, AG, GT, etc.] or a single base), and the control regions, which have an average nucleotide composition for the respective $X$.

**Acknowledgements**

This project is supported by NSF Career award MCB-0643542. We thank
Peter Bazeley, University of Toledo, for his computational support and discussion of our algorithms.
References


Chapter Four

Construction of a database of mammalian orthologous introns of 5 species (DOMINO5).

Authors

Ashwin Prakash, and Alexei Fedorov
Abstract

No well documented case of intron gain exists between fully sequenced mammalian genomes, and there are only a few intron losses reported [1, 2]. Therefore, introns are ‘fossilized’ in mammalian genes, and the exon–intron gene structure of mammals is highly conserved. Thus we can define ‘orthologous introns’ as introns from orthologous genes that have the same position relative to the coding sequence. Orthologous introns are valuable tools in a comparative study of mammalian introns such as ours, for this reason, we created a database of orthologous introns comprising human, cow, dog, mouse and rat species. An existing Mammalian orthologous database (MOID), consists of human, rat and mouse introns, but we created a new database using the current build of each of the five mammalian genomes (GRCH37, bosTau5, canFam2, mm9 and rn4). The inclusion of cow and dog into the database, served to increase the sensitivity in identification of orthologous introns between the common ancestor of all mammals, in comparison with the MOID database.

Methods

Characterizing Homologous Proteins

The first step in the construction of the DOIMINO5 was to identify all homologous proteins between human and each of the other species. This was performed using BLAST-2 analysis [3] and the protein databases for each organism was obtained from the FASTA formatted protein database in latest version of the exon-intron database (EID). Identifying homologous proteins between each non-human specie, to a single
protein in the human, involved several cycles of BLAST, while using a human protein as the query, against each of the non-human protein database. The series of BLAST searches would yield five files, each containing homologous proteins between human and one of the non-human species.

In order to rule out false positive hits thrown out by the BLAST searches, we only retained BLAST hits with a bit-score of 80 or over. There were several instances where a protein had more than one strong hit, which was likely to be a hit within several isoforms of the same gene within the other species, this would be a source of redundancies. In order to filter out such instances reverse BLAST of the proteins from the other species was performed against the human protein database, using BLAST-2, an illustration of this is provided in Figure 4-1.

Then the two BLAST search results, human with each of the other species and vice versa, were cross referenced generating a final set, retaining only those instances where each protein or its isoform in human has no more than one corresponding protein or isoform in the other species. The BLAST results were finally organized into four files, each of which would contain the alignments of homologous proteins between human and one of the other species. These four files were further processed to obtain orthologous introns between each pair of species.
Fig 4-1: Identification of Isoform association between two species.

Sequential BLAST analysis between protein query from the human with the non-human protein database, followed by protein query from non-human against the human protein database, can be used to elucidate the true association between isoforms of protein in either species.

Orthologous introns from homologous proteins

Orthologous introns are introns from homologous genes between species, where the introns share the same position relative to the reading frame. It is to be noted here that the definition excludes intron length or sequence information as determinants of orthology, but defines their position and phase as the sole criteria (Fig 4-2.1). In order to define all such instances within homologous proteins between the species, we will use a perl program which will map all intron positions and phase within the BLAST-2 protein alignments (Fig 4-2.2). Using this mapping of introns within the protein alignments, another perl program will select for only those instances where introns share common position and phase in both the species. The five resulting sets of homologous proteins
between human and each of the other four species, are cross-referenced to retain only introns which are common between all the species to create a five way table with unique instances of orthologous introns between the species with no redundancies. This process will finally lead to a single file in which each line contains introns identifiers, for corresponding orthologous introns from each of the five species in tab delimited format, in the form of a five way table.

![Diagram of orthologous introns between species](image)

**Fig 4-2.1: Identification of Orthologous introns between species.**

Orthologous introns are determined by their position within homologous proteins between all the five species, and the definition [4] is irrespective of their size or sequence conservation. These introns lie between corresponding highly conserved exons across species, and their location within the mRNA is conserved.
**Fig 4-2.2: Processing Blast output for homologous proteins.**

The BLAST output which will contain homologous protein alignments will be processed using a perl program to map all intron positions within the proteins of both the species as shown as numbers above and below the protein sequence here, the numbers in red signify the intron phase. Introns are defined as orthologous when they share position as well as phase between species, and all instances which do not satisfy this criteria are excluded from DOMINOS.

The five way table of non-redundant orthologous introns will be used to extract the nucleotide sequence information in FASTA format for each intron from the exon-intron database, and placed in groups such that the sequences are placed in groups of introns, each group consisting of orthologous introns from the five species. The groups will be arranged in order of their location in the human genome. The file with all the orthologous introns arranged in groups will be a flat file database which will be available for download to enable users to perform large-scale analyses of the features in the database.

All the sequences of orthologous introns within the groups will be in FASTA format, simplifying their use in many applications. We will also provide intron phase information, intron position, length, and scientific name of the organism, within the command line of each intron. The introns within each group will be arranged such that
the human intron will be followed by introns of other species in alphabetical order of their genus (i.e. Human (Homo sapiens) followed by, Cow (Bos taurus), Dog (Canis Familiaris), Mouse (Mus musculus) and Rat(Rattus Norvegicus)). All of these features are accessible through easily parseable handles. A query-able database will be developed to provide easy accessibility to intron groups and better navigation through the database.

**Alignment of orthologous introns**

These groups will be further processed to create multiple alignments, using the MAFFT local stand alone program based on a Fast Fourier Transform (FFT) algorithm [5]. The program takes about 10 seconds to align each group of introns with an average nucleotide length of 2000 bp [5]. Of over 200,000 human introns in the exon-intron database [6], if even half are found to have orthologs common among all the five species, the alignment process would require several weeks of computational processing. In order to expedite the process we will develop a program which will divide the groups of introns into 15 equal groups, and in turn submit each group to be processed on a separate node within a cluster of 15 CPUs, and retrieve the results and assimilate it to reconstruct one file consisting of multiple alignments of all the groups of orthologous introns. This parallel processing is likely to bring down the time required to about a 15th of the time required if processed individually. Some pitfalls in the multiple alignment process could be very wide variations in length of each intron within a group, poor sequencing of segments within introns resulting in multiple ambiguous nucleotides (N's) in a sequence and poor sequence homology between the introns, we will exclude all such cases. The alignments of all the orthologous intron groups will be available for download to public users as a
flat file in FASTA format.
References


3. Fedorov A., Fedorova L. “An Intricate Mosaic of Genomic Patterns at Mid-range Scale” Chapter3 in “Genome-scale Pattern Analysis in the Post-ENCODE Era.”


Chapter Five

Association of genomic Mid Range Inhomogeneity with DNA conformations

Authors
Ashwin Prakash and Alexei Fedorov
Introduction

DNA can adopt a number of different structures apart from the B-form described by Watson and Crick [1]. Regions of nucleotide composition inhomogeneity could give rise to noncanonical DNA secondary structures having a range of biological functions. This study deals with the evolution and distribution of putative Z-DNA and putative H-DNA that are among the less understood DNA structures, and also looks at the distribution of other regions of nucleotide inhomogeneity.

With the advent of single-crystal X-ray diffraction on specific molecules in the 1970s it became possible to define the structure of DNA. The first single-crystal X-ray structure of a DNA fragment — a self-complementary DNA hexamer d(CG)3 — surprisingly showed a left-handed double helix with two anti-parallel chains that were held together by Watson-Crick base pairs [1] instead of the familiar right-handed B-DNA double helix that had been the focus of molecular biology. The alternative structure with a zig-zag arrangement ((RY)n repeats) of its backbone pointed to an unusual function for this form of DNA and was named Z-DNA. Z-DNA was confirmed by examining the Raman spectra of DNA in 4M sodium chloride solution [3]. The conversion of B-DNA to Z-DNA was associated with a ‘flipping over’ of the base pairs so that they had an upside down orientation relative to that of B-DNA [4]. In the presence of a high-salt solution, the electrostatic repulsion of the phosphate residues is vastly decreased, and Z-DNA becomes the stable conformation, while under normal cellular conditions, the electrostatic repulsion tends to push the molecule into the B-DNA conformation. The net result of this reorganization was that the phosphate groups were closer together in Z-DNA than in B-DNA [4]. The sequences that most readily underwent this conversion had alternations of purines and
pyrimidines, especially of C and G. This change also occurred easily with alternations of CA on one strand and TG on the other [5, 6].

Another type of noncanonical DNA structure, an intramolecular triplex or H-DNA, forms at regions containing mirror repeat symmetry, where one half of the DNA strand can dissociate into single strands by using the energy provided by supercoiling [7]. One of the single strands can then swivel its backbone parallel to the purine-rich strand in the remaining duplex, forming a three-stranded helix in this half of the region, leaving its complementary strand unpaired [8, 9]. Unlike other DNA secondary structure-forming sequences, which are typically located in intergenic or intronic regions, H-DNA forming sequences have so far been found most frequently in promoters and exons and have been found to be involved in regulating the expression of several disease-linked genes [10, 11]. Most studies have focused on H-DNA’s regulatory role in transcription, and potential contribution to genetic instability has also been examined.

The alternation of purines and pyrimidines that give rise to stable Z-DNA conformation, as well as the purine-rich stretch of nucleotides which characterize H-DNA give rise to nucleotide compositional patterns which would show up as corresponding MRI features. Z-DNA can be characterized as a special case of G+C/G+T/A+C/A+T – rich MRI features, while H-DNA can be characterized as G+A/C+T – rich MRI features and can be mapped out using the MRI analyzer programs.

Considering mid-range genomic compositional patterns from the most
commonly studied case: G + C-rich and A + T-rich regions. These regions of various
lengths from thirty to several thousand nucleotides are 4–20 times over-represented in
the mammalian genomes compared to a random expectation [12, 13]. It has been
shown recently that CpG dinucleotide without methylation exhibit structural
abnormalities in the DNA helix. Particularly, they are one of the most frequent sites
for DNA backbone cleavage by hydroxyl radicals [14, 15] and during the sonication
of double-stranded DNA [16]. G+C-rich regions may be present in A-form DNA even
in aqueous solutions [17,18,19]. A special form of DNA which is an intermediate
between A- and B-forms, has been characterized in G + C-rich sequences with
methylated cytosines [20]. The crucial involvement of cytosine methylation in the
regulation of gene expression is well described in a number of reviews including
some recent ones [21, 22, 23]. Thus, here we also study the distribution of G + C-rich
and A + T-rich regions.

Another kind of MRI we study in this paper are G+T-rich and A+C rich
regions. It should be noted at this point that complementary strands of G + T-rich
regions are naturally A + C-rich regions so these features co-exist with each other and
we consider them interchangeably with respect to their description in the literature.
Bechtel and co-authors demonstrated that G + T-regions are about five times more
abundant in the mammalian genomes compared to random expectation [13]. Despite
their abundance, G+T-rich motifs are much less investigated than other regions with
extremes in base compositions, and in this study we attempt to map these features to
gain an understanding of their distribution so as to shed light on possible functional
roles of these regions.
Results

Mid range inhomogeneous structures in introns of DOMINO5

Except few ultra conserved sections within them, it is very hard to identify long stretches of orthologous DNA within intergenic regions, and thus in this study we focus on orthologous introns. The Database of Orthologous Mammalian Introns of five species (DOMINO5) contains 61,633 orthologous introns each of human, cow, dog, mouse and rat. These introns have been directly aligned to each other, and thus serve as a valuable resource for a comparative study of compositional features in the genomes of the five species. From the multiple aligned introns in DOMINO5 we extracted all the regions with inhomogeneous nucleotide compositions for each of the five species, satisfying the criteria for MRI (X-Rich MRI regions, where X was G+C, G+T or A+T content), using a modified version of the MRI analyzer program. We also developed programs to predict the sequences containing putative Z-DNA and putative H-DNA features within introns of all the five species (see methods). In order to exclude the compositional features found in extra-long [25] introns and decrease the number of false positives, we only considered introns, which were less than 6000 nucleotides in length, this included 46,696 (more than 75% introns) of the introns in DOMINO5. We used a minimum window size of 100 nucleotides in order to characterize MRI, putative Z-DNA and putative H-DNA features. The threshold percentage above which a region is recognized as a feature and is defined as X-Rich was established individually for each feature.
Putative Z-DNA structures within DOMINO5 introns

The Z-DNA like features were characterized by sequences over 100 nucleotides in length enriched with at least 70% purine-pyrimidine alternating (RpY/YpR) dinucleotide repeats. The putative Z-DNA features were computed within DOMINO5 introns, using the Z-DNA Analyzer.java program. Figure 5-1 depicts a putative Z-DNA structure within sequences of introns within DOMINO5.

**Figure 5-1: Putative Z-DNA sequences in multiple alignments.** The figure depicts sequence information of putative Z-DNA structures within the multiple alignment of human, cow, dog, mouse and rat introns within the DOMINO5 database from intron 19 of the DIP2C gene (human protein id: NP_055789.1) on chromosome 10. The regions highlighted in color within each sequence depict the Z-DNA like structure forming sequence, and each species has a specific color to depict putative Z-DNA feature in it.
Of the 61,633 introns within DOMINO5, 6744 human introns were found to contain at least one putative Z-DNA structure within them. While 3787 putative Z-DNA structures were found within 3031 human introns that were shorter than 6000 nucleotides long. 152 of these introns shorter than 6000 nucleotids have at least one Z-DNA like structure within orthologous introns in all the five species.

**Figure 5-2: Z-DNA structures in introns less than 60000 nucleotides long within quintiles.** The bar-graph represents the relative abundance of Z-DNA like structures within each quintile of a gene, each quintile is depicted on the X axis (1<sup>st</sup> quintile is the first 20% of gene length, 2<sup>nd</sup> quintile 21% to 40% and so on) and the Y axis depicts the proportion (depicted as percentage) of the 264 Z-DNA like structures within each quintile in 152 orthologous intron groups which contain at least one Z-DNA like structure in all five species in the group.
The probability of finding at least one Z-DNA like structure in all five species within a group of orthologous introns (excluding extra long introns) shorter than 6000 nucleotides in a random expectation model is 1 in 218,182 and so we can clearly observe that there is a strong enrichment of Z-DNA like structures within orthologous introns. The 264 putative Z-DNA structures within human introns shorter than 6000 nucleotides where each of the five species contain at least one Z-DNA like structure were then mapped onto their corresponding genes in order to study their position within these genes. The results suggest (Figure 5-2) that putative Z-DNA structures tend to be enriched in the 3’ end of the gene, while they tend to avoid the first quintile. A study of the 3787 putative Z-DNA structures in all human introns from DOMINO5 shorter than 6000 nucleotides in length, shows a similar distribution of Z-DNA like structures (Figure 5-3) where these structures tend to be enriched in the last two quintiles (two-fifth of the 3’ end) of the gene, while they avoid the 5’ end.

![Diagram](image)

**Figure 5-3:** Z-DNA structures in all human introns within quintiles. The bar-graph
represents the relative abundance of Z-DNA like structures within each quintile of a gene, each quintile is depicted on the X axis (1\textsuperscript{st} quintile is the first 20\% of gene length, 2\textsuperscript{nd} quintile 21\% to 40\% and so on) and the Y axis depicts the proportion (depicted as percentage) of the Z-DNA like structures within each quintile in orthologous intron groups which contain at least one Z-DNA like structure in human.

We then studied the degree of sequence conservation of the 264 Z-DNA like structures, using the five way alignment of the orthologous introns. The results suggest that 87.5\% of the putative Z-DNA structures had less than 30\% sequence similarity (Figure 5-4), which is very similar to the average degree of conservation expected to be seen between the five orthologous introns. This suggests that there does not seem to a significant selection pressure acting to preserve the sequences of the Z-DNA like structures.
**Figure 5-4: Z-DNA structures by sequence conservation.** This bar-graph represents the distribution of the 264 Z-DNA like structures based on their degree of sequence conservation among all the five species in DOMINO5 introns given as percentage identity. The X-Axis represents the percentage of conservation within each feature, while the Y-Axis represents the percentage of features which have that level of conservation.

Though the data suggests that sequence conservation of putative Z-DNA structures is not higher than average levels of conservation for introns, we investigated if the position of these structures was conserved within the introns in the five species. The results (Figure 5-5 and 5-6) suggest a surprisingly high degree of positional conservation between Z-DNA like structures, with 46.59% of the structures having the same position in all the five species and over 81.44% structures being preserved in human and at least two of the other four species. This suggests that there is a significant selection pressure favoring maintenance of the position of Z-DNA like structures within the introns.
Figure 5-5 and 5-6: Z-DNA structures by preservation of position. The bar-graph represents the percentage distribution of putative Z-DNA features which overlap with the features in orthologous introns of other species. The X-Axis indicates the number of species which have a feature in that particular position, and the Y-Axis represents the percentage of features. The pie chart represents the degree of preservation of position, where the presence of overlap of features between human and at least two other species is considered to be significantly highly preserved.

Putative H-DNA structures within DOMINO5 introns

Putative H-DNA features were characterized by sequences over 100 nucleotides in length, enriched with at least 80% purine-purine (ApG/GpA) or pyrimidine-pyrimidine (TpC/CpT) dinucleotide repeats. The H-DNA like features were computed within DOMINO5 introns, using the H-DNA_analyzer.java program. Figure 5-7 depicts putative H-DNA structure within intron sequences of introns within DOMINO5.
Figure 5-7: Putative HDNA sequences in multiple alignments. The figure depicts sequence information of putative H-DNA structures within the multiple alignment of human, cow, dog, mouse and rat orthologous introns within the DOMINO5 database from intron 12 of the RUFY2 gene (human protein id: NP_060457.4) on chromosome 10. The regions highlighted in color within each sequence depict the H-DNA structure forming sequence, and each species has a specific color to depict putative H-DNA feature in it.

Of the 61,633 introns within DOMINO5, 5576 human introns were found to contain at least one H-DNA like structure within them. While 3213 H-DNA structures were found within 2733 human introns which were less than 6000 nucleotides long. 260 of these introns have at least one H-DNA like structure within corresponding orthologous introns in all the five species.
Figure 5-8: **HDNA structures in introns less than 60000 nucleotides long within quintiles.** The bar-graph represents the relative abundance of HDNA like structures within each quintile of a gene, each quintile is depicted on the X axis (1st quintile is the first 20% of gene length, 2nd quintile 21% to 40% and so on) and the Y axis depicts the proportion (depicted as percentage) of the 386 HDNA like structures within each quintile in 260 orthologous intron groups which contain at least one HDNA like structure in all five species in the group.

The probability of finding at least one H-DNA like structure in all five species within a group of orthologous introns (excluding extra long introns) in a random expectation model is 1 in 369,530 and so we can clearly observe that there is a strong enrichment of H-DNA like structures within orthologous introns. The 386 H-DNA structures within human introns shorter than 6000 nucleotides, where all five species contain H-DNA like structures were then mapped onto their corresponding genes in order to study their position within these genes. The results suggest (Figure 5-8) that
H-DNA like structures does not have any preference to be enriched at a particular part within a gene; however there seems to be a slight avoidance of the second quintile of the gene. A study of 3210 H-DNA like structures in all human introns shorter than 6000 nucleotides however shows a slightly different distribution (Figure 5-9) where H-DNA like structures tend to be enriched in the final two-fifth of the 3’ end of the gene, while they avoid the 5’ end.

![Abundance of HDNA like structures within the gene](image)

**Figure 5-9: HDNA structures in all human introns within quintiles.** The bar-graph represents the relative abundance of HDNA like structures within each quintile of a gene, each quintile is depicted on the X axis (1st quintile is the first 20% of gene length, 2nd quintile 21% to 40% and so on) and the Y axis depicts the proportion (depicted as percentage) of the 3213 HDNA like structures within each quintile in orthologous intron groups which contain at least one HDNA like structure in humans.

We then studied the degree of sequence conservation of the 386 H-DNA
structures, using the five-way alignment of the orthologous introns. The result suggests that 71.5% of the H-DNA-like structures had less than 30% sequence conservation (Figure 5-10), which is very similar to the average degree of conservation expected to be seen between the five orthologous introns. This suggests that there does not seem to be a significant selection pressure acting to preserve a majority of the H-DNA-like structures. However, there were a small number (4.9%) of the structures which have greater than 90% conservation.

**Figure 5-10: HDNA structures by sequence conservation.** This bar-graph represents the distribution of the 386 H-DNA-like structures based on their degree of sequence conservation among all the five species in DOMINO5 introns given as percentage identity. The X-Axis represents the percentage of conservation within each feature, while the Y-Axis represents the percentage of features which have that level of conservation.

Though the data suggests that sequence conservation of H-DNA-like structures...
is not higher than average levels of conservation for introns, we investigated the relative conservation of position of the structures in the introns among the five species. The results (Figure 5-11 & 5.12) suggest a surprisingly high degree of positional conservation between H-DNA like structures among the orthologous introns, with 58.5% of the structures having the same position in all the five species and over 88.34% structures being preserved in human and at least two of the other four species. This suggests that there is a significant selection pressure in order to maintain the position of H-DNA like structures within the introns.

**Figure 5-11 and 5.12: HDNA structures by preservation of position.** The bar-graph represents the percentage distribution of putative H-DNA features which overlap with the features in orthologous introns of other species. The X-Axis indicates the number of species which have a feature in that particular position, and the Y-Axis represents the percentage of features. The pie chart represents the degree of preservation of position, where the presence of overlap of features between human and at least two other species is considered to be significantly highly preserved.
GC-Rich MRI features within DOMINO5 introns

GC-rich MRI features were characterized by sequences over 100 nucleotides in length, enriched with at least 75% G+C nucleotide composition. The GC-MRI features were computed within DOMINO5 introns, using the GCMRI_analyzer.java program. Figure 5-13 depicts a MRI feature within intron sequences of introns within DOMINO5.

Figure 5-13: GC-MRI features in multiple alignments. The figure depicts sequence information of GC-MRI features within the multiple alignment of human, cow, dog,
Of the 61,633 introns within DOMINO5, 2492 human introns were found to contain at least one GC-MRI feature within them. While 2335 GC-MRI features were found within 1754 human introns which were less than 6000 nucleotides long, of which 188 introns have at least one feature within corresponding orthologous introns in all the five species. The probability of finding at least one GC-MRI feature in all five species within a group of orthologous introns (excluding extra long introns) in a random expectation model is 1 in 58,198,621 and so we can clearly observe that there is a strong conservation of GC-MRI feature within orthologous introns.

![Abundance of GC-MRI features within the gene](image)

**Figure 5-14:** GC-MRI features in introns less than 60000 nucleotides long within
quintiles. The bar-graph represents the relative abundance of GC-MRI features within each quintile of a gene, each quintile is depicted on the X axis (1st quintile is the first 20% of gene length, 2nd quintile 21% to 40% and so on) and the Y axis depicts the proportion of GC-MRI features within each quintile in orthologous intron groups which contain at least one GC-MRI feature in all five species in the group, depicted as a percentage.

The 308 GC-MRI features within short human introns where all five species contain GC-MRI features were then mapped onto their corresponding genes in order to study their position within these genes. The result suggests (Figure 5-14) that GC-MRI features have a significant enrichment in the first quintile in the gene and the occurrence drastically declines in other parts of the gene. A study of 2335 GC-MRI features in human introns shorter than 6000 nucleotides also shows a similar distribution of GC-MRI features (Figure 5-15) where GC-MRI features tend to be enriched in the first quintile in the gene, while they avoid the 3’ end.
Figure 5-15: **GC-MRI features in all human introns within quintiles.** The bar-graph represents the relative abundance of GC-MRI features within each quintile of a gene, each quintile is depicted on the X axis (1st quintile is the first 20% of gene length, 2nd quintile 21% to 40% and so on) and the Y axis depicts the proportion of GC-MRI features within each quintile in orthologous intron groups which contain at least one GC-MRI feature in human, depicted as a percentage.

We then studied the degree of sequence conservation of the 308 GC-MRI features, using the five way alignment of the orthologous introns. The result suggest that 73.7% of the GC-MRI features had greater than 30% sequence conservation (Figure 5-16) and 35.39% having greater than 90% conservation, which is significantly higher than the average degree of sequence conservation seen between the five orthologous introns. This suggests that there is a significant selection pressure acting to preserve a majority of the GC-MRI features.
Figure 5-16: GC-MRI features by sequence conservation. This bar-graph represents the distribution of the 308 GC-MRI features based on their degree of sequence conservation among all the five species in DOMINO5 introns given as percentage identity. The X-Axis represents the percentage of conservation within each feature, while the Y-Axis represents the percentage of features which have that level of conservation.

As we found the data suggests that sequence conservation of GC-MRI features is significantly higher than average levels of conservation for introns, we also investigated to see if there was conservation of position of the features within the introns among the five species. The results (Figure 5-17 & 5.18) suggest a similarly high degree of positional conservation between GC-MRI features, with 32.14% of the structures having the same position in all the five species and over 82.79% structures being preserved in at least three of the five species including human. This suggests that there is a significant selection pressure in order to maintain the GC-MRI features
at particular positions within the introns.

![Bar graph and pie chart](image)

**Figure 5-17 and 5.18: GC-MRI features by preservation of position.** The bar-graph represents the percentage distribution of GC-MRI features which overlap with the features in orthologous introns of other species. The X-Axis indicates the number of species which have a feature in that particular position, and the Y-Axis represents the percentage of features. The pie chart represents the degree of preservation of position, where the presence of overlap of features between human and at least two other species is considered to be significantly highly preserved.

**GT-Rich MRI features within DOMINO5 introns**

GT-rich MRI features were characterized by sequences over 100 nucleotides in length, enriched with at least 75% G+T nucleotide composition. The GT-MRI features were computed for both strands of DNA within DOMINO5 introns, using the GTMRI_analyzer.java program. Figure 5-19 depicts a MRI feature within intron sequences of introns within DOMINO5.
Figure 5-19: **GT-MRI features in multiple alignments.** The figure depicts sequence information of GT-MRI features within the multiple alignment of human, cow, dog, mouse and rat orthologous introns within the DOMINO5 database from intron 10 of the FAF1 gene (human protein ID: NP_008982.1) on chromosome 1. The regions highlighted in color within each sequence depict the GT-MRI sequence, and each species has a specific color to depict GT-MRI feature in it.

Of the 61,633 introns within DOMINO5, 6724 human introns were found to contain at least one GT-MRI feature within them. While 3606 GT-MRI features were found within 3130 human introns which were less than 6000 nucleotides long, of which 49 introns have at least one within corresponding orthologous introns in all the five species. The probability of finding at least one GT-MRI feature in all five species within a group of orthologous introns (excluding extra long introns) in a random expectation model is 1 in 266,528 and so we can clearly observe that there is a strong conservation of GT-MRI feature within orthologous introns.
Figure 5-20: GT-MRI features in introns less than 60000 nucleotides long within quintiles. The bar-graph represents the relative abundance of GT-MRI features within each quintile of a gene, each quintile is depicted on the X axis (1st quintile is the first 20% of gene length, 2nd quintile 21% to 40% and so on) and the Y axis depicts the proportion of GT-MRI features within each quintile in orthologous intron groups which contain at least one GT-MRI feature in all five species in the group, depicted as a percentage.

The 128 GT-MRI features within short human introns where all five species contain GT-MRI features were then mapped onto their corresponding genes in order to study their position within these genes. The result suggests (Figure 5-20), that GT-MRI features do not have any specific position preference within a gene however there seems to be a slight avoidance of the first quintile at the gene. A study of 3606 GT-MRI features in all human introns larger than 6000 nucleotides however shows a
slightly different distribution (Figure 5-21) where the features tend to be enriched in the final two-fifth of the 3’ end of the gene, while they avoid the 5’ end.

![Abundance of GT-MRI features within the gene](image)

**Figure 5-21: GT-MRI features in all human introns within quintiles.** The bar-graph represents the relative abundance of GT-MRI features within each quintile of a gene, each quintile is depicted on the X axis (1st quintile is the first 20% of gene length, 2nd quintile 21% to 40% and so on) and the Y axis depicts the proportion of GT-MRI features within each quintile in orthologous intron groups which contain at least one GT-MRI feature in human, depicted as a percentage.

We then studied the degree of sequence conservation of the 128 GT-MRI features, using the five way alignment of the orthologous introns. The result suggest that 73.44% of the GT-MRI features had less than 30% sequence conservation (Figure 5-22), which is very similar to the average degree of sequence conservation seen between the five orthologous introns. This suggests that there does not seem to be a
significant selection pressure acting to preserve the sequence of a majority of the GT-MRI features.

Figure 5-22: **GT-MRI features by sequence conservation.** This bar-graph represents the distribution of the 128 GT-MRI features based on their degree of sequence conservation among all the five species in DOMINO5 introns given as percentage identity. The X-Axis represents the percentage of conservation within each feature, while the Y-Axis represents the percentage of features which have that level of conservation.

As we found the data suggests that sequence conservation of GT-MRI features is similar to average levels of conservation for introns, we also investigated to see if there was conservation of position of the features within the introns among the five species. The results (Figure 5-23 & 5.24) suggest a relatively higher degree of positional conservation between GT-MRI features, with 19.53% of the structures...
having the same position in all the five species and over 56.25% structures being preserved in at least three of the five species including human. This suggests that there is only a weak selection pressure to maintain the GT-MRI features at particular positions within the introns.

Figure 5-23 and 5.24: GT-MRI features by preservation of position. The bar-graph represents the percentage distribution of GT-MRI features which overlap with the features in orthologous introns of other species. The X-Axis indicates the number of species which have a feature in that particular position, and the Y-Axis represents the percentage of features. The pie chart represents the degree of preservation of position, where the presence of overlap of features between human and at least two other species is considered to be significantly highly preserved.

AT-Rich MRI features within DOMINO5 introns

AT-rich MRI features were characterized by sequences over 100 nucleotides in length, enriched with at least 83% A+T nucleotide composition. The AT-MRI features were computed for both strands of DNA within DOMINO5 introns, using the ATMRI_analyzer.java program. Figure 5-25 depicts a MRI feature within intron sequences of
Figure 5-25: AT-MRI features in multiple alignments. The figure depicts sequence information of AT-MRI features within the multiple alignment of human, cow, dog, mouse and rat orthologous introns within the DOMINO5 database from intron 1 of the SEC23IP gene (human protein id: NP_009121.1) on chromosome 10. The regions highlighted in color within each sequence depict the AT-MRI sequence, and each species has a specific color to depict AT-MRI feature in it.

Of the 61,633 introns within DOMINO5, 8206 human introns were found to contain at least one AT-MRI feature within them. While 5016 AT-MRI features were
found within 4232 human introns which were less than 6000 nucleotides long, of which 116 short introns have at least one within corresponding orthologous introns in all the five species. The probability of finding at least one AT-MRI feature in all five species within a group of orthologous introns (excluding extra long introns) in a random expectation model is 1 in 266,528 and so we can clearly observe that there is a strong conservation of AT-MRI feature within orthologous introns.

Figure 5-26: **AT-MRI features in introns less than 60000 nucleotides long within quintiles.** The bar-graph represents the relative abundance of AT-MRI features within each quintile of a gene, each quintile is depicted on the X axis (1st quintile is the first 20% of gene length, 2nd quintile 21% to 40% and so on) and the Y axis depicts the proportion of AT-MRI features within each quintile in orthologous intron groups which contain at least one AT-MRI feature in all five species in the group, depicted as a percentage.
The 165 AT-MRI features within short human introns where all five species contain AT-MRI features were then mapped onto their corresponding genes in order to study their position within these genes. The result suggests (Figure 5-26), that AT-MRI features tend to be enriched in the third and fourth quintile there also seems to be a slight avoidance of the first two quintiles of the gene. A study of 5016 AT-MRI features in all short human introns shows a similar distribution (Figure 5-21) where the features tend to be enriched in the fourth quadrant towards the 3’ end of the gene, while they avoid the 5’ end.

Figure 5-27: AT-MRI features in all human introns within quintiles. The bar-graph represents the relative abundance of AT-MRI features within each quintile of a gene, each quintile is depicted on the X axis (1st quintile is the first 20% of gene length, 2nd quintile 21% to 40% and so on) and the Y axis depicts the proportion of AT-MRI features within each quintile in orthologous intron groups which contain at least one AT-MRI feature in all five species in the group, depicted as a percentage.
We then studied the degree of sequence conservation of the 165 AT-MRI features, using the five way alignment of the orthologous introns. The result suggests that 66.67% of the AT-MRI features had less than 30% sequence conservation (Figure 5-22), which is very similar to the average degree of sequence conservation seen between the five orthologous introns. This suggests that there does not seem to be a significant selection pressure acting to preserve the sequence of a majority of the AT-MRI features.

**Figure 5-28: AT-MRI features by sequence conservation.** This bar-graph represents the distribution of the 165 AT-MRI features based on their degree of sequence conservation among all the five species in DOMINO5 introns given as percentage identity. The X-Axis represents the percentage of conservation within each feature, while the Y-Axis represents the percentage of features which have that level of conservation.
As we found the data suggests that sequence conservation of AT-MRI features is significantly higher than average levels of conservation for introns, we also investigated to see if there was conservation of position of the features within the introns among the five species. The results (Figure 5-23 & 5.24) suggest a low degree of positional conservation between AT-MRI features, with 10.91% of the structures having the same position in all the five species and only 43.64% structures being preserved in at least three of the five species including human. This suggests that there is no significant selection pressure to maintain the AT-MRI features at particular positions within the introns.

**Figure 5-29 and 5-30: AT-MRI features by preservation of position.** The bar-graph represents the percentage distribution of AT-MRI features which overlap with the features in orthologous introns of other species. The X-Axis indicates the number of species which have a feature in that particular position, and the Y-Axis represents the percentage of features. The pie chart represents the degree of preservation of position, where the presence of overlap of features between human and at least two other
species is considered to be significantly highly preserved.

Methods

Database of Orthologous Mammalian Introns of 5 species (DOMINO5)

Introns from the DOMINO5 database, discussed in chapter 4, were used for the study of conservation of Z-DNA, H-DNA and other MRI features. The intron groups were divided into extra long and others based on the length of the human intron in each group, where if the human intron was greater that 6000 nucleotides long it was referred to as extra long intron. Over 75% (46,696) of the DOMINO5 introns were found to be shorter than 6000 nucleotides in length. The DOMINO5 alignments were used to calculate sequence identity and positional conservation between features across the five species.

Characterizing Z-DNA features in DOMINO5

Z-DNA features were characterized using the program Z-DNA_analyzer.java within the DOMINO5 introns. Z-DNA features were defined as stretch of sequences within each intron that was over 100 nucleotides in length, and are enriched with at least 70% purine-pyrimidine alternating (RpY/YpR) dinucleotide repeats. The program uses a stretchy window approach such that once a 100 nucleotide long Z-DNA feature with a dinucleotide repeat frequency greater than the threshold is identified; the program tries to incorporate the maximum possible nucleotides downstream from the features until the dinucleotide frequency drops below the threshold. The start and end positions of these features were calculated for each species and used for further statistical evaluation. As the DOMINO5 alignments were used in the study, upon
identification of each Z-DNA feature in human, the program also calculated the percentage conservation of nucleotides within the feature.

Characterizing H-DNA features in DOMINO5

H-DNA features were characterized using the program H-DNA_analyzer.java within the DOMINO5 introns. H-DNA features were defined as sequences within introns at least 100 nucleotides in length, enriched with at least 80% purine-purine (ApG/GpA) or pyrimidine-pyrimidine (TpC/CpT) dinucleotide repeats. The program uses a stretchy window approach such that once a 100 nucleotide long H-DNA feature with a dinucleotide repeat frequency greater than the threshold is identified; the program tries to incorporate the maximum possible nucleotides downstream from the features until which the dinucleotide frequency drops below the threshold. The start and end positions of these features were calculated for each species and used for further statistical evaluation. As the DOMINO5 alignments were used in the study, upon identification of each H-DNA feature in human, the program also calculated the percentage conservation of nucleotides within the feature.

Characterizing MRI features in DOMINO5

MRI features were characterized using java versions of the MRI analyzer program mentioned in chapter 1. The MRI features were computed on each intron within the DOMINO5 database. MRI features were defined as sequences within introns at least 100 nucleotides in length, enriched with specific nucleotide content above a defined threshold set for each content type. The program uses a stretchy window approach such that once a 100 nucleotide long MRI feature with a content richness greater than or equal to the threshold is identified; the program tries to incorporate the maximum
possible nucleotides downstream from the features until which the richness drops below the threshold. The start and end positions of these features were calculated for each species and used for further statistical evaluation. As the DOMINO5 alignments were used in the study, upon identification of each MRI feature in human, the program also calculated the percentage conservation of nucleotides within the feature. The regions with GC-richness or AT-richness were considered separately in our study and their distribution was calculated separately, while GT-rich and AC-rich regions were considered together for our analysis, because an AC-rich region on one strand of DNA would correspond to a GT-rich region on the opposite strand, and would hence be under similar evolutionary selection pressures and hence could be combined for our analyses. In each content type the thresholds were selected such that the number of intron groups contained at least one MRI feature of a given content type, in introns of all five species within that group.

**Calculation of sequence and positional conservation**

The coordinates of the various features were used to map them onto the multiple alignment of introns in DOMINO5 and for each feature in the human intron, sequence similarity was calculated between the five species and used as a percentage to arrive at statistics for conservation of these features. Positional conservation for the features between the five species was defined as the overlap of the features in all five species in the multiple alignment. This would be possible only in cases where introns from all five species in the group have at least one feature within their intron.

**Calculating distribution of features within genes**

Another aspect in the study of the features was to compute the relative
distribution of the features within the host gene. In order to understand this distribution, we developed several perl programs in order to mine the length of the host gene, and then to accurately map each of the features based on their position within the gene. The positions within the gene were divided into quintiles, and the distribution of feature within each of the five quintiles was calculated in order to study any positional bias in the distribution of the gene.

**Graphical representation of the features**

A perl program was developed in order to map each feature within the host introns and then output a html file which would contain the multiple alignment of the intron group and a color marking for the start and stop of each feature within the intron of each specie. This helps us to graphically visualize the data, and html files for each feature time where intron length was less than 6000 nucleotides and all five species had atleast one such feature is available for download at http://bpg.utoledo.edu/~aprakash/DNAfeatures/.

**Discussion**

Simple repeats are abundant in the genomes of diverse animals and plants. In rodents, 2.4% of the euchromatic part of their genome is represented by simple repeats, which is two times bigger than the length of all protein-coding sequences [24]. Under certain conditions those repeats which are at mid-rande scales can exist in a variety of non-canonical DNA conformations in vivo associated with specific genomic functions. This study sheds light on the distribution and degree of conservation of these MRI features.
The results of this study suggest that introns which form a major proportion of actively transcribed section of DNA, are extremely enriched with several MRI features. These MRI features cause the DNA within these regions to exist in various non-canonical conformations under certain conditions. We are just beginning to appreciate how Multifarious DNA structures provide additional structural and functional dimensions to chromatin organization and gene expression. Several studies of DNA Conformation and Transcription have been performed to date to solve the riddle 'written' in DNA conformation ("conformation code"). The presence of these non-canonical conformations within the gene could have an impact on binding of transcription factor binding [26], provide torsional strain relief (supercoiling) while DNA transcription occurs [27]. Z-DNA formed after transcription initiation in some cases may be bound by RNA modifying enzymes, such as ADAR1, which then alter the sequence of the newly-formed RNA. A poxvirus virulence factor, called E3L, mimicked a mammalian protein that binds Z-DNA. When expressed in human cells, E3L increases by five- to ten-fold the production of several genes that block a cell’s ability to self-destruct in response to infection [28, 29]. The potential to form a Z-DNA structure also correlates with regions of active transcription [30]. It has also been shown that naturally occurring H-DNA, which are abundant can regulate the expression of several disease-linked genes and that H-DNA is a causative factor in mutagenesis and not just the end product [7].

This study reveals that several of these putative Z-DNA and H-DNA regions tend to be abundant in introns. The study also reveals that these putative non-canonical regions tend to be distributed more frequently towards the end of genes,
which could probably serve the purpose of decreasing torsional strain during the
process of transcription of the host gene. This would help the region hosting this gene
to undergo separation of the two strands of DNA without causing strand breakage
downstream of the gene. However the only exception we see is in the case of GC-
MRI regions which seem to be very highly enriched at the 5'-end of the gene. This
could probably because many of the G+C rich regions which were predicted were
CpG islands which span well into the first intron of a gene. The presence of these GC-
MRI regions could result in epigenomic modifications, especially methylation or
demethylation which would have a significant impact on the level of expression the
host gene.

The study also revealed that most of the MRI regions tend to have low levels
of sequence conservation, which mirror the average conservation of nucleotides that
one would expect to find within an intron. As discussed in previous chapters; regions
which are in the MRI range in length usually have sequences which never repeat
within the genome of the organism and so are usually unique, thus supporting the use
of nucleotide content to define these regions rather than the actual sequence. The
selection pressure on these regions are mainly due to their ability to form non-
canonical DNA conformations, and thus would allow considerable latitude as to actual
sequences but however would emphasize the conservation of their position within the
gene. This is exactly what is observed in the study as despite very low sequence
conservation of these features, all of them seem to have a very high degree of
positional conservation across the five species. The GC-MRI regions being an
exception to this in that in addition to positional conservation they also have very high
sequence conservation. This can be explained in that if their function is to undergo
methylation or other epigenomic modifications it is important to have a specific sequence pattern such as CpG etc which would afford a high degree of sequence similarity between these regions. Another exception is AT-MRI regions which tend to have very low sequence and positional conservation, which could probably point to the lack of a functional role for these regions within introns.

In summary the study validates the use of studying regions of nucleotide inhomogeneity in order to uncover as yet unknown functions within regions of DNA which do not code for proteins. The abundance of these MRI regions within introns which span >30% of the genome, suggests yet another layer of information within DNA which provides a mechanism of controlling the levels of expression of genes within the genome.
References


Chapter Six

Critical association of ncRNA with introns

Authors

David Rearick†, Ashwin Prakash†, Andrew McSweeny†, Samuel S. Shepard, Larisa Fedorova and Alexei Fedorov*

† Equal Author Contribution.
Introduction to Chapter Six

Association of Mid Range Inhomogeneity

with non coding RNA
Biological significance of MRI in the formation of various DNA conformations has been demonstrated in the previous chapter. In this chapter we focus on the significance of MRI on RNA secondary structures, and in turn the impact this has on the production of non-coding RNA (ncRNA). Support for the biological significance of nucleotide composition within the MRI range can be inferred from the size of highly conserved small non-coding RNAs involved in RNA interference (RNAi). For example, pre-microRNAs (pre-miRNAs) and pre small interfering RNAs (pre-siRNAs) are encoded by highly conserved regions with no fewer than 30 nucleotides, which place them well within the MRI range. miRNAs have been found in all forms of life on this planet from viruses, slime molds and protists to plants and animals [1, 2, 3] (Griffiths-Jones 2004; Griffiths-Jones et al. 2006; Griffiths-Jones et al. 2008) and are consistently within MRI ranges irrespective of species.

It is logical that miRNAs and siRNAs are highly specific because they have the potential to completely or partially silence gene expression for any mRNA matching their sequence. This indicates that the nucleotide composition of miRNA and siRNA are central in their ability to perform their functions within the cell, this gives rise to the strong compositional biases where they are located in the genome, and given their length it would be safe to assume that these functional ncRNAs would give rise to Mid range inhomogeneity.

Secondary structures (SS) are crucial in the biosynthesis, localization and function of ncRNA. A study by Bechtel, et.al showed the association of energetically strong SS with the phenomenon of MRI. Distinct parts of mRNAs and introns have large variations in nucleotide compositions, ranging from 35% to 60% G+C.
content.[4, 5]. The Bechtel, et.al study deals with energetically strong local SS which have a high probability of occurrence in vivo. The study demonstrated that the randomization of natural intron sequences while preserving SRI fails to preserve these strong secondary structures. While most of the strong SS are well preserved when the natural sequence is randomized by preserving MRI; using the MRI randomizer program from the GMRI web resource. This clearly proves the fact that preservation of MRI features is essential for the presence of strong SS which constitutes many of the functional ncRNA within them. Put otherwise, the presence of functional ncRNAs within introns, and their high degree of conservation across species, produces regions of nucleotide compositional bias, which manifests itself as MRI features.

Bechtel et.al, also demonstrated that there is a significant enrichment of MRI features within naturally occurring intronic sequences, as opposed to what would be expected at random. This suggests that there might also be an associated enrichment of ncRNA within introns. The presence of ncRNA within introns is a scarcely dealt topic as most studies of ncRNAs have focused on the ncRNA in intergenic regions transcribed directly from the DNA. The intergenic ncRNA are thought to possess their own promoter regions, and are independently transcribed, essentially constituting ncRNA genes. Intronic ncRNA have received little attention, and yet spliceosomal introns which are transcribed with their host genes, and are eventually spliced out to produce mRNA, provide for an excellent medium for production of ncRNA. Intronic ncRNA would not need a separate promoter for their production and also have the advantage of being able to be produce in sync with the host genes. The production of intronic ncRNA with their host genes provides a simple control for the levels of ncRNA. This finely controled ncRNA in turn could affect other proteins within the
pathway in which the host gene is involved, thus providing a finely tuned gene expression system for several genes with little regulatory mechanism.

There are several examples of ncRNA occurring within introns, such as in the case of HB-II 85 small nucleolar RNA (snoRNA) which are abundant in the introns of SNRPN genes [6]. The presence of ncRNAs within introns has clear implications for the biogenesis of both mature small RNAs and host mRNA. The balance of the interactions between the processing or ribonucleoprotein assembly of intronic noncoding RNAs and the splicing process can regulate the levels of ncRNA and host mRNA. The processing of snoRNAs - both intronic and non-intronic - is well characterised in yeast, plants and animals and provides a basis for examining how intronic plant miRNAs are processed. Apart from the vast repertoire of already characterized ncRNA and their families. There exists a vast pool of ncRNAs which are undiscovered, especially the large ncRNA molecules, most of which are uncharacterized and whose functions are yet to be clearly understood. The prediction of traditional ncRNA is mainly based on the assumption that the functional roles of ncRNA requires them to have very high degree of sequence preservation, and thus can be seen to be well preserved across most species. As mentioned above introns, provide an attractive medium for biosynthesis of ncRNA and in this study we try to study the association of known and predicted ncRNA with mammalian introns. This study deals with the distribution of the vast repertoire of already known ncRNA families within introns, and also uses predictive programs to understand the association of putative long ncRNA and putative short interfering RNA (putative siRNA) within introns.
References


Manuscript

Critical association of ncRNA with introns
Abstract

It has been widely acknowledged that non-coding RNAs are master-regulators of genomic functions. However, the significance of the presence of ncRNA within introns has not received proper attention. ncRNA within introns are commonly produced through the post-splicing process and are specific signals of gene transcription events, impacting many other genes and modulating their expression. This study, along with the following discussion, details the association of thousands of ncRNAs—snoRNA, miRNA, siRNA, piRNA and long ncRNA—within human introns. We propose that such an association between human introns and ncRNAs has a pronounced synergistic effect with important implications for fine-tuning gene expression patterns across the entire genome.

INTRODUCTION

Spliceosomal introns are ubiquitous elements of nuclear genomes. Their evolutionary rise is associated with the origin of eukaryotes [1, 2]. Recently, a new conception of the co-evolution of introns and nucleus-cytosol compartmentalization has been detailed [3]. The existence of introns allows for the alternative splicing of pre-mRNA molecules, thus serving to increase both protein diversity and specialization within the proteome [4, 5]. Additional intron functions have been reviewed [6]. However, the use of introns is a double-edged sword for organisms enriched with these elements, since they require complex processing that can lead to serious problems when splicing goes awry. Particularly, large intron sizes in vertebrate and other complex organisms incur several drawbacks including waste of energy, delay in protein production and increased vulnerability to splicing errors [7]. Having acknowledged intron roles, we will focus solely on the non-random presence of non-
protein-coding RNAs (ncRNAs) inside these gene elements. At the dawn of small ncRNA discovery, John Mattick first proposed the hypothesis that introns contain information valuable to gene regulation and called it ‘informational RNA’ [8]. Since that time a whole new field of RNomics has emerged for the investigation of ncRNAs in genetic regulation. A positive correlation between the number of ncRNAs and the complexity of an organism is evident, while the number of protein-coding genes is relatively constant from worms to humans [9]. Non-coding RNAs consist of a diverse group of short molecules including miRNAs, siRNAs, snoRNAs and piRNAs as well as various long ncRNAs. They are involved in a spectrum of regulatory processes within the nucleus and cytoplasm indispensable for the proper organization and functioning of every eukaryotic cell [see reviews [10, 11]]. The present study demonstrates how intimately ncRNAs are associated with introns.

MATERIALS AND METHODS

Databases

For the localization of small RNAs within the human genome we used our human Exon–Intron Database (EID), release 36.1 [12] and the NCBI human genome sequence, build 36.1.

Statistics on snoRNA were obtained from snoRNA-LBME-db database, version 3 [13]. This is a manually curated database with stringent requirements for experimental verification of each deposited sequence.

A comprehensive set of 462 pre-miRNA was obtained from miRBase [14]. Pre-miRNA sequences contained within this database all represent miRNA sequences that have been published in peer reviewed journals. ‘Each sequence represents a
predicted hairpin portion of the transcript [14].

A comprehensive set of 33051 human piRNA sequences was obtained from RNAdb [15]. This set of human piRNA were obtained from one laboratory using a pyrosequencing technique [16]. The authors provided experimental validation that their sequences are significantly enriched with PIWI-associated small RNA molecules (piRNA).

Complete sets of functional non-coding RNAs for human (124591 entries) and mouse (110495 entries) were obtained from functional ncRNA database (fRNAdb) [17].

**Sequence processing**

Sequences of small ncRNA were matched with the human genome using PERL regular expressions. piRNAs that had perfect matches to multiple locations within the genome were called ‘multi-match’ and were not counted in the distributions for exons, introns, or intergenic regions. The remaining ‘single-match’ ncRNA sequences that had only one perfect match to an exon or intron (transcribed strand) in the human EID were considered to be either exonic or intronic. Those single-match ncRNA sequences that were perfectly matched to complementary sequences of exons or introns from EID were designated as being ‘complementary to’ exons or introns, respectively. All other small ncRNA locations (i.e. outside of exons and introns as well as their complementary strands) were considered to be ‘intergenic’.

**miRNA**

Distances between miRNA sequences were determined using the chromosomal positions given in the miRNA annotations [14].
siRNA

In order to computationally assess the ability of human introns to produce endogenous siRNA the siRNA.pl Perl program—a modified version of the snoTARGET program (http://bpg.utoledo.edu/~dbs/snotarget)—was used. The siRNA.pl program scans the entire set of human introns, searching for stem–loop hairpin structures with perfect stems spanning at least 21 nt and with short (0–80 nt) loops. In order to understand the association of these hairpin structures with repetitive elements, we scanned the introns using siRNA.pl after masking them by Repeatmasker [18] followed by the trf (‘tandem repeats finder’) program for masking tandem repeats [19]. In order to evaluate the statistical association of hairpins with introns, a search for hairpin structures was undertaken within three control sets. The control sets were generated using our web application ‘SRI-generator’ [20] and consisted of randomized nucleotide sequences that maintained the oligonucleotide frequency composition and length of the natural set of introns. Statistical significance for the comparison of hairpin distribution between introns and control sets was established using the Fisher exact test. Similar analysis was performed within exons and intergenic regions, and the frequencies of occurrence of perfect stems were compared to those found in introns, using the chi-square test. Evolutionary conservation of the hairpins was examined by performing a BLAST search against cow, mouse and rat orthologous introns [21].

piRNA

A comprehensive set of 33051 human piRNA sequences was processed by first removing sequences with ambiguous nucleotides (e.g. ‘n’), yielding 32439 remaining
sequences. From this set, 5274 sequences had zero matches to the human genome and were removed from consideration; 22835 sequences had exactly one match and were named ‘single-match’; while each of the remaining 4330 sequences had multiple exact matches to different genomic locations and were named ‘multi-match.’ Furthermore, we showed that among the 22835 single-match piRNAs, 3138 sequences were redundant, i.e. were mapped exactly within the same site of the human genome as at least one other piRNA from this group. We removed all redundant sequencing creating the final set of 19697 single-match piRNAs, which was used for the calculation of distributions within exons, introns and intergenic regions.

The combined 19697 single-match and the 4330 multi-match sets were analyzed for their association with repetitive elements from Repbase [22] using the BLAST program without filters. Of the 24027 piRNA sequences, 1249 demonstrated significant similarity ($e < 10^{-4}$) to known repeats (5.2%). The corresponding random set was analyzed under identical conditions, yielding 1776 sequences demonstrating significant similarity ($e < 10^{-4}$) to human repeats (13.2%). Due to the short length of these sequences, a substantial number of false negatives are expected. These results were also confirmed by using RepeatMasker to mask piRNA and random sequences under sensitive conditions using the slow search option.

Among the final set of 19697 single-match sequences, 15047 were characterized as intergenic piRNAs and 4650 piRNAs were mapped within the exons and/or introns of protein-coding genes or their complementary strands. From the latter group, 300 piRNA corresponded to loci containing both exons and introns (i.e. overlapping splicing junctions, overlapping genes on opposite strands or alternate transcripts) and were excluded from the calculations regarding the exon/intron
In order to determine if there were positional preferences for piRNAs within introns, we divided each intron into quintiles (20% portions) based on the entire length of the given intron. Each piRNA sequence was assigned to a quintile based on its position within an intron. The total number of occurrences was calculated for each quintile. The positional preference of piRNAs within mRNAs was determined in a similar manner. The calculation of the standard error of means was determined using the Binomial distribution.

**Analyzing the distribution of ‘random’ ncRNAs within genomic regions**

We created a PERL program for the selection of 13500 random positions along the entire human genome. From these positions, 30-bp long sequences were collected and listed as a set of 13500 ‘random’ ncRNAs. Each of these random sequences was aligned to the entire human genome using the same protocol as for real piRNA (see previous paragraph). Among them, 2068 random sequences matched to several genomic locations and were grouped as ‘multi-match’. Each of the remaining 11432 sequences had a single match to the genome. Alignment with BLAST demonstrated that 1776 random sequences out of 13500 [13.2 ± 0.3% standard error (SE)] had a significant similarity to repetitive elements ($e < 10^{-4}$). The same proportion among the real set of piRNA comprised 5.2 ± 0.14% SE (1249/24027).

SE for each percentage was calculated using the formula $SE_p = \sqrt{p(1-p)/n}$, where $p$ is the sample proportion and $n$ is the sample size, using the Binomial distribution. A chi-square test was used to compare the distribution of piRNA sequences classified as exonic, intronic and intergenic to the distribution of 11432 randomly placed sequences within these genomic regions.
Long intronic ncRNA

A total of 63077 groups of orthologous introns for five mammalian species (human, mouse, rat, dog, cow) was obtained from the latest release (July 2010) of our Mammalian Orthologous Intron Database [21], available on our website (www.bioinfo.utoledo.edu/domino5). We defined ‘orthologous introns’ as introns from orthologous genes that have the same position and phase relative to the coding sequence.

Each group of orthologous intron sequences from the five species was aligned using MAFFT, a stand-alone program which can align a set of sequences flanking around alignable domains [23] (using the L-INS-I parameters: mafft –localpair –maxiterate 1000 input_file > output_file). A Perl program was developed to process the obtained alignments and investigate the degree of conservation among the different species. The program required that each conserved intronic region (CIR) spanned at least 400 nt in length, so as to exclude small ncRNAs from our results (Explanations in MOID web page). Additionally, CIRs qualified as evolutionarily conserved only if they had at least 50% sequence identity among the five species. This threshold was chosen to be high enough so that regions of identity occurring by chance would be eliminated, and yet low enough to take into consideration the wide degree of divergence among the five species. Various filters were applied to reduce the possibility of the conserved segment being a part of an alternatively spliced exon, as explained in [24].

The corresponding human and mouse sequences of the CIRs with masked repeats (RepeatMasker, version-3.2.8) were compared to the respective Functional non-coding RNA database, fRNAdb, [17] using the stand-alone BLAST program. The
results were parsed to enumerate the overlap with ncRNA, in instances where the BLAST score was more than 80 bits (e-value < $2 \times 10^{-16}$).

**Statistics**

Statistical analysis with the chi-square test and Fisher exact test was performed using the R package (v2.7.1).

**Programs**

The new release of our snoRNA.r3.pl mentioned in the results section is available on our website (http://www.utoledo.edu/med/depts/bioinfo/database.html). All programs used to perform calculations were written in Perl and are available upon request.

**RESULTS**

**snoRNAs, a byproduct of intron splicing in animals**

All known snoRNAs in vertebrates (and possibly in Drosophila) are a byproduct of splicing because they are created by the exonucleolytic processing of debranched introns after their excision from the pre-mRNA [25]. The vast majority of animal snoRNAs have been found inside the introns of protein-coding genes, while only a few of them have been reported to be inside the introns of ncRNAs transcribed by RNA polymerase II [26, 27]. The current release of the snoRNA-LBME-db database, version 3, contains 402 experimentally confirmed human snoRNAs [13]. The majority of them are involved in the chemical modification of 184 bases of ribosomal 28S, 18S and 5.8S rRNAs and 33 bases of spliceosomal U1, U2, U4, U5,
U6 and U12 snoRNAs. Moreover, 136 snoRNAs in this database belong to so-called orphan molecules that do not display antisense elements compatible with a modification for rRNA or snRNA. In addition to the described sample of natural snoRNAs, there are many computationally predicted snoRNA-like sequences within human introns whose existence have not been confirmed experimentally and therefore, are not featured in snoRNA-LBME-db. These snoRNA-like sequences have been identified inside genomes using several computational approaches [21, 28–31]. The computationally predicted sequences possess all the major characteristics of natural snoRNAs such as conserved sequence motifs (boxes) and secondary structures; hence, a portion of them could represent uncharacterized natural snoRNAs. Supplementary Table S1 contains a list of 324 novel C/D-box snoRNA-like sequences within human introns produced by our snoRNA.r3.pl, program [21]. We project that the total number of snoRNA-like sequences in the human genome may exceed 1000. The facts testify that the presence of introns in animals is crucial for the biosynthesis of snoRNAs.

<table>
<thead>
<tr>
<th>Genomic region</th>
<th>Number of pre-miRNA</th>
<th>Percentage of pre-miRNA</th>
<th>Number of random sequence</th>
<th>Percentage of random sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intergenic</td>
<td>237</td>
<td>51.3 ± 2.3</td>
<td>8480</td>
<td>74.2 ± 0.4</td>
</tr>
<tr>
<td>Intron</td>
<td>206</td>
<td>44.6 ± 2.5</td>
<td>2779</td>
<td>24.3 ± 0.4</td>
</tr>
<tr>
<td>Exon</td>
<td>19</td>
<td>4.1 ± 0.9</td>
<td>173</td>
<td>1.5 ± 0.1</td>
</tr>
<tr>
<td>Total</td>
<td>462</td>
<td>100</td>
<td>11,432</td>
<td>100</td>
</tr>
</tbody>
</table>

**Table 6.1: Distribution of human pre-miRNAs within exons, introns and intergenic regions.** The distribution of pre-miRNA within specific genomic regions is compared to the estimated probability of localization within these regions, calculated by classifying 11 432 sequences from randomly chosen locations within the human genome. Both transcribed and complementary strands are represented for exons and introns. Percentages are shown ± (SE).
miRNAs are significantly enriched in the transcribed strands of human introns

Table 6.1 illustrates the distribution of all known human pre-miRNAs from miRBase within the introns and exons of protein-coding genes as well as the regions between these genes, which we refer to as intergenic regions. The data demonstrates a preference of pre-miRNA to exist inside introns and exons over intergenic regions. The bias of pre-miRNA to favor genic regions while avoiding intergenic regions is statistically significant ($X^2_{21\text{df}} = 117.6; P < 2.2 \times 10^{-16}$). Among the 19 pre-miRNAs found inside exons, 5 occur within the complementary strand of the Retrotransposon-like (RTL1) gene. Their function is associated with the chromosomal methylation and regulation involved in imprinting of the RTL1 locus [32]. Only two other exonic pre-miRNAs correspond to coding regions (Supplementary Table S2), while the rest correspond to 5’- or 3’-UTRs. Ten of these are found on the transcribed strand and four are found on the complementary strand.

The distribution of pre-miRNAs inside introns is shown in Table 6.2. The data demonstrates a strong preference of pre-miRNAs to associate with the transcribed strand of introns (87%) while 13% are associated with the complementary strand. Twenty-four percent of intronic pre-miRNAs are found in clusters (two or more pre-miRNAs inside the same intron) while the majority (76%) of these pre-miRNAs are sparsely populated (one pre-miRNA per intron). In intergenic regions, there is a stronger tendency for several pre-miRNAs to be located in close proximity to each other (64% of intergenic pre-miRNAs are separated from each other by <5 kb). The largest cluster of the pre-miRNAs exists in human chromosome 19, where 42 different pre-miRNAs were found within a 150-kb region (Supplementary Table S2). Tables 6.1 and 6.2 demonstrate that 39% of all human pre-miRNAs originate from the
transcribed strand of introns, while a random distribution would put 12% on the transcribed strand as well as 12% on the complementary strand of introns. A chi-square test confirms that this association of pre-miRNA with the transcribed strand of introns is statistically significant ($\chi^2_{1 df} = 63.2, P = 1.9 \times 10^{-15}$). The calculated association of pre-miRNAs with introns is likely to be underestimated due to the dearth of information on introns located between 5'- or 3'-untranslated exons. For example, the IC-SNURF-SNRPN gene has 137 introns within the 3'-untranslated portion of the gene, which includes 94 orphan snoRNAs [33]. Moreover, the introns in the untranslated part of this gene have not been annotated properly (GenBank NG_002690.1) and thus are not classified as intronic elements. Due to such inaccuracies, we reason that 50% of all human miRNA correspond to the transcribed strand of introns and are byproducts of splicing, as is also the case of snoRNAs.

<table>
<thead>
<tr>
<th>Orientation and grouping</th>
<th>Number of pre-miRNAs (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA strand</td>
<td></td>
</tr>
<tr>
<td>Transcribed</td>
<td>179 (86.9 ± 2.4)</td>
</tr>
<tr>
<td>Complementary</td>
<td>27 (13.1 ± 2.4)</td>
</tr>
<tr>
<td>Pre-miRNA clustering</td>
<td></td>
</tr>
<tr>
<td>One per intron</td>
<td>157 (76.2 ± 3.0)</td>
</tr>
<tr>
<td>In clusters (≥ 2)</td>
<td>49 (23.8 ± 3.0)</td>
</tr>
</tbody>
</table>

Table 6.2: The distribution of pre-miRNAs inside introns. The data represents intronic pre-miRNA among transcribed and complementary strands as well as the tendency for pre-miRNA to form clusters within introns. A cluster is defined as any intron containing more than one pre-miRNA, irrespective of strand orientation. Percentages are shown ± SE.
Table 6.3: Distribution of human piRNAs within human genome. The distribution of piRNA within specific genomic regions is compared to the estimated probability of localization within these regions, calculated by classifying 11,432 sequences from randomly chose locations within the human genome. Both transcribed and complementary strands are represented for exons and introns. Percentages are shown ± SE.

<table>
<thead>
<tr>
<th>Classification</th>
<th>Number of piRNA (%)</th>
<th>Number of random sequences (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intergenic</td>
<td>15,047 (76.4 ± 0.3)</td>
<td>8,480 (74.0 ± 0.4)</td>
</tr>
<tr>
<td>Intron</td>
<td>2,349 (11.9 ± 0.2)</td>
<td>2,779 (24.3 ± 0.4)</td>
</tr>
<tr>
<td>Exon</td>
<td>2,001 (10.2 ± 0.2)</td>
<td>173 (1.5 ± 0.1)</td>
</tr>
<tr>
<td>Intron/Exon</td>
<td>300 (1.5 ± 0.1)</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>Total</td>
<td>19,697 (100)</td>
<td>11,432 (100)</td>
</tr>
</tbody>
</table>

Table 6.4: The distribution of piRNAs inside introns. The data represents the distribution of intronic piRNA among transcribed and complementary strands as well as the tendency for piRNA to form clusters within introns. A cluster is defined as any intron containing more than one piRNA, irrespective of strand orientation. Percentages are shown ± SE.

<table>
<thead>
<tr>
<th>Orientation and grouping</th>
<th>Number of piRNAs (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA strand</td>
<td></td>
</tr>
<tr>
<td>Transcribed</td>
<td>1,623 (69.1 ± 1.0)</td>
</tr>
<tr>
<td>Complementary</td>
<td>726 (30.9 ± 1.0)</td>
</tr>
<tr>
<td>piRNA clustering</td>
<td></td>
</tr>
<tr>
<td>One per intron</td>
<td>1,043 (44.4 ± 1.0)</td>
</tr>
<tr>
<td>In clusters (≥2)</td>
<td>1,306 (55.6 ± 1.0)</td>
</tr>
</tbody>
</table>
piRNAs are twice as abundant in the transcribed strand of introns as the complementary strand

piRNA sequences were mapped to the human genome and classified as multi-match or single-match. Single-match sequences were further filtered and classified as intronic, exonic or intergenic. Table 6.3 shows these results along with the distribution of 11432 randomly placed 30-nt long sequences within the human genome (‗Materials and Methods‘ section). A comparison of the estimated percentage of piRNA that is repetitive (5.2 ± 0.14% SE) with the estimated percentage of randomly located sequences mapping to repetitive regions (13.2 ± 0.3% SE) suggests that piRNA do not preferentially associate with repetitive elements, in contrast to prior observations [34].

The estimated percentage of piRNA that is intergenic (76.4 ± 0.3% SE) does not differ significantly with the estimated percentage of the human genome that is intergenic (74.2 ± 0.4% SE). We observed that 82.1% (16176/19697) of human piRNA are produced from intergenic regions (15047 piRNAs) or from the complementary strands of exons and introns (1129 piRNAs). The rest of the piRNAs are almost equally produced from the transcribed strand of exons (1598/19697; 8.1 ± 0.2% SE) and introns (1623/19697; 8.2 ± 0.2% SE). Finally, 300 piRNAs (1.5 ± 0.1% SE) overlap both exons and introns (‗Materials and Methods‘ section). Examining the data for intronic and exonic DNA from the human genome, piRNAs are significantly more likely than expected ($X^2/df = 1353.2; P < 2.2 \times 10^{-16}$) to reside in exons rather than introns, given that introns are on average, approximately 15 times larger than exons. The piRNAs classified as intronic predominantly mapped within the transcribed strand (Table 6.4). This non-random association ($X^2/df = 177.0, P < 2.2 \times 10^{-16}$) of piRNAs with the transcribed strand is slightly higher for exons (79.9 ± 1% SE) than for introns (69.1 ± 1% SE). Only 90 piRNAs overlap with the exon–intron splice sites.
and, intriguingly, are not preferentially associated with the transcribed strand (63 of them are associated with the complementary strand, Supplementary Table S3). Among the entire set of 32439 human piRNA, we found 36 sequences with a perfect match to regions overlapping the exon–exon splice junctions of mRNA, suggesting that these piRNA are produced from mature mRNA. All 36 piRNAs were found on the transcribed strand (Supplementary Table S3). The observed dearth of piRNAs overlapping exon–exon splice junctions might be explained by the tight binding of splicing proteins at splice sites with mRNA in accordance with the NMD theory [35]. This possible protection of mRNA and pre-mRNA by splicing proteins from endonucleolytic cleavage might explain the deficiency of piRNAs corresponding to exon–intron splice sites. Finally, exonic piRNAs tend to be in the internal mRNA regions and notably avoid the 3′-end (Figure 6-1A), while intronic piRNAs avoid both the 5′- and 3′-termini and prefer to localize within the central regions of introns (Figure 6-1B).
Figure 6-1: Distribution of piRNA along mRNA and introns. (A) piRNA location along each mRNA was determined by dividing mRNA into five equal segments. The total number of piRNAs within each quintile was determined. (B) The location of piRNA along introns was determined by dividing each intron into quintiles and calculated as in (A). Vertical bars show the standard error of the means.

In summary, we saw a 2.2-fold prevalence of piRNA on the transcribed strand of introns over the complementary strand. A significant enrichment of piRNA within the central regions of introns (Figure 6-1B) was observed, suggesting that a fraction of piRNAs are likely to be produced from post-spliced introns.

**Putative endogenous siRNAs within introns**

The number of endogenous siRNA molecules identified so far is quite small [36], therefore any analysis to map their positions within introns and exons at this stage would be uninformative. We instead performed a computational approach in order to assess the ability of human introns to produce endogenous siRNAs. Since hairpin siRNAs are derived from perfect double-stranded segments of RNA, we
examined the occurrences of such hairpins within the entire set of human introns which could hypothetically produce siRNAs. This computation resulted in the characterization of 8053 intronic hairpin structures within 6163 introns. These hairpins had perfect stems spanning at least 21 nt in length and a short interlude of 0-to 80-nt long loops. A vast majority of these hairpins are associated with inverted DNA repeats, while only 507 represent unique genomic hairpin sequences unrelated to repetitive DNA (Supplementary Table S4 and Supplementary Figure S1). Similar searches within the three control randomized nucleotide sequence sets derived from naturally occurring introns, yielded no hairpin structures within them. Therefore we infer that there is a statistically significant enrichment of hairpins among natural introns (Fisher exact test, $P < 2 \times 10^{-16}$). No evolutionary conservation of the non-repeat-associated set of hairpins with rodent, dog or cow genomes was found. A similar search for perfect stems within exons (total length: 58366965 nt) yielded zero occurrences of perfect stems not associated with DNA repeats. Comparison of hairpin occurrence with introns and exons suggests a significant enrichment of stems within introns compared to exons ($X^2_{1df} = 26.5, P = 2.6 \times 10^{-07}$). In a representative sample of intergenic regions (total length: 35374166 nt) there were 23 stems, which were unassociated with DNA repetitive elements, suggesting that the frequency of perfect stems in intergenic regions is similar to that within introns ($X^2_{1df} = 1.9, P = 0.17$). It is unlikely that evolutionarily conserved endogenous stem–loop (cis–trans) siRNAs are produced from introns. Nonetheless, introns might still be a source for endogenous siRNA that are derived from repetitive genomic elements, perhaps inhibiting their propagation.
Long ncRNAs inside introns

Current estimates suggest that 95% of the human genome is transcribed and produces a vast number of ncRNAs involved in different biological processes [11]. Traditionally, ncRNAs are divided into short (<200 nt) and long (>200 nt) categories according to their length [37, 38]. According to Qureshi, Mattick and Mehler, ‘a major function of long ncRNAs (lncRNAs) appears to modulate the epigenetic status of proximal and distal protein-coding genes through cis- and trans-acting mechanisms’ [39]. A considerable proportion of lncRNA exhibit low sequence conservation during evolution [37, 39, 40]. However, in 2009, it was shown that a particular type of lncRNA, known as lincRNA (long intergenic ncRNA, large intervening ncRNA) is highly conserved in mammals [40]. Intriguingly, there are also numerous evolutionarily conserved regions in mammalian introns that match the size range of lncRNAs. Recently Louro and co-authors described evolutionary conserved intronic lncRNA sequences from mouse and human [41]. Figure 6-2 demonstrates 13 long conserved regions within one of the largest mammalian introns, intron 3 of Heparanase-2 gene. The average size of these 13 conserved regions is 600 nt, although the size depends on the choice and number of species analyzed. For an example, Supplementary Figure S2 illustrates the alignment of one such conserved region from intron 3 of Heparanase-2. When the introns of a larger selection of vertebrates were aligned, the length of the conserved region became only 100 bp (Supplementary Figure S2A), while in the alignment of a smaller group of closely related species (human–mouse–cow–dog) the evolutionary conservation of the region extended to as much as 750 bp (Supplementary Figure S2B).
Figure 6-2: Evolutionarily conserved regions within the third intron of the Heparanase 2 (HPSE2) gene. The intron–exon structure of HPSE2 is shown at the top, with vertical lines depicting exons. In the bottom diagram, the cylinders depict 13 highly conserved regions with the coordinates specified below.

Using the latest release (July 2010) of our Mammalian Orthologous Intron Database (21), we performed a large-scale bioinformatic investigation of the distribution of long evolutionarily conserved regions within the entire set of 63077 introns from 8161 human genes that have orthologs in each of the four mammalian species: mouse, rat, cow and dog. Only aligned segments >400 nt with at least 50% identity within five mammalian species were taken into account. Furthermore, computational filters removed alignments that could be associated with alternative splicing (‘Materials and Methods’ section). This computation revealed 9833 CIRs with lengths exceeding 400 bp. Since there are several stringent criteria defining orthologous introns, their entire set comprises approximately one-third of the total number of human introns (approximately 180000). Therefore, the entire number of large CIRs in the human genome may be as large as 30000. When the threshold for the alignment length was increased to 600 nt, 4848 CIRs were registered. Previous work in our lab showed that distribution of conserved regions within introns is uneven and, in particular, depends on the gene function [24]. Such an abundant and uneven distribution of CIRs is in
complete accordance with the previously published results by Sironi et al. [42].

Here we present computations in order to check our hypothesis that some CIRs might represent lncRNAs. This hypothesis is strengthened by the recent experimental findings that a fraction of lncRNA is found inside introns (39,41,43–45). BLAST analysis of our 9883 large CIRs (>400 bp and >50% identity) cross-referenced with all known human and mouse ncRNAs from Functional RNA Database (fRNAdb) [17] revealed hundreds of matches between them. Particularly, we found that 415 mouse large non-coding RNA sequences experimentally obtained under the FANTOM3 project and five additional mouse ncRNAs from other sources overlap with the CIRs (Supplementary Table S5). Seventy-seven percent of these 420 mice non-coding RNAs correspond to the transcribed strand of introns, while the remaining 23% correspond to the intronic complementary strand. However, in control calculations with ‘random CIRs’ sequences having the same length and number as natural CIR set, yet placed randomly along the orthologous introns, 438 mouse ncRNA from FANTOM3 dataset matched random CIRs. Moreover, in 86% cases they occur in the transcribed strand of the introns. These results are in accord with the claim of Guttman et al. [40] that ‘current [lincRNA] catalogues may consist largely of transcriptional noise, with a minority of bona fide functional lincRNAs hidden amid this background’.

The human database of experimentally verified large ncRNA is many times smaller than the corresponding mouse set, yet the human ncRNA database contains thousands of putative computer-predicted sequences that have not been predicted for mouse. BLAST analysis of the human ncRNA sequences revealed that 1268 putative ncRNA obtained by RNAz program; 485 putative ncRNA obtained by EvoFold program; and 18 experimentally verified large ncRNA overlap with our entire set of
9833 CIRs (Supplementary Table S6). Not surprisingly, our long intronic conserved regions correspond to 1753 putative ncRNA predicted by RNAz and EvoFold, since the latter algorithms are heavily based on evolution conservation. The EvoFold program considers the evolutionary conservation of RNA secondary structures and, therefore, is capable of predicting the DNA strand which gives rise to the putative ncRNA, since conservation of secondary structure may be strand-specific. However, in many cases it is problematic to infer the orientation of ncRNA when both strands have conserved secondary structures. Among the 485 predicted ncRNA (EvoFold) that overlap with our CIR set, $60.0 \pm 2.2\%$ SE correspond to the transcribed intronic strand, while $40.0 \pm 2.2\%$ SE to the complementary strand. In control calculations with ‘random CIRs’ they matched only eight EvoFold-predicted sequences and 76 RNAz-predicted sequences from the entire human fRNA database.

The strong preference of ncRNA from intronic regions to be associated with the transcribed strand is in accordance to Nakaya et al. [46], who examined 5678 wholly intronic human ‘mRNA clusters’ computed from GenBank entries. They found that 74% of these non-coding ‘mRNA clusters’ correspond to transcribed strand of introns while 26% correspond to the complementary strand.

We conjecture that among large CIRs there may be found thousands of long functional ncRNAs originated through the post-splicing processing.

**DISCUSSION**

Our calculations demonstrate that human introns may potentially contain thousands of ncRNAs—snoRNAs, miRNAs, piRNAs and, presumably, lincRNA-like molecules. Specifically, introns are enriched with ncRNAs, which mildly regulate gene expression (miRNA and orphan snoRNA). According to Selbach et al. [47] and
Baek et al. [48], an individual miRNA modulates (predominantly down-regulates) the expression of hundreds of genes, although modestly (1.5- to 2-fold). Gene array experiments with knockout mice lacking orphan snoRNAs from the IC-SNURF-SNRPN locus revealed that such snoRNAs do not abruptly shut down or turn on genes, but rather, mildly change the expression of dozens of them [49]. Lastly, the most abundant group of small ncRNAs in humans (piRNAs), whose functions are restricted to a very specific tissue (spermatocytes), do not show a preference to be either within or outside introns. Recent articles speculate that the role of piRNAs is to defend the genome against transposable elements [50]; however, the high percentage of piRNAs not associated with repetitive elements suggests other undefined roles. This idea is supported by a new study demonstrating that piRNAs are also expressed in somatic tissues [51].

Non-coding RNAs regulate gene expression through two major pathways: (i) through transcriptional gene silencing (TGS) occurring within the nucleus, when ncRNAs, after their transcription and processing, are involved in chromatin changes and (ii) through post-transcriptional gene silencing (PTGS) occurring within the cytoplasm, when ncRNAs direct the RISC complex to target mRNAs for either cleavage or translational arrest [52]. The TGS pathway is very actively employed in plants and therefore is the most studied pathway in this taxon. Mi et al. [53] characterized more than 300000 Arabidopsis siRNAs, which are associated with nucleus-localized AGO4 protein and are specifically involved in chromatin changes and methylation. In mammals, the majority of siRNA and miRNA are associated with PTGS, which is the most studied pathway in this group. However, some mammalian miRNA are also involved in chromatin methylation and remodeling. For example, five RTL1-associated miRNAs control imprinting of the RTL1 gene [54]. In addition,
numerous mammalian piRNA and lincRNA also work through the TGS pathway [55]. We see, therefore, that both TGS and PTGS are actively engaged in higher eukaryotes. When intronic ncRNAs (such as miRNA) work via PTGS, they regulate the production of hundreds of different proteins [47, 48] some of which could include transcription factors. These transcription factors will in turn modulate the expression of other genes, (although not necessarily the parent gene that initiated this regulation event). However, auto-regulatory feedback loops within the PTGS pathway are not uncommon and have been known since the discovery of miRNAs. One of the first described miRNAs in Caenorhabditis elegans was let-7. Let-7 is regulated by a double-negative feedback loop where the miRNA inhibits the expression of lin-28 and lin-41, while the expression of these target genes inhibits let-7 [56]. Another well-known example is an intron of the Arabidopsis Dicer gene containing miRNAs that regulate the expression of its own gene [57]. Under the TGS pathway, an intronic ncRNA usually regulates the expression of its host and, potentially, neighboring genes. The regulation of multiple genes via the TGS pathway has not yet been well studied and therefore cannot be ruled out.

How precise should the regulation of genes be in healthy humans? It is well established that within the same cell type and developmental stage there is extensive individual variability in gene expression [58]. In many cases the expression levels of genes are heritable and population-specific [58]. From the perspective of thermodynamics, gene expression is a fundamentally stochastic process, with randomness in transcription and translation leading to cell-to-cell variations in mRNA and protein levels [59]. Raj and Oudenaarden emphasize that the stochastic nature of gene expression has important consequences for cellular function, being beneficial in some contexts and harmful in others [59]. In this respect, genetic diseases provide
invaluable insight into genomic operation. A majority (87% by our estimate) of the prevalent human genetic autosomal diseases are recessive, which means that one healthy copy of a gene can substitute for two functional copies without much harm. In heterozygous individuals; that is, having one mutant and one normal gene, the expression level of the corresponding protein is often reduced by up to one-half of the average level. Considering the effect of gene overproduction, when the expression level of a large group of genes is even mildly up-regulated, the consequence is usually quite devastating as observed in various cases of human trisomy. One of the most common trisomies is Down syndrome where three copies of chromosome 21 (or a portion of 21) occur in the patient’s karyotype. The phenotype is characterized with some impairment of cognitive ability and physical growth as well as facial abnormalities. A partial trisomy of chromosome 21 can be as small as 2–3 Mb, representing 200 genes with expression levels being elevated 1.5 times on average [60]. Perturbed expression of genes on other autosomes as a result of trisomies, such as chromosomes 8, 12, 13 and 18, cause more severe conditions such as Warkany syndrome, chronic lymphocytic leukemia, Patau syndrome and Edward’s syndrome respectively [60]. Partial trisomies of these chromosomes produce milder symptoms. The two most frequent autosomal trisomies in humans, 16 and 22, are the most common chromosomal causes of spontaneous first trimester abortions [61]. Partial trisomies of the remaining chromosomes are less common and often result in conditions ranging from few phenotypic symptoms, as in the case of Cat eye syndrome (22pter→q11), to lethal birth defects as in the case of chromosome 14 [62]. Therefore, even mild up-regulation of a large group of genes is usually deleterious to an organism [60]. In 2002 Yan et al. [63] showed that mammals, similar to plants, have allele-specific expression (ASE) of genes also known as allelic imbalance. This
heritable allelic variation in gene expression was shown to be a common phenomenon within the human genome [64]. De la Chapelle emphasizes the surprising extent of genomic regulation resulting from ASE [65]. Many types of ASE dramatically influence susceptibility to disorders such as cancer, autoimmune diseases and diabetes [66, 65]. It is well documented that ASE is governed by cis-regulatory elements, yet the particular type and location of these elements is yet to be verified and therefore is debatable. De la Chapelle argues that the cis-elements responsible for ASE are likely to be miRNA and lincRNAs [65]. From this standpoint, intronic ncRNA are outstanding candidates for the regulation of allelic imbalance via the TGS pathway. The aforementioned example of the five miRNAs that shut down the expression of the maternal RTL1 allele validates the ability of ncRNAs to have allele-specific precision [54].

Despite the permissible variations in the expression of many individual genes, the entire ensemble of genes must be highly coordinated. Only minor fluctuations in the expression of a number of genes are allowed in healthy humans. Such coordinated regulation of thousands of genes in a cell is unimaginable without numerous feedback loops engaged in the gene expression system. Intronic ncRNAs are perfect elements for such a feedback regulation system. Indeed, intronic ncRNAs are co-produced with the mRNA of their host genes. When a host gene is silent, its pool of ncRNAs is also not produced. However, during transcription, the production of intronic ncRNAs is strictly proportional to the expression level of the host gene. It becomes clear that the fundamental significance of many introns is to provide regulatory ncRNAs for the fine control of genes within complex higher organisms. This view of the subtle yet inextricable value of introns in genomic functioning is what we term the Symbiotic Intron Hypothesis. This hypothesis proposes a ‘non-selfish’ harmony between genes,
introns and ncRNAs within higher eukaryotes. Genes provide space for introns inside of them. In turn, introns act as hosts for regulatory ncRNAs. Finally, ncRNAs provide essential regulation for the expression of genes. We conclude, therefore, that there is a natural symbiosis, between genes, introns and ncRNAs—a symbiosis that is only just beginning to be discovered and properly appreciated.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

FUNDING
National Science Foundation Career award ‘Investigation of intron cellular roles’ (grant number MCB-0643542). Funding for open access charge: National Science Foundation Career award ‘Investigation of intron cellular roles’ (grant number MCB-0643542).

Conflict of interest statement. None declared.

ACKNOWLEDGEMENTS
D.R. and A.M. were responsible for the computational processing and analysis of the miRNA and the piRNA sections of the article. D.R. was responsible for co-referencing CIRs with the mouse database and the analysis of Figure 2. A.M. was responsible for the statistical analysis conducted through the article. A.P. was responsible for the processing and analysis of the siRNA and long ncRNA sections of the article, and was responsible for creating the Mammalian Orthologous Intron Database, their multiple alignments and the websites. S.S.S. was responsible for RepeatMasking of introns for the siRNA and piRNA sections, editing and providing
guidance while writing the draft. A.F. and L.F. supervised the project, provided guidance and wrote the draft. All authors have read and approved the article.
REFERENCES


inhomogeneity correlates with an abundance of RNA secondary structures. BMC Genomics, 9, 284.


Source Code

Programs used in this project can be downloaded from GMRI web tool kit:

1) ZDNA_Analyzer.java
2) HDNA_Analyzer.java
3) GCMRI_Analyzer.java
4) GTMRI_Analyzer.java
5) ATMRI_Analyzer.java
6) Feature_visualizer.pl
7) MRI_Analyzer.pl
8) SRI_Analyzer.pl
9) MRI_Visualizer.pl
10) SRI_Generator.pl
11) MRI_Generator.pl
import java.io.*;
import java.util.regex.*;

class seqProcess {
    String seq;
    int count(String s, String a) {
        // count(<sequence>, <string of characters to count>) outputs an int
        this.seq = s;
        int nt = 0;
        for(int i = 0; i < seq.length(); i++) {
            for(int j = 0; j < a.length(); j++) {
                if(seq.charAt(i) == a.charAt(j)) {nt = nt + 1;}
            }
        }
        return nt;
    }

    String adjust_window(String s, int a, int b) {
        // adjust_window(<sequence>, <start>, <needed window length>)
        this.seq = s;
        int start = a;
        int win_len = b;
        String win = "";
        for (int i = win_len; i <= (seq.length() - start); i++) {
            win = seq.substring(start, start + i);
            seqProcess obj = new seqProcess();
            int not_nt = obj.count(win, ".");
            if (win_len == i - not_nt) {
                break;
            }
        }
        return win;
    }
}
int stretchWindow(String s, int a, int b, int c) {
    int win_len = b;
    this.seq = s;
    int start = a;
    int upthrsh = c;
    for (int j=win_len; j<seq.length()-start; j++) {
        String window = seq.substring(start, start+j);
        //System.out.println(window + "t" + seq.length() + "t" + start + "t" + j);
        seqProcess str = new seqProcess();
        int zdna = str.calcDNfreq(window, j);
        float ratio = (float) zdna / (float) j * 100;
        if(ratio < upthrsh) { win_len = j-1; break; }
    }
    return win_len;
}

int calcDNfreq(String s, int a) {
    int win_len = a;
    this.seq = s.replace("\-", ");
    String RY[] = {"at", "gt", "ac", "gc"};
    String YR[] = {"ta", "tg", "ca", "cg"};
    int zdna = 0;
    for (int i = 0; i<=seq.length()-2; i++) {
        String DN = seq.substring(i, i+2);
        for(int j = 0; j<4; j++) {
            if(DN.equals(RY[j])) {
                zdna++;
            } else if(DN.equals(YR[j])) {
                zdna++;
            }
        }
    }
    return zdna;
}

String org_name(int a) {
    String name;
    switch (a) {
        case 2:
            name = "hs";
            break;
    }
break;
case 3:
    name = "bt";
    break;
case 4:
    name = "cf";
    break;
case 5:
    name = "mm";
    break;
case 6:
    name = "rn";
    break;
default:
    name = "org";
    break;
}
return(name);
}

class ZDNA_Analyzer
{
    public static void main (String args[])
    {
        try{
            BufferedReader in = new BufferedReader(new FileReader(args[0]));
            //---------------------------------------------------------------------------
            String line;
            int $win_len = Integer.parseInt(args[1]);
            int $upthr = Integer.parseInt(args[2]);
            //int $lothr = Integer.parseInt(args[3]);
            String ORG = args[3];
            int $count = 0;
            int $organism = 0;
            String comment = "";
            String elements = "";
            int $intlen =0;
            //--------------------------------------------------------------------------------
while ((line = in.readLine())!= null) {
    String window;
    if($count >= 1) {
        $count++;
        String organism;
        line = line.replace("-", " ");
        int $regions = 0;
        for(int $i = 0; $i<=(line.length()-$win_len); $i++) {
            window = line.substring($i, $i+$win_len);
            boolean a = window.startsWith("-");
            if(a == true){continue;}
            seqProcess o = new seqProcess();
            organism = o.org_name($count);
            if(organism.equals(ORG)) {$intlen = line.length();}
            int $zdna = o.calcDNfreq(window, $win_len);
            if($zdna >= $upthr) {
                $regions++;
                //organism = o.org_name($count);
                int $new_len = o.stretchWindow(line, $i, $win_len, $upthr);
                int $end = $i+$new_len-1;
                elements = elements+"\n"+organism+"t"+"t"+$i+"t"+$end;
                //System.out.println(organism+"t"+$i+"t"+$end);
                $i+=$new_len;
            }
        }
        //if($regions>0){
        //    $organism++;
        //}
        if($organism==5) {
            if($intlen >0) {
                System.out.println(comment+" \nHuman Intron Length:"+$intlen+" \n"+elements+" \n\n");
                $organism=0;
            }
        }
    }
}
boolean b = line.startsWith(">");
if(b == true) {
    comment = line;
    //System.out.println(comment);
    $organism=0;
    $count++;
    continue;
}
if($count==6) {$count=0; elements=""; $intlen=0;}
}
catch(Exception e) {
    System.err.println(e);
    System.err.println("ERROR!!\n$ java class <In_FILE> <Window size> <UP_Threshold> <Low_Threshold> <X-type> ");
}
}
//Program Name: HDNA_Analyzer.java
//Author: Ashwin Prakash
//License: GPLv3.0

import java.io.*;
import java.util.regex.*;

class seqProcess
{
    String seq;

    int count(String s, String a) {
        //count(<sequence> , <string of characters to count>) outputs an int
        this.seq = s;
        int nt = 0;
        for(int i = 0; i < seq.length(); i++) {
            for(int j =0; j< a.length(); j++){
                if(seq.charAt(i) == a.charAt(j)){nt = nt+1;}
            }
        }
        return nt;
    }

    String adjust_window(String s, int a, int b) {
        // adjust_window(<sequence> ,<start>, <needed window length>)
        this.seq = s;
        int start = a;
        int win_len = b;
        String win = "";
        for (int i = win_len; i<=(seq.length()-start); i++){
            win = seq.substring(start, start+i);
            seqProcess obj = new seqProcess();
            int not_nt = obj.count(win, "-" );
            if(win_len == i-not_nt){
                break;
            }
        }
        return win;
    }
}

151
int stretchWindow(String s, int a, int b, int c) {
    int win_len = b;
    this.seq = s;
    int start = a;
    int upthrsh = c;
    for (int j=win_len; j<seq.length()-start; j++) {
        String window = seq.substring(start, start+j);
        //System.out.println(window+"t"+seq.length()+"t"+start+"t"+j);
        seqProcess str = new seqProcess();
        int hdna = str.ca
cDNfreq(window, j);
        float ratio = (float) hdna/ (float) j*100;
        if(ratio < upthrsh) { win_len = j-1; break;}
    }
    return win_len;
}

int calcDNfreq(String s, int a) {
    int win_len = a;
    this.seq = s.replace("-", "");
    String RY [] = {"at", "gt", "ac", "gc"};
    String YR [] = {"ta", "tg", "ca", "cg"};
    int hdna = 0;
    for (int i = 0; i<=seq.length()-2; i++) {
        String DN = seq.substring(i, i+2);
        for(int j = 0; j<4; j++) {
            if(DN.equals(RY[j])){
                hdna++;
            } else if(DN.equals(YR[j])){
                hdna++;
            }
        }
    }
    return hdna;
}

String org_name(int a){
    String name;
    switch (a){
    case 2:
class HDNA_analyzer
{
    public static void main (String args[])
    {
        try{
            BufferedReader in = new BufferedReader(new FileReader(args[0]));
            String line;
            int $win_len = Integer.parseInt(args[1]);
            int $upthr = Integer.parseInt(args[2]);
            String ORG = args[3];
            int $count = 0;
            int $organism = 0;
            String comment = "";
            String elements = "";
            int $intlen =0;
        }
    }
}
while ((line = in.readLine())!= null) {
    String window;
    if($count >= 1) {
        $count++;
        String organism;
        line = line.replace("-", "");
        int $regions = 0;
        for(int $i = 0; $i<=(line.length()-$win_len); $i++) {
            window = line.substring($i, $i+$win_len);
            boolean a = window.startsWith("-");
            if(a == true) {continue;}
            seqProcess o = new seqProcess();
            int $dna = o.calcDNfreq(window, $win_len);
            if($dna >= $upthr) {
                $regions++;
                organism = o.org_name($count);
                int $new_len = o.stretchWindow(line, $i, $win_len, $upthr);
                int $end = $i+$new_len-1;
                elements = elements+"\n"+organism+"t"+$i+"t"+$end;
                if(organism.equals(ORG)) {$intlen = line.length();}
                //System.out.println(organism+"t"+$i+"t"+$end);
                $i+=$new_len;
            }
        }
        if($regions>0) {
            $organism++;
        }
    }
    if($organism==5) {
        if($intlen >0) {
            System.out.println(comment+"\nHuman Intron Length:"+$intlen+"\n"+elements+"\n\n");
        }
        $organism=0;
    }
boolean b = line.startsWith(">");
if(b == true) {
    comment = line;
    //System.out.println(comment);
    $organism=0;
    $count++;  
    continue;
}
if($count==6) {$count=0; elements="", $intlen=0;}
}catch(Exception e){
    System.err.println(e);
    System.err.println("ERROR!!\n java class <In_FILE> <Window size> <UP_Threshold> <Low_Threshold> <X-type>");
}
```java
import java.io.*;
import java.util.*;
import java.lang.*;

class seqProcess {
    String seq;
    int $Rhdna=0;
    int $Yhdna=0;

    int count(String s, String a) {
        //count(<sequence>, <string of characters to count>) outputs an int
        this.seq = s;
        int nt = 0;
        for(int i = 0; i < seq.length(); i++) {
            for(int j = 0; j < a.length(); j++){
                if(seq.charAt(i) == a.charAt(j)) {nt = nt+1;}
            }
        }
        return nt;
    }

    int stretchWindow(String s, int a, int b, int c) {
        int win_len = b;
        this.seq = s;
        int start = a;
        int upthrsh = c;
        for (int j=win_len; j<seq.length()-start; j++) {
            String window = seq.substring(start, start+j);
            seqProcess str = new seqProcess();
            int gc_count = str.count(window, "gc");
            float gc_ratio = (float) gc_count * 100/ (float) j;
        }
    }
}
```
if(ct_ratio<upthrsh)
{
    win_len=j-1; break;
}
}
return win_len;

String adjust_window(String s, int a ,int b){
    // adjust_window(<sequence> ,<start>, <needed window length>)
    this.seq = s;
    int start = a;
    int win_len = b;
    String win = "";
    for (int i = win_len; i<=(seq.length()-start); i++){
        win = seq.substring(start, start+i);
        if(win.startsWith("-")){break;}
        seqProcess obj = new seqProcess();
        int not_nt = obj.count(win, "- ");
        if(win_len == i-not_nt){
            break;
        }
    }
    return win;
}

String org_name(int a){
    String name;
    switch (a){
    case 0:
        name = "hs";
        break;
    case 1:
        name = "bt";
        break;
    case 2:
name = "cf";
break;

case 3:
    name = "mm";
    break;

case 4:
    name = "rn";
    break;

default:
    name = "org";
    break;
}
return(name);
}
}

class GCMRI_analyzer{
    public static void main(String [] args) {
        try{
            //-------------------------------------------------------------------------
            BufferedReader in = new BufferedReader(new FileReader(args[0]));
            int $win_len = Integer.parseInt(args[1]);
            int $thr = Integer.parseInt(args[2]);

            $thr = (int) ($thr*$win_len/100);
            String line;
            String comment="";
            Vector<String> seq = new Vector<String>();
            seqProcess obj = new seqProcess();
            int $linCount=0;
            int $intcountr=0;
            //-------------------------------------------------------------------------

            while((line = in.readLine())!= null) {
                String window;
                boolean c = line.startsWith(">");
                if(c == true) {
                    //$/intcountr++;
                    //if($intcountr==20) {break;}
                    comment = line;
                    //-------------------------------------------------------------------------
                    seq.add(line);
                    $linCount++;
                    //-------------------------------------------------------------------------
                }
            }
            //-------------------------------------------------------------------------
        }
    }
}
$linCount=1;
continue;
}
if($linCount>=1){
$linCount++;
seq.add(line);
if($linCount==7){
    String output ="";
    int $hum_hit=0;
    int $organisms=0;
    int $length_sequence=0;
    for(int $i=0; $i<seq.size(); $i++) {
        String intron = seq.get($i);
        intron = intron.replace("-", "");
        int $regions=0;
        String org = obj.org_name($i);
        if(org.equals("hs")) {
            $length_sequence = intron.length();
            for(int $j=0; $j<(intron.length()-$win_len; $j++) {
                window = obj.adjust_window(intron, $j, $win_len);
                if(window.startsWith("-")) {continue;}
                int $gc_count = obj.count(window, "gc");
                if($gc_count>$thr) {
                    $regions++;
                    int $adj_len = obj.stretchWindow(intron, $j, $win_len, $thr);
                    int $end = $i+$adj_len-1;
                    window = window.replace("-", "");
                    output = output+org+" "+$j+" +(\$adj_len+$j-1)+"n"
                    $j += $adj_len;
                }
            }
            if($regions>0){
                $organisms++;
            }
        }
    }
    if($organisms>0){
        if($length_sequence>0) {
            System.out.println(comment+"Human Intron Length:"+$length_sequence+"n"+output);
        }
    }
}
} 
} 
} 
seq.clear(); 
$linCount=0; 
continue;
 
} 
} 
} 
}catch(Exception e) {
    System.out.println(e); 
} 
} 
} 
}
import java.io.*;
import java.util.*;
import java.lang.*;

class seqProcess
{
    String seq;
    int $Rhdna=0;
    int $Yhdna=0;

    int count(String s, String a) {
        //count(<sequence>, <string of characters to count>) outputs an int
        this.seq = s;
        int nt = 0;
        for(int i = 0; i < seq.length(); i++) {
            for(int j =0; j< a.length(); j++){
                if(seq.charAt(i) == a.charAt(j)){nt = nt+1;}
            }
        }
        return nt;
    }

    int stretchWindow(String s, int a, int b, int c){
        int win_len = b;
        this.seq = s;
        int start = a;
        int upthrsh = c;
        for (int j=win_len; j<seq.length()-start; j++){
            String window = seq.substring(start, start+j);
            seqProcess str = new seqProcess();
            int at_count = str.count(window, "ac");
            float at_ratio = (float) at_count *100/ (float) j;
            if(at_ratio<upthrsh){
                win_len =j-1; break;
            }
        }
        return win_len;
    }
}
String adjust_window(String s, int start, int win_len) {
    // adjust_window(<sequence>, <start>, <needed window length>)
    this.seq = s;
    int start = a;
    int win_len = b;
    String win = "";
    for (int i = win_len; i <= (seq.length() - start); i++) {
        win = seq.substring(start, start + i);
        if (win.startsWith("-")) {break;}
        seqProcess obj = new seqProcess();
        int not_nt = obj.count(win, "-");
        if (win_len == i - not_nt) {
            break;
        }
    }
    return win;
}

String org_name(int a) {
    String name;
    switch (a) {
    case 0:
        name = "hs";
        break;
    case 1:
        name = "bt";
        break;
    case 2:
        name = "cf";
        break;
    }
case 3:
    name = "mm";
    break;

case 4:
    name = "rn";
    break;

default:
    name = "org";
    break;
}
return(name);

class ATMRI_analyzer{
    public static void main(String [] args){
        try{
            BufferedReader in = new BufferedReader(new FileReader(args[0]));
            int $win_len = Integer.parseInt(args[1]);
            int $thr = Integer.parseInt(args[2]);

            $thr = (int) ($thr*$win_len/100);
            String line;
            String comment="";
            Vector<String> seq = new Vector<String>();
            seqProcess obj = new seqProcess();
            int $linCount=0;
            int $intcount=0;
            //---------------------------------------------------------------

            while((line = in.readLine())!= null){
                String window;
                boolean c = line.startsWith(">");
                if(c == true) {
                    //$intcount++;
                    //if($intcount==20){break;}
                    comment = line;
                    $linCount=1;
                    continue;
                }
            }
            //---------------------------------------------------------------

            return;
        }
    }
}
if($linCount>=1){
    $linCount++;  
    seq.add(line);  
    if($linCount==7){
        String output ="";  
        int $hum_hit=0;  
        int $organisms=0;  
        int $length_sequence=0;  
        for(int $i=0; $i<seq.size(); $i++) {
            String intron = seq.get($i);  
            intron = intron.replace("-", "");  
            int $regions=0;  
            String org = obj.org_name($i);  
            if(org.equals("hs")) {$length_sequence = intron.length();}
            for(int $j=0; $j<(intron.length())-$win_len; $j++) {  
                window = obj.adjust_window(intron, $j, $win_len);  
                if(window.startsWith("-")) {continue;}  
                int $at_count = obj.count(window, "at");  
                if($at_count>$thr) {
                    $regions++;  
                    int $adj_len = obj.stretchWindow(intron, $j, $win_len, $thr);  
                    int $end = $i+$adj_len-1;  
                    window = window.replace("-", "");  
                    output = output+org+" "+$j+" "+($adj_len+$j-1)+"n";  
                    $j += $adj_len;
                }
            }
            if($regions>0) {
                $organisms++;  
            }
        }
        if($organisms>0) {
            if($length_sequence>0) {  
                System.out.println(comment+"nHuman Intron Length:"+$length_sequence+"n"+output);
            }
        }
    }
}
seq.clear();
$linCount=0;
continue;
}
}
}
catch(Exception e) {
    System.out.println(e);
}
}
import java.io.*;
import java.util.*;
import java.lang.*;

class seqProcess {
    String seq;
    int $Rhdna=0;
    int $Yhdna=0;

    int count(String s, String a) {
        //count(<sequence> , <string of characters to count>) outputs an int
        this.seq = s;
        int nt = 0;
        for(int i = 0; i < seq.length(); i++) {
            for(int j =0; j< a.length(); j++){
                if(seq.charAt(i) == a.charAt(j)){nt = nt+1;}
            }
        }
        return nt;
    }

    int stretchWindow(String s, int a, int b, int c){
        int win_len = b;
        this.seq = s;
        int start = a;
        int upthrsh = c;
        for (int j=win_len; j<seq.length()-start; j++){
            String window = seq.substring(start, start+j);
            seqProcess str = new seqProcess();
            int gt_count = str.count(window, "gt");
            int ac_count = str.count(window, "ac");
            float gt_ratio = (float) gt_count *100/ (float) j;
        }
    }
}

//Program Name:  GTMRI_Analyzer.java
//Author:  Ashwin Prakash
//License:  GPLv3.0
float ac_ratio = (float) ga_count *100/ (float) j;
if(gt_ratio<upthrsh && ac_ratio<upthrsh){
    win_len=j-1; break;
}
}
return win_len;

String adjust_window(String s, int a ,int b){
    // adjust_window(<sequence>,<start>, <needed window length>)
    this.seq = s;
    int start = a;
    int win_len = b;
    String win = "";
    for (int i = win_len; i<=(seq.length()-start); i++){
        win = seq.substring(start, start+i);
        if(win.startsWith("-")){break;}
        seqProcess obj = new seqProcess();
        int not_nt = obj.count(win, "-"REFERRED"");
        if(win_len == i-not_nt){
            break;
        }
    }
    return win;
}

String org_name(int a){
    String name;
    switch (a){
    case 0:
        name = "hs";
        break;
    case 1:
        name = "bt";
        break;
    case 2:
        name = "cf";
        break;
    case 3:
name = "mm";
break;
case 4:
    name = "rn";
break;
default:
    name = "org";
break;
}
return(name);
}

class GTMRI_analyzer {
    public static void main(String [] args){
        try{
            //-------------------------------------------------------------------------
            BufferedReader in = new BufferedReader(new FileReader(args[0]));
            int $win_len = Integer.parseInt(args[1]);
            int $thr = Integer.parseInt(args[2]);
            $thr = (int) ($thr*$win_len/100);
            String line;
            String comment="";
            Vector<String> seq = new Vector<String>();
            seqProcess obj = new seqProcess();
            int $linCount=0;
            int $intcountr=0;
            //-------------------------------------------------------------------------

            while((line = in.readLine())!= null) {
                String window;
                boolean c = line.startsWith(">");
                if(c == true) {
                    //$intcountr++;
                    //if($intcountr==20) {break;}
                    comment = line;
                    $linCount=1;
                    continue;
                }
            }
        }
    }
}
if($linCount>=1){
    $linCount++;
    seq.add(line);
    if($linCount==7){
        String output ="";
        int $hum_hit=0;
        int $organisms=0;
        int $length_sequence=0;
        for(int $i=0; $i<seq.size(); $i++) {
            String intron = seq.get($i);
            intron = intron.replace("-", ");
            int $regions=0;
            String org = obj.org_name($i);
        if(org.equals("hs")){
                    $length_sequence = intron.length();
            }
            for(int $j=0; $j<intron.length()-$win_len; $j++){
                window = obj.adjust_window(intron, $j, $win_len);
            if(window.startsWith("-")){continue;}
                int $ct_count = obj.count(window, "gt");
            if($ct_count>$thr){
            $regions++;
        int $adj_len = obj.stretchWindow(intron, $j, $win_len, $thr);
            int $end = $i+$adj_len-1;
            window = window.replace("-", ");
        output = output+org+" "+$j+" "+($adj_len+$j-1)+"n";
        $j += $adj_len;
}        if($regions>0){
            $organisms++;
        }
    }
}
    if($organisms>0){
        if($length_sequence>0)
            System.out.println(comment+"nHuman Intron Length:"+$length_sequence+"n"+output);
    }
}
seq.clear();
$nlinCount=0;
continue;

} catch(Exception e) {
    System.out.println(e);
}
}
#!/usr/bin/perl

open(FILE, "$ARGV[0]") || die "Can not open $ARGV[0]\n";
$aln_file = $ARGV[1];

$/ = "\n";
print "<html><body>

while(<FILE>){
    $count++;
    chomp($_);
    my @line = split("\n", $_);
    print "<table width=\"875\">"

    if($_ =~ /chromosome\s+(.*);\s+\w+;\s+(INTRON_\d+\-[A-Z]*_NT_\d+)/) {
        my $alnmnt = `grep -A6 "$2" $aln_file/chr$1_result_int_grp_aln.fa`;
        my @seq = split("\n", $alnmnt);
        print "<table width=\"875\"><tr><td><font face="monospace" size=\"4\">$line[0]</font></td></tr>"
        print "<table width=\"875\">"

        $track_hs = &add_track($_, $seq[1], "hs");
        $track_bt = &add_track($_, $seq[2], "bt");
        $track_cf = &add_track($_, $seq[3], "cf");
        $track_mm = &add_track($_, $seq[4], "mm");
        $track_rn = &add_track($_, $seq[5], "rn");

        &format_print($track_hs, $track_bt, $track_cf, $track_mm, $track_rn, $seq[1], $seq[2], $seq[3], $seq[4], $seq[5], $seq[6]);
        print "</table></body></html>
    
    }

}


sub format_print{
    my $track_hs = $_[0]; my $track_bt = $_[1]; my $track_cf = $_[2];
    my $track_mm = $_[3]; my $track_rn = $_[4];

    my $hs = $_[5]; my $bt = $_[6]; my $cf = $_[7]; my $rn = $_[8]; my $mm = $_[9]; my $star = $_[10];
    my $nths = $ntbt = $ntcf = $ntmm = $ntrn = 0;

    for(my $i = 0; $i<length($hs); $i+=80) {
        my $subtrk_hs = substr($track_hs, $i, 80);
        my $subtrk_bt = substr($track_bt, $i, 80);
        my $subtrk_cf = substr($track_cf, $i, 80);
        my $subtrk_mm = substr($track_mm, $i, 80);
        my $subtrk_rn = substr($track_rn, $i, 80);

        my $sub_hs = substr($hs, $i, 80);
        my $sub_bt = substr($bt, $i, 80);
        my $sub_cf = substr($cf, $i, 80);
        my $sub_mm = substr($mm, $i, 80);
        my $sub_rn = substr($rn, $i, 80);

        $nths += $sub_hs =~ tr/acgt/acgt/;
        $ntbt += $sub_bt =~ tr/acgt/acgt/;
        $ntcf += $sub_cf =~ tr/acgt/acgt/;
        $ntmm += $sub_mm =~ tr/acgt/acgt/;
        $ntrn += $sub_rn =~ tr/acgt/acgt/;

        &color_print($subtrk_hs, $sub_hs, "#8B0000", "H", $nths, "hs", "#F4A460");
        &color_print($subtrk_bt, $sub_bt, "yellow", "B", $ntbt, "bt", "#20B2AA");
        &color_print($subtrk_cf, $sub_cf, "#00008B", "C", $ntcf, "cf", "#BDB76B");
        &color_print($subtrk_mm, $sub_mm, "brown", "M", $ntmm, "mm", "#8FBC8F");
        &color_print($subtrk_rn, $sub_rn, "green", "R", $ntrn, "rn", "#CD853F");
        $sub_star =~ s/ /&nbsp/; /font face="monospace" size="4" color="white">&nbsp</font>/g;
        print "<tr><td></td><td><font face="monospace" size="4">$sub_star</font></td><td></td></tr>");
        print "<tr><td></td><td></td><td></td></tr>");
        print "<tr><td></td><td></td><td></td></tr>");
    }
}

172
sub color_print{
    my $track = $_[0];
    my $seq = $_[1];
    my $color = $_[2];
    my $char = $_[3];
    my $nt = $_[4];
    my $org = $_[5];
    my $bgcolor = $_[6];

    my @trk = split("", $track);
    my @seq = split("", $seq);
    print "<tr><td><font face="monospace" size="4">$org</font></td><td>
        for my $j (0..$#seq){
            if($trk[$j] eq $char){
                print "<font style="background-color:$bgcolor;" face="monospace" size="4" color="$color">$seq[$j]</font>";
            }else{
                print "<font face="monospace" size="4">$seq[$j]</font>";
            }
        }
        print "</td><td><font face="monospace" size="4">$nt</font></td></tr>";
}

sub add_track{
    my $intron = $_[0];
    my $seq = $_[1];
    my $org = $_[2];
    my $track = "";
    my %coord = ();

    if($org eq "hs"){ $char = "H"; }
    elsif($org eq "bt") { $char = "B"; }
    elsif($org eq "cf") { $char = "C"; }
    elsif($org eq "mm") { $char = "M"; }
    elsif($org eq "rn") { $char = "R"; }
}
my @entry = split("\n", $intron);
for my $i(1..$#entry) {
    if($entry[$i] =~ /^[\d]+\s+([\d]+)\s+([\d]+)\s+$/) {
        #print "$1\t$2\t$3\n";
        if($1 eq $org) {
            $coord{$2} = $3;
        }
    }
}
for(my $x=1; $x<=length($seq); $x++) {
    my $fragmnt = substr($seq, 0, $x);
    my $nt = $fragmnt =~ tr/acgt//;
    if($coord{$nt}) {
        $pos = $nt;
        #print "$nt-$coord{$nt}-\$x\n";
        $track .= $char;
    } elsif($pos>0 && $nt<$coord{$pos}) {
        #print "$nt\t", length($fragmnt), "\n";
        $positive++;
        $track .= $char;
    } else {
        $track .= " ";
        $pos = 0;
    }
    #print "$nt\t\$x\t", length($track), "\n";
}
return $track;
}