Computational analysis of the evolution of non-coding genomic sequences

Arnab Saha Mandal

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

Recommended Citation
http://utdr.utoledo.edu/theses-dissertations/144

This Thesis 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 Thesis

entitled

Computational Analysis of the Evolution of Non-Coding Genomic Sequences

by

Arnab Saha Mandal

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the

Master of Science Degree in Biomedical Sciences

__________________________
Dr. Alexei Fedorov, Committee Chair

__________________________
Dr. Sadik Khuder, Committee Member

__________________________
Dr. Robert Trumbly, Committee Member

__________________________
Dr. Edwin Sanchez, Committee Member

__________________________
Dr. Patricia R. Komuniecki, Dean
College of Graduate Studies

The University of Toledo

August, 2013
An Abstract of

Computational Analysis of the Evolution of Non-coding Genomic Sequences

by

Arnab Saha Mandal

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the
Master of Science Degree in Biomedical Sciences

The University of Toledo

August, 2013

The precipitous drop in costs and increase in sequencing efficiency; with simultaneous
development of annotation technologies have made way for the availability of numerous
eukaryote whole genome sequences. The efficient use of such cornucopia of genomic
data housed in several databases has also necessitated increased application of
bioinformatic tools and amalgamation of several interdisciplinary studies. This thesis,
constituted by three chapters, provides examples of such use of genomic data from three
different perspectives. Firstly, a study of population genomes from more than a thousand
people provided by the 1000 genomes consortium is conducted across genomic loci
corresponding to four genes in different chromosomes. Genetic variants across
population genomes were also examined in the light of biased gene conversion theory.
Secondly, the profound non-randomness replete within the three billion nucleotides of
human genome is also quantitatively characterized for arrangements of specific
nucleotide patterns, which reflect on unknown signals critical for human genome
organization. Thirdly, evolutionary studies on groups of more than 40,000 orthologous
plant introns are conducted to reveal conserved sequence elements in non-coding DNA,
which are in turn characterized and compared across animals. The functional significance of conserved regions of non-coding DNA is ascertained on a case by case basis. A novel putative tRNA-like sequence is also discovered in this process within five plant introns.
I dedicate this thesis to His Holiness and the Almighty Lord Shiva for bestowing his kind blessings onto me at every stage of my life. I also dedicate this work to my parents Tushar Kanti Saha Mandal and Bela Saha Mandal who have been the only visible manifestations of God that I have known and who always gave everything they had so that I could live with plenty.
Acknowledgements

I take this opportunity to thank my advisor Dr. Alexei Fedorov for being an excellent mentor, a great teacher and an ideal role model for instilling in me the love for science towards dedicated pursuits of uncovering truths. I am deeply indebted to my teachers Dr. Robert Blumenthal and Dr. Sadik Khuder for several invaluable professional and personal lessons, without which the degree would have been far off from fruition. I also thank Dr. Robert Trumbly, Dr. Edwin Sanchez and Dr. Tammy Morrish for their help in my research also for the excellent lessons I learnt from their classes. I also thank Dr. Jason de Koning at the University of Calgary for his invaluable help towards my writing and for inspiring me with several ideas towards a bioinformatics oriented future career. The support of JoAnne Gray for helping me at different stages of the graduate program is also deeply appreciated.

I thank my fiancée and the love of my life Amrita Pal for being an apostle of love, support and understanding despite being continents apart, for being the more matured partner whenever stress got the better of me and for always believing in my abilities and promises that I have made to her. I also thank my colleagues Shuhao Qiu, Ahmed AlKhuder, Xi Cheng and other co-authors for their collaborative efforts in my research projects and more importantly, for fostering strong bonds of friendship and camaraderie. Lastly, I thank my niece Soumili Gorai, born to my sister last year as an angel of good luck and hope, whose presence has always made her uncle (me) a happier person.
# Table of Contents

Abstract ............................................................................................................................................... iii

Acknowledgements ........................................................................................................................ vi

Table of Contents ........................................................................................................................... vii

List of Tables .................................................................................................................................... xi

List of Figures ..................................................................................................................................... xii

List of Abbreviation .......................................................................................................................... xiv

1 **Chapter 1. What Do 1000 Genomes Tell Us About Biased Gene Conversion Theory**

1.1 Abstract ....................................................................................................................................... 2

1.2 Introduction ................................................................................................................................ 3

1.2.1 Tsunamic wave of genomic data ............................................................................................. 3

1.2.2 Identification and characterization of haplotypes ..................................................................... 4

1.2.3 Biased Gene Conversion theory ............................................................................................... 6

1.3 Materials and Methods ............................................................................................................... 9

1.3.1 Identification of haplotypes ..................................................................................................... 10

1.3.2 Testing Biased Gene Conversion theory .................................................................................. 12

1.4 Results ......................................................................................................................................... 13

1.4.1 Haplotype classification ......................................................................................................... 13

1.4.1.1 HMOX1 haplotypes ........................................................................................................... 14
1.4.1.2 HMOX2 haplotypes .............................................. 15
1.4.1.3 AGT haplotypes .............................................. 15
1.4.1.4 ZMAT5 haplotypes .......................................... 16
1.4.2 Instances of Biased Gene Conversion .......................... 17
1.5 Conclusions ............................................................. 20
1.6 Discussions .............................................................. 20
1.7 Additional files ......................................................... 29
1.8 References ................................................................. 31

2 Chapter 2. Critical associations and non-randomness of inhomogeneous genomic compositions at middle-range distance ................................................................. 38

2.1 Abstract ................................................................. 39
2.2 Introduction ............................................................ 40
  2.2.1 Unusual DNA conformations ................................. 40
  2.2.2 Classification of MRI regions ................................. 41
  2.2.3 Properties of MRI regions ..................................... 42
  2.2.4 Genomic signatures .............................................. 45
2.3 Materials and Methods .............................................. 46
2.4 Results ................................................................. 48
2.5 Conclusions ............................................................ 59
2.6 Discussions ............................................................ 60
2.7 References ............................................................... 62

3 Chapter 3. Bioinformatics analysis of plant orthologous introns: identification of an intronic tRNA-like sequence ................................................................. 64
3.1 Abstract........................................................................................................66
3.2 Background..................................................................................................67
3.3 Results.........................................................................................................69
  3.3.1 Databases of plant orthologous introns.................................................69
  3.3.2 Conserved intronic regions in dicotyledonous plants.........................70
  3.3.3 Conserved intronic regions in flowering plants...............................71
  3.3.4 Conserved intronic regions common for moss and angiosperms......73
  3.3.5 Characterization of tRNA-like plant intronic putative ncRNA........75
  3.3.6 Intronic tRNAs of vertebrates.............................................................77
3.4 Discussions.................................................................................................78
3.5 Conclusions.................................................................................................81
3.6 Methods.......................................................................................................82
  3.6.1 Databases of plant orthologous introns.................................................82
  3.6.2 Computational analysis of RNA secondary structures.......................83
  3.6.3 BLAST analysis of conserved plant intronic regions.........................84
  3.6.4 Phylogeny of Arabidopsis tRNAs with CPIR-1 sequence....................84
  3.6.5 Mapping gene functions from Gene Ontology database...................85
  3.6.6 Computation of intronic tRNAs in humans and mouse......................86
  3.6.7 Plant growth conditions and RNA analysis.......................................86
  3.6.8 Northern blot experiments.................................................................87
3.7 Competing interests....................................................................................88
3.8 Authors’ contributions..............................................................................88
3.9 Acknowledgements....................................................................................89
3.10 Figures.................................................................................................89
3.11 References.............................................................................................96
3.12 Additional files.......................................................................................102
A  Appendix A...............................................................................................103
B  Appendix B...............................................................................................119
## List of Tables

1.1 Nucleotide conversions in formations of putative heteroduplexes ..................26
2.1 The ten longest contigs of the human genome ..........................................51
2.2 The ten longest contigs of the mouse genome ..........................................52
2.3 MRI signatures compared against genomic signatures .............................53
2.4 Comparison of MRI signatures of humans vs mouse ...............................58
A.1 HMOX1 genotypes (annotated) for 1000 genomes population .............103
A.2 HMOX2 genotypes (annotated) for 1000 genomes population .............104
A.3 AGT genotypes (annotated) for 1000 genomes population .................105
A.4 ZMAT5 genotypes (annotated) for 1000 genomes population .............106
A.15 Explanations for putative heteroduplex formations ...............................112
B.1 Comparison of tRNA features against CPIR-1 sequence .......................119
B.2 Characterization of human tRNAs inside functional genes .....................121
## List of Figures

1-1 Haplotypes illustration from four different genes under examination..............18
1-2 Heteroduplexes from HMOX1 gene................................................................................21
1-3 Heteroduplexes from HMOX2 gene..............................................................................22
1-4 Heteroduplexes from AGT gene....................................................................................23
1-5 Heteroduplexes from ZMAT5 gene..............................................................................25
2-1 Possible MRI regions.....................................................................................................43
2-2 Workflow representation for MRI calculations...............................................................50
2-3 MRI signatures after randomization of K-rich region....................................................55
2-4 MRI signatures after randomization of W-rich region....................................................56
2-5 MRI signatures after randomization of Z-rich region....................................................57
3-1 Multiple alignment of plant orthologous introns with long conserved regions…90
3-2 Multiple alignment of plant orthologous introns containing CPIR-1 sequence….92
3-3 Secondary structures of conserved region of CPIR-1 in five plant species...........93
3-4 Northern blot analysis of total RNA from Arabidopsis.................................................95
A.5 Pairwise LD for SNPs from HMOX1 gene.................................................................107
A.6 Pairwise LD for SNPs from HMOX2 gene.................................................................108
A.7 Pairwise LD for SNPs from AGT gene.................................................................109
A.8 Pairwise LD for SNPs from ZMAT5 gene.................................................................110
A.9 Complex recombinations outside the scope of heteroduplexes.........................111
B.3 Evolutionary analysis of CPIR-1 sequence........................................122
List of Abbreviations

BGC……………………………Biased Gene Conversion
BLAST…………………………Basic Local Alignment Search Tool
CIWOG…………………………Common Introns Within Orthologous Genes
CPIR…………………………Conserved Plant Intronic Region
EID…………………………Exon-Intron Database
EST…………………………Expressed Sequence Tags
GO…………………………Gene Ontology
LD…………………………Linkage Disequilibrium
MRI…………………………Mid Range Inhomogeneity
SNP…………………………Single Nucleotide Polymorphism
Chapter 1

What Do 1000 Genomes Tell Us About Biased Gene Conversion Theory?

Arnab Saha-Mandal¹, Xi Cheng², Shuhao Qiu², Jasmine Serpen³, Alexei Fedorov¹,⁴*

¹Program in Bioinformatics and Proteomics/Genomics, University of Toledo, Health Science Campus, OH 43614, USA.

²Program in Biomedical Sciences, University of Toledo, Health Science Campus, OH 43614, USA.

³Sylvania Southview High School, Sylvania, OH 43560, USA.

⁴Department of Medicine, University of Toledo, Health Science Campus, OH 43614, USA.

*To whom correspondence should be addressed. Tel +419-383-5270; Fax +419-383-3102; e-mail: Alexei.fedorov@utoledo.edu
1.1 Abstract

Using computational approaches, we mapped the SNPs from “1000 genomes” datasets obtained from international consortium onto four human genes of HMOX1, HMOX2, AGT and ZMAT5. More than half of these SNPs were found to be rare ones with frequencies less than 1%. Linkage disequilibrium was determined for small subgroups of SNPs with higher frequencies (>10%) that illustrated the presence of major haplotypes (3 in HMOX1, 3 in HMOX2, 5 in AGT and 5 in ZMAT5) which were found to be mutually exclusive, stable, and characterized by strong pairwise correlations ($R^2 > 0.8$) between any two SNPs within the same haplotype. While a vast majority of the alleles within the population (>90%) could be associated to a particular major haplotype, a small minority of them exhibited cross mixing of haplotypes due to recombination events. Special cases among 1000 Genomes genotype data were examined for possible short-range recombinations, where a span of 1-5 kb of a haplotype allele could be inserted inside another mutually exclusive haplotype allele. Such short-range exchanges of genetic materials may be explained by the formation of DNA heteroduplexes that form the crux of Biased Gene Conversion (BGC) theory. Our investigations for the aforementioned genes revealed a total of 67 cases involving possible heteroduplexes. 38 of them corresponded to conversion of G/C pairs to A/T pairs and 17 of them corresponded to conversion of A/T pairs to G/C pairs. We did not find strong evidence in support of BGC theory that heteroduplexes may cause a significant shift toward GC-richness. Our observations indicating an opposite bias towards AT-richness suggests that biased gene conversion theory should be accountable for possible exceptions and reconsiderations.
1.2 Introduction

1.2.1 Tsunamic wave of genomic data

The increasingly low cost and high throughput of next generation DNA sequencing technologies has made way for numerous large-scale sequencing projects of tremendous importance. These include the recent generation of more than 1000 human genomes as a reference population for future work [1], and the prospectively targeted sequencing of 10,000 vertebrate species [2]. The improvements in whole-genome sequencing methodologies and annotation technologies have also initiated population sequencing projects planned worldwide [3] and made thousands of patient exomes across many laboratories throughout the world available. The amount of genomic data generated in the next few years is therefore likely to usher in a new “golden age of genetics”. This will create an urgent need for efficient computational tools to translate such wealth of genomic data into useful biological and biomedical insights. Human beings are known be genetically very diverse. An average pair of human genomes differs at millions of position (Ahmed et.al, unpublished). One of the key scientific challenges of the post genomic era has been the identification of variants that matter to human health. That is, which variants impact an individual’s risk for developing diseases or modify their ability to metabolize drugs and successfully respond to medical intervention. In this regard, techniques have been sought to understand the genetic architecture of humans. Tremendous importance has therefore been placed on exploitation of the vast amount of genomic information through phased haplotypes, which identify the alleles that are co-located on the same locus of a chromosome [4]. A second meaning of the term phased
haplotype is a set of single-nucleotide polymorphisms (SNPs) on a single chromosome that are associated statistically.

1.2.2 Identification and characterization of haplotypes

DNA sequencing methods are largely based on assembling sequence reads to infer the whole genome sequence of organisms. Mammals being diploid organisms present obstacles in analyzing whole genome sequences to distinguish and infer alleles located specifically on one chromosome out of a chromosome pair. This problem is also aggravated by the fact that sequence reads are extremely shorter compared to whole genome lengths which can be up to billions of nucleotides. The understanding of genetic architecture of an organism must therefore depend on methods that could uniquely specify the genome composition pertaining to different chromosomes. Thus, methods for determining alleles co-located and statistically associated to each other (phased haplotypes) as well as those that are mutually exclusive and exist on different chromosomes (unphased haplotypes) have been highly sought for, both experimentally and computationally. As large-scale sequencing methods bring us to the “post-genomic” era, the determination of haplotype phase has garnered particular importance through several collaborative and independent efforts throughout the world. Applications of haplotype phasing include understanding relationships between genetic variations and diseases [5], imputation of missing genetic variants [6-8], using statistical techniques to resolve raw microarray data into genotypes [1, 9-11], detecting errors in individual genotypes [12], inferring lineages and patterns of human ancestry [13], inferring recombination hotspots [14], detecting fixation of mutations from recurrent mutational
patterns to infer species specific fitnesses [14], detecting signatures for forces of molecular evolution which may be attributable to neutralism, random drift or natural selection [15] and studying allele-specific differential gene expression [16].

Clark’s algorithm [17] was the first published method for haplotype phase inference using PCR-amplified samples of diploid population. The method was based on utilizing unambiguous haplotypes (from individuals homozygous at one or more loci) and parsimony (finding solutions that utilize the least number of unique haplotypes). Clark’s method was useful for identification of haplotypes having very tightly linked polymorphisms. The Expectation-Maximization (EM) [18] algorithm embarked on treating all possible haplotype configurations as equally likely and was useful in haplotype phasing from small number of SNPs that were less tightly linked. The Approximate Coalescent Models characterized new haplotypes being formed from old haplotypes through processes of mutations and recombinations [19]. These models provided a breakthrough in the use of Hidden Markov Models (HMM) for population-based statistical phasing of haplotypes. Subsequently, several methods implementing HMM for haplotype phasing came up out of which the most cited would be those of PHASE [20], MACH [8], BEAGLE [21] and IMPUTE2 [22]. The most recent development in the computational phasing of haplotypes would be the use of Identity by Descent (IBD) [14] information where two haplotypes are identical by descent if they are identical copies of a haplotype inherited from a common ancestor. IBD was successfully implemented in a large sample of Icelandic population. The pros and cons of all the aforementioned methods in comparison with one another with regard to haplotype phasing are discussed in the review paper by Browning and Browning [4]. It is also
worthwhile to note that the computational phasing accuracy depends on a number of factors. These include sample size, marker density, genotype accuracy, relatedness in the sample, ethnicity and allele frequency. The use of phase information from short sequence reads provides more information for haplotypes phasing than whole genome experimental phasing, the latter having increased cost and complexity.

1.2.3 Biased Gene Conversion Theory

Genomic DNA is characterized by profound non-randomness in its nucleotide composition and the first breakthrough achievement to illustrate such phenomenon for mammalian genomes at macromolecular level took place in 1976. Bernardi and colleagues first uncovered the compositional non-uniformity within amniote (mammals and birds) genomes using thermal melting and density gradient centrifugation [23-25]. The DNA fragments extracted by the gradient centrifugation came to be known as isochors [26], which are genomic stretches greater than 300 kb in length with high degree of uniformity in guanine and cytosine (GC) composition. In the last decade, the origin of isochors was attempted to be explained through the proposition of biased gene conversion hypothesis [27]. Biased gene conversion happens during meiotic recombination of parental alleles during sexual reproduction, typically embarking on short scale exchange of genetic material through the formation of DNA heteroduplexes [28]. BGC proposes a bias in the repair mechanism acting upon mismatches acquired inside a heteroduplex, during meiotic crossovers of homologous chromosomes. The molecular machinery of the repair mechanism (MMR) [29] prefers replacement of mismatches with alleles G/C instead of A/T giving rise to a bias in the nucleotide composition of the resulting gamete.
This phenomenon proposed to explain the GC richness of the human genome, the fixation of AT->GC mutations [30] and evolution of GC-rich isochors in mammalian genomes. It was proposed to occur in excess in regions of recombination hotspots and fast evolving regions of the human genome [31]. However, a quantitative picture of such biased conversions within human genome is lacking at present. In addition, BGC is currently a controversial hypothesis with several supporting and opposing viewpoints from scientists and research groups all over the world.

The supporting arguments in favor of GC Biased Gene Conversion (gBGC) suggest that it is responsible for increasing the GC content of recombining DNA over evolutionary time and is thought to explain the evolution of GC content in mammals and yeasts [32]. Along with molecular forces of mutational bias, random genetic drift and purifying selection of functionally deleterious alleles, gBGC is considered vital for shaping evolution [33]. Visible manifestations of gBGC in non-mammalian genomes would encompass firstly the affected patterns of codon usage and amino acid usage in yeast [34]. Secondly the results of next generation sequencing on Arabidopsis that revealed excess of gene conversion events (90%) in recombination [35] indicated GC bias. Thirdly, the maintenance of the heterogeneous GC content of the honey bee genome despite an overall A/T mutation bias was explained as an aftermath of events resulting from biased gene conversion together with recombination [36]. In the context of human genome, gBGC involving amino-acid altering polymorphisms in human populations during meiotic recombination has been proven to contribute to the spreading of deleterious mutations in human populations [37]. gBGC has also been quantitatively studied using a substitution model for DNA sequence evolution that used a series of
lineage-specific likelihood ratio tests to evaluate sequence alignments for evidence of changes in mode of selection, action of gBGC, or both [31]. It has also been stated that gBGC is widespread in human genome and is highly enriched in recombination hotspots based on data from 1000 genomes [38]. This study however considered only 179 individuals from 4 populations based on 1000 genomes low-coverage pilot project data instead of all 1092 individuals that are presently genotyped by the 1000 genomes consortium. Moreover, the study explored biased gene conversion hypothesis in the vicinity of recombination hotspots which may not be applicable at the unit of each gene, since a gene contains transcription start sites, exons, introns and untranslated 3’ ends all of which are not enriched with recombination hotspots. Indeed, several critics to the gBGC hypothesis have also emanated from research laboratories across the world. For example, gBGC may not be able to explain evolutionary phenomena such as that reported for Fetuin-A gene [39]. Within the human genome, a DNA-repair bias proposed by the biased gene conversion hypothesis could not explain the absence of GC-bias in several regions [40]. Although, a recombination initiation bias explained the absence of GC-bias, a clear picture providing a comparative strength of molecular forces of recombination initiation bias against the biased gene conversion phenomenon or explaining the preferential likelihood of a particular event in preference to other was not proposed. In a different study, by comparing the corresponding allele frequency spectra of AT-rich vs. GC-rich introns, the influence of directional selection or biased gene conversion on the mutational pattern was ruled out in several Drosophila species [41]. Additionally, a direct phylogenetic approach to examine nucleotide replacements produced by nonallelic gene conversion in Drosophila Melanogaster and primate genomes were examined that ruled
out GC-biased gene conversion and attributed previously observed GC biases due to positive selection [42]. Another inconsistency in gBGC model was pointed out with regard to the chicken genome due to a negative correlation between substitution and recombination rate [43]. These premises formed the rationale in conducting this project. Our goal was to carry out a quantitative bioinformatic investigation within human genomes to ascertain the validity of gBGC proposition i.e. contributing to overall GC richness within human genomes. We examined gBGC in the context of short-range exchanges of haplotypes where a 1-5 kb fragment of a haplotype allele could be inserted into another allele. In this respect, we use a pipeline of in-house perl scripts and manual examinations of output to determine haplotypes specifically present in four genes namely HemeOxygenase1 (HMOX1), Heme Oxygenase2 (HMOX2), Angiotensinogen (AGT) and Zinc Finger Matrin-Type 5 (ZMAT5) based on 1000 genomes datasets. These genes have been historically studied at our lab [44-47]. Subsequently, we identified, examined and characterized several putative DNA heteroduplexes within the population alleles. HMOX1 and HMOX2 are functional isoforms that are essential enzymes in heme catabolism cleaving heme to form biliverdin and are responsible for maintaining iron homeostasis in our body. AGT encodes for angiotensinogen precursor, is expressed in the liver and is cleaved by the enzyme renin in response to lowered blood pressure. Mutations in this gene are associated with susceptibility to essential hypertension and can cause disorders in renal tubular development. In addition, another gene ZMAT5 was randomly chosen for analysis.

1.3 Materials and Methods
1.3.1 Identification of haplotypes

Genotype datasets for 1092 human genomes were downloaded from the 1000 genomes ftp site (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521/) as Variant Call Format (VCF) files version 4.1 that contained a total of 38.2M SNPs, 3.9M short indels and 14K deletions for all the human chromosomes. Distribution of mutations were investigated in four genes: 1) Heme oxygenease-1 (HMOX1), 2) Heme oxygenase-2 (HMOX2), 3) Angiotensinogen (AGT) and 4) Zinc finger matrin-type 5 (ZMAT5). Using the GenBank [48] coordinates as limits for HMOX1 (Chr22: 35777060..35790207), HMOX2 (Chr16: 4524719..4560348), AGT (Chr1: 230838269..230850336, complement) and ZMAT5 (Chr22: 30126945..30162969, complement) respectively, the VCF files for corresponding chromosomes were processed using perl script Geneexcision.pl, whereby only the genetic variants inclusive of, but not crossing the above limits were retained. Within each of the resultant files corresponding to a particular gene, subgroups of SNPs having a frequency of occurrence > 10% (based on allele frequency annotation in 1000 genomes datasets) were considered. Corresponding to those SNPs, genotype annotations for each population sample with respect to homozygous variants (1|1), heterozygous (0|1 or 1|0) or homozygous presence of reference alleles (0|0) was further processed using perl script Popvariance.pl, that exemplified the annotations into values of 2 (homozygous allele), 1 (heterozygous allele) and 0 (reference allele) arranged in microsoft excel tables (A1 – A4). The excel tables were analyzed for pairwise correlation coefficients of linkage disequilibrium using the formula from Sved J.A. [49]

\[ \tau_{AB}^2 = \frac{D_{AB}^2}{P_A (1-P_A)P_B (1-P_B)} \]  

\text{(1)}
where $r_{AB}^2$ implies the squared correlation coefficient for two alleles A and B in different genomic loci and $P_A$ or $P_B$ denotes the probability of occurrence of alleles A or B respectively. $D_{AB}$ is the linkage disequilibrium coefficient calculated using the formula from *Lewontin and Kojima* [50]

$$D_{AB} = P_{AB} - P_A P_B$$

...(2)

where $P_{AB}$ denotes the joint probability estimate of two different alleles A and B existing together in a population sample. For any two given alleles A and B, only the mutated states were considered as cases (increment of two for homozygous mutated, one for heterozygous mutated and none for reference allele) for computation of probability measures in the aforementioned formulas 1 and 2. In-house perl script *LinkageRsquare.pl* performed the calculations for all possible pairwise correlation coefficients which are displayed for each of the four genes in Appendix A (A5-A8). Groups of mutations that were found to exhibit strong correlation coefficients of linkage disequilibrium $r_{AB}^2 > 0.8$ were categorized as haplotypes, colored differently for further analysis. Further processing of tables A1-A4 using perl script *homozygous.pl* was done to calculate the number of people and the associated homozygosity, heterozygosity or reference allele attributable to a cluster of SNPs specified by the user according to genomic positions/SNPs ID. The results are presented in tables A9-A12. Visual examinations of the tables revealed exceptions to the categorization of haplogroups by the correlation coefficients of linkage disequilibrium alone and are described further in the results section.
1.3.2 Testing Biased Gene Conversion theory

Tables (A1-A4) generated from *Popvariance.pl* were examined for co-existence and frequency of haplotypes. While a vast majority of the alleles within the population (>90%) could be associated to a particular major haplogroup, a small minority of them exhibited cross mixing of haplotypes due to recombination events. Special cases were examined for possible short-range exchanges, where a span of 1-5 kb of a haplotype allele could be inserted inside another mutually exclusive haplotype allele. Such events may be explained by double recombination events or formation of heteroduplexes. Preliminary investigations for such possible cases revealed 83 polymorphisms (inclusive of SNPs, insertions and deletions) that exhibited transfer from one haplotype into another and yet there existed several cases of “complex” recombinations that are out of the scope of the paper since they are unlikely to be formed by heteroduplex mechanism owing to their size (> 5kb) (Figure A13). The statistical likelihood that encapsulates the relative evidence for each possible genotype call is expressed as genotype likelihood and is annotated by the 1000 genomes consortium. Genotype likelihoods are expressed as logarithms to the base 10, corresponding to each polymorphism for every individual. When a threshold of -0.02 (or greater than $10^{-0.02} = 95.4\%$ probability) was applied on the aforementioned 83 cases involving heteroduplexes and only single nucleotide polymorphisms were considered, we obtained a total of 67 cases which we considered statistically reliable and thus processed further. Thirty-eight of them corresponded to conversion of G/C pairs to A/T pairs and 17 of them corresponded to conversion of A/T pairs to G/C pairs. We used a Monte Carlo simulation available from [http://bpg.utoledo.edu/~afedorov/lab/prog/montecarlo.html](http://bpg.utoledo.edu/~afedorov/lab/prog/montecarlo.html) to determine the statistical
significance associated with the outcome of our investigations of biased gene conversion hypothesis.

1.4 Results

1.4.1 Haplotype classification

The distribution of variants within the four genes under investigation illustrated the numbers of genetic variants as 212 for HMOX1, 489 for HMOX2, 215 for AGT and 451 for ZMAT5 respectively. More than 50% of them showed a frequency of occurrence less than 1% in the population as were evident from the allele frequency annotation in 1000 genomes database as well as the absence of polymorphism (0|0 annotation) for a majority of population samples. Smaller subgroups of genetic variants within each gene that had a frequency of occurrence > 10% were isolated that yielded 32 mutations for HMOX1, 64 for HMOX2, 25 for AGT and 83 for ZMAT5. The genetic variants were analyzed for existence of possible haplotypes using mathematical equations of linkage disequilibrium embedded through our Perl programs. Even though a correlation cut-off > 0.8 resulted in the identification of several haplotypes, visual examinations of allelic compositions of population samples were necessary for determining phased haplotypes since the formula could not explain within which one out of the two alleles a polymorphism would exist, if it exhibited heterozygosity. The execution of our perl script Popvariance.pl yielded Tables A1-A4 which were subsequently manually inferred for possible haplotype associations not evident from correlation tables (A5-A8). 3-5 major haplotypes were discovered in each gene that further encompassed several sub-haplotypes. We have
reported one or two sub-haplotypes from genes HMOX1, AGT and ZMAT5. Their in-depth analysis is out of the focus of the project and hence left out. The major haplotypes characterized more than 90% of the population alleles and were also found to be mutually exclusive, the pairwise correlations between any two mutations belonging to different haplotype groups $r_{AB}^2$ being $< 0.2$. Our results were also confirmed from a public software SNAP [51] that used the snpIDs as input and calculated pairwise correlation coefficients of linkage disequilibrium between all possible pairs of them. From the tables A1-A4, 10% of the remaining alleles could be associated to events involving recombinations due to cross-mixing of haplotype alleles. With respect to our correlation cut-off $r_{AB}^2 > 0.8$ for identification of haplotypes, we found several exceptions to our threshold.

1.4.1.1 HMOX1 haplotypes

The haplotypes for HMOX1 are presented in Figure 1A. There are a total of three major haplotypes colored red, blue and green. Within the HMOX1 gene, the SNP rs55798039(CT/c) was found to have a pairwise correlation coefficient $\approx 0.71$ to $0.79$ with other SNPs belonging to blue haplotype and was thus categorized within the same haplotype as the second blue SNP. Also, three SNPs rs79074644(T/c), snpID n/a:pos35785282 (CTTCT/c), rs5750104(T/a) colored black in Figure 1A were found to weakly correlate with each other ($r_{AB}^2 \approx 0.4$), yet demonstrating a high frequency of occurrence (32%-42%) within the population samples and avoidance with the red haplotype. The three black SNPs were therefore labelled as a minor haplotype that coexisted with some (blue or green) and avoided other haplotype (red).
1.4.1.2 HMOX2 haplotypes

The haplotypes for HMOX2 are presented in Figure 1B. There are a total of three major haplotypes colored red, blue and green. Within the HMOX2 gene, the SNP rs11643057(T/c) was found to exhibit a correlation coefficient $r_{AB}^2 \approx 0.72$ to 0.83 with other SNPs belonging to the green haplotype and hence was labeled as the 16th green SNP. The first 31 SNPs of the blue haplotype have strong pairwise correlation coefficients with each other ($r_{AB}^2 > 0.8$). Two mutations, namely rs7192051 (G/t) and another with uncharacterized identifier but located in position 4547786 (AG/a) within the HMOX2 gene exhibited correlation coefficients $r_{AB}^2 \approx 0.70$ to 0.83 with other SNPs belonging to the blue haplotype. Thus, they are labeled as 32nd and 33rd SNPs respectively of the blue haplotype.

1.4.1.3 AGT haplotypes

The haplotypes for AGT are presented in Figure 1C. Within the AGT gene, the SNP rs3789670(T/c) was found to possess a pairwise correlation coefficient $r_{AB}^2 \approx 0.69$ to 0.77 with other orange SNPs and hence was labelled as the first orange SNP. The fourth green SNP rs1078499 (G/a) had a weak correlation with other green SNPs ($r_{AB}^2 \approx 0.6$) since this SNP can exist as a separate haplotype allele in addition to existing with the other green SNPs. Infact, visual examinations revealed that the fourth green SNP existed as a heterozygous allele in several populations and on one occasion, as a homozygous allele (within individual NA19750 belonging to Mexican individual ethnicity from Los Angeles, California). It was also seen that haplotypes like orange, yellow and green associated to the blue haplotype. Particularly, the orange and yellow haplotypes never
exist independently but do so in association with the blue haplotype allele. Green exists as a separate haplotype exhibiting homozygosity within 20 individuals (Table A11) and also in association with blue haplotype exhibiting homozygosity within 27 individuals (Table A11). Since, the allele frequencies for blue are different from orange or yellow, the pairwise correlation coefficients of any yellow with blue or any orange with blue is very poor (≈ 0.16 for yellow with blue and ≈ 0.18 for orange with blue) but yellow or orange always exists with blue. There are 37 individuals homozygous with yellow and blue alleles together and 24 individuals homozygous with orange and blue alleles together (Table A11). In contrast, there are no individuals having exclusively yellow or orange with the absence of blue. This can also be attributed to the higher allele frequency of blue haplotype which might have appeared earlier in evolutionary history and accumulated yellow, orange and green SNPs within different population groups at different periods of time. The resulting blue haplotype that has been under evolutionary pressure ceases to exist on its own and associates with either yellow or green or orange. Haplotypes sharing a common ancestor are known as haplogroups. Based on this rationale, combinations of alleles representing blue + yellow, blue + green, blue + orange and blue + yellow haplotypes could be characterized as different haplogroups. Figure 1C represents haplotypes and haplogroups for AGT.

1.4.1.4 ZMAT5 haplotypes

The haplotypes for ZMAT5 are presented in Figure 1D. Within the ZMAT5 gene, the haplotype groups green and purple have strong coefficients of correlation ≈ 0.6 to 0.8 and are yet categorized as separate haplotypes. The number of individuals homozygous with
green are 226 (Table A12) and with purple are 230 (Table A12). The common amongst the population samples tantamount to 185 (Table A12) which not only explains a strong correlation between green and purple, but also illustrates the rationale of having green and purple as two separate haplotypes. There are 41 homozygous individuals having green haplotype without purple and 45 homozygous individuals having purple without green. Indeed, for this reason, green only, purple only and combination of green + purple alleles could be classified as a haplogroup. The SNP rs5997518(G/a) has a pairwise correlation coefficient ≈ 0.6 with other orange SNPs and is thus categorized as a member of the orange haplotype as the fifth orange SNP. Being close to the 5’ end of the gene, it is subject to recombination with other haplotypes that decreases its measure of association with its original haplotype cluster, the orange group. There are 16 individuals homozygous with green and orange and 3 individuals homozygous with only orange haplotype due to which they are classified as minor haplotypes in Figure 1D.

1.4.2 Instances of Biased Gene Conversion

The biased gene conversion hypothesis was tested by manual examinations of tables A1-A4. The interpretation of haplotypes, their relative associations and exclusivity were utilized in examining cases of short-range exchange of haplotypes. We obtained 83 mutations (SNPs and indels) involved in 75 individuals that exhibited possible heteroduplexes indicating biased gene conversion as the causal factor. With respect to a polymorphic site that is characterized by a particular reference allele (eg. A) and another mutated allele (eg. a), the biallelic composition of alleles at that position could be attributable to three different combinations (namely AA, Aa and aa for our particular case
Figure 1-1. Exon intron structure of the human genes A) HemeOxygenase 1 (HMOX1), B) HemeOxygenase 2 (HMOX2), C) Angiotensinogen (AGT) and D) Zinc finger matrin type-5 (ZMAT5) with mapped most frequent SNPs (stars) characterized in 1000 genomes project that revealed the existence of major and minor haplotype groups. The major haplotypes exhibit homozygosity within several individuals while minor haplotypes remain as either heterozygous alleles or are present in homozygous combinations in very few individuals. The SNPs within each of the above genes exhibiting strong linkage disequilibrium with each other are marked with the same color. However, same colors for SNPs in different genes should not be interpreted as the same haplotype since each gene is an independent study. Also, some haplotypes may prefer co-existence (eg. Green and Blue in figure C) or avoidance with other haplotypes (eg. Black avoids Red in figure...
A). The SNP identifiers annotated by the 1000 genomes consortium for all the SNPs mapped above are listed in Table A14.

). The probabilities of such possible genotypes are annotated in the 1000 genomes database using a likelihood measure expressed as logarithmic values to the base ten. This would imply that a likelihood value of -0.00 should correspond to a $10^{0.00}$ or > 99% probability and a likelihood value of -0.03 should correspond to a $10^{-0.03}$ or > 93% probability. When a threshold of -0.02 or > 95% probability was applied on the aforementioned 83 cases involving heteroduplexes and only single nucleotide polymorphisms were considered, we obtained a total of 67 cases within 52 individuals. All those 67 cases for all the four genes are presented in Figures 2-5, with supplementary explanations in Table A.15 and Table 1 enlisting the nucleotide conversions taking place. 38 of those cases corresponded to conversion of G/C pairs to A/T pairs and 17 of them corresponded to conversion of A/T pairs to G/C pairs. The remaining 12 cases represented conversions of either A to T, G to C, T to A or C to G that are out of the context of examination of biased gene conversion hypothesis. We used our own monte carlo simulation program available from http://bpg.utoledo.edu/~afedorov/lab/prog/montecarlo.html to conduct a statistical evaluation of our results. In our monte carlo program, the null hypothesis was based on equal probabilities (0.5) of allele conversions from G/C to A/T as well as from A/T to G/C. We had a total of 55 cases involving the aforementioned allele conversions using the rationale of which, we performed 1 million simulation experiments each with 55 coin tosses. A fair coin would be most likely to produce 27-28 heads and same number of tails
upon 55 coin tosses. From our observations, we had a total of 38 cases involving G/C -> A/T and 17 cases involving A/T -> G/C. We thus simulated 38 (or more) and the remaining as possible numbers of heads and tails respectively in our simulator and concluded a prevalence of conversions of G/C alleles to A/T alleles. It testified a prevalence of GC-> AT transitions over AT ->GC transitions with a p-value of 0.016. Our data reject the hypothesis that there is a significant bias of AT -> GC conversions over GC -> AT ones in heteroduplexes.

1.5 Conclusions

In the light of our findings, we conjecture that gBGC cannot explain the GC richness of human genomes. The formation of GC rich isochors in mammals is thus, still an open scientific question.

1.6 Discussions

In addition to pure algorithmic characterization of haplotypes, manual curation of datasets improves the overall picture. Since the annotated genotype information is subject to errors in sequencing and low confidence calls for SNPs, (evident from low genotype likelihoods in 83 cases involving heteroduplexes, thereby reducing it to 67) the information in the database could be corrected over time, along with increasing the volume of datasets by sequencing additional individuals (as targeted up-to 2500). From the analysis of four human genes, we also found that gene conversion may be biased towards AT richness. We therefore propose that the GC biased gene conversion theory
Figure 1-2. Population samples from 1000 genomes corresponding to HMOX1 gene (35777060..35790207) in chr22 reveal instances of heteroduplex formations due to short-scale recombination events, wherein mutations jump from one haplotype to another. The corresponding regions have been shown by black bars. The mutations diagrammatically represented in between the two alleles are equally probable in either of them. Only point mutations have been analyzed irrespective of some indels that also occupy positions within the gene. Some empty stars are indicative of absence of mutation expected at a particular position and the values in brackets denote the likelihood values of the SNPs expressed as logarithms to the base 10. Even though several SNPs underneath a black bar represent a local recombination event, only those having a likelihood > -0.02 have been labeled and analyzed further. Explanations for each of the above case are provided in Table A.15.
Figure 1-3. Population samples from 1000 genomes corresponding to HMOX2 gene (4524719..4560348) in chr16 reveal instances of heteroduplex formations due to short-scale recombination events, wherein mutations jump from one haplotype to another. The corresponding regions have been shown by black bars. The mutations diagrammatically represented in between the two alleles are equally probable in either of them. Only point mutations have been analyzed irrespective of some indels that also occupy positions within the gene. Some empty stars are indicative of absence of mutation expected at a particular position and the values in brackets denote the likelihood values of the SNPs expressed as logarithms to the base 10. Even though several SNPs underneath a black bar represent a local recombination event, only those having a likelihood > -0.02 have been labeled and analyzed further. Explanations for each of the above case are provided in Table A.15.
Figure 1-4. Population samples from 1000 genomes corresponding to AGT gene (230838269..230850336) in Chr1 reveal instances of heteroduplex formations due to short-scale recombination events, wherein mutations jump from one haplotype group to
another. The corresponding regions have been shown by black bars. The mutations
diagrammatically represented in between the two alleles are equally probable in either of
them. Only point mutations have been analyzed irrespective of some indels that also
occupy positions within the gene. Some empty stars are indicative of absence of mutation
expected at a particular position and the values in brackets denote the likelihood values
of the SNPs expressed as logarithms to the base 10. Even though several SNPs
underneath a black bar represent a local recombination event, only those having a
likelihood > -0.02 have been labeled and analyzed further. Explanations for each of the
above case are provided in Table A.15.
Population samples from 1000 genomes corresponding to ZMAT5 gene (30126945..30162969) in Chr22 reveal instances of heteroduplex formations due to short-scale recombination events, wherein mutations jump from one haplotype group to another. The corresponding regions have been shown by black bars. The mutations diagrammatically represented in between the two alleles are equally probable in either of them. Only point mutations have been analyzed irrespective of some indels that also occupy positions within the gene. Some empty stars are indicative of absence of mutation expected at a particular position and the values in brackets denote the likelihood values of the SNPs expressed as logarithms to the base 10. Even though several SNPs underneath a black bar represent a local recombination event, only those having a likelihood $> 0.02$ have been labeled and analyzed further. Explanations for each of the above case are provided in Table A15.
## HMOX1

<table>
<thead>
<tr>
<th>Population Sample</th>
<th>SNP ID</th>
<th>Position</th>
<th>Nucleotide Change</th>
<th>Local GC %</th>
</tr>
</thead>
<tbody>
<tr>
<td>HG00593</td>
<td>rs5755717</td>
<td>35786713</td>
<td>C→A</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>rs5755718</td>
<td>35786722</td>
<td>C→T</td>
<td>60</td>
</tr>
<tr>
<td>HG01494</td>
<td>rs8139532</td>
<td>35779568</td>
<td>A→G</td>
<td>30</td>
</tr>
<tr>
<td>NA11930</td>
<td>rs5995098</td>
<td>35787167</td>
<td>C→G</td>
<td>40</td>
</tr>
<tr>
<td>NA11994</td>
<td>rs11913209</td>
<td>35778727</td>
<td>T→A</td>
<td>50</td>
</tr>
<tr>
<td>NA19355</td>
<td>rs79074644</td>
<td>35785208</td>
<td>C→T</td>
<td>20</td>
</tr>
<tr>
<td>NA12761</td>
<td>rs2269534</td>
<td>35786226</td>
<td>C→T</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>rs5755717</td>
<td>35786713</td>
<td>C→A</td>
<td>60</td>
</tr>
</tbody>
</table>

## HMOX2

<table>
<thead>
<tr>
<th>Population Sample</th>
<th>SNP ID</th>
<th>Position</th>
<th>Nucleotide Change</th>
<th>Local GC %</th>
</tr>
</thead>
<tbody>
<tr>
<td>HG00129</td>
<td>rs11643057</td>
<td>4544505</td>
<td>C→T</td>
<td>30</td>
</tr>
<tr>
<td>HG00130</td>
<td>rs7194923</td>
<td>4530919</td>
<td>T→C</td>
<td>40</td>
</tr>
<tr>
<td>HG00134</td>
<td>rs11643057</td>
<td>4544505</td>
<td>C→T</td>
<td>30</td>
</tr>
<tr>
<td>HG00355</td>
<td>rs8046295</td>
<td>4528634</td>
<td>C→T</td>
<td>70</td>
</tr>
<tr>
<td>HG01149</td>
<td>rs7192051</td>
<td>4542117</td>
<td>T→G</td>
<td>50</td>
</tr>
<tr>
<td>HG01259</td>
<td>rs7200280</td>
<td>4530862</td>
<td>C→T</td>
<td>60</td>
</tr>
<tr>
<td>NA18505</td>
<td>rs9929475</td>
<td>4540885</td>
<td>A→G</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>rs55939215</td>
<td>4541390</td>
<td>C→G</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>rs12443748</td>
<td>4541831</td>
<td>C→T</td>
<td>70</td>
</tr>
<tr>
<td>NA18516</td>
<td>rs7200280</td>
<td>4530862</td>
<td>C→T</td>
<td>60</td>
</tr>
<tr>
<td>NA18952</td>
<td>rs7195361</td>
<td>4530806</td>
<td>T→A</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>rs7194923</td>
<td>4530919</td>
<td>T→C</td>
<td>40</td>
</tr>
<tr>
<td>NA19096</td>
<td>rs4785968</td>
<td>4535307</td>
<td>T→A</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>rs8063084</td>
<td>4536555</td>
<td>C→T</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>rs8051695</td>
<td>4536621</td>
<td>T→C</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>rs8053770</td>
<td>4539030</td>
<td>T→A</td>
<td>50</td>
</tr>
<tr>
<td>NA19108</td>
<td>rs8046295</td>
<td>4528634</td>
<td>C→T</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>rs11863671</td>
<td>4528636</td>
<td>C→T</td>
<td>70</td>
</tr>
<tr>
<td>NA19672</td>
<td>rs8046295</td>
<td>4528634</td>
<td>C→T</td>
<td>70</td>
</tr>
<tr>
<td>NA19657</td>
<td>rs4786502</td>
<td>4537104</td>
<td>T→C</td>
<td>0</td>
</tr>
<tr>
<td>NA20336</td>
<td>rs78386763</td>
<td>4525477</td>
<td>A→T</td>
<td>10</td>
</tr>
</tbody>
</table>
### AGT

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample</th>
<th>SNP ID</th>
<th>Position</th>
<th>Nucleotide Change</th>
<th>Local GC %</th>
</tr>
</thead>
<tbody>
<tr>
<td>HG00406</td>
<td>rs2493134</td>
<td>230849359</td>
<td>C→T</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>HG00418</td>
<td>rs699</td>
<td>230845794</td>
<td>G→A</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>NA18592</td>
<td>rs11568056</td>
<td>230842497</td>
<td>C→T</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>NA12006</td>
<td>rs2148582</td>
<td>230849799</td>
<td>G→A</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>NA12717</td>
<td>rs11122576</td>
<td>230846679</td>
<td>T→C</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rs11568028</td>
<td>230847244</td>
<td>C→T</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rs11122578</td>
<td>230847789</td>
<td>G→A</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>NA18489</td>
<td>rs3789678</td>
<td>230849482</td>
<td>T→C</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rs5049</td>
<td>230850083</td>
<td>C→T</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>NA18873</td>
<td>rs2478539</td>
<td>230844772</td>
<td>T→G</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>NA18949</td>
<td>rs3789670</td>
<td>230843714</td>
<td>C→T</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>NA18990</td>
<td>rs3827749</td>
<td>230841559</td>
<td>G→A</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>NA19028</td>
<td>rs2478543</td>
<td>230844310</td>
<td>C→T</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>NA19074</td>
<td>rs3827749</td>
<td>230841559</td>
<td>G→A</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rs3789666</td>
<td>230843438</td>
<td>G→C</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rs3789667</td>
<td>230843446</td>
<td>G→A</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>NA19445</td>
<td>rs699</td>
<td>230845794</td>
<td>G→A</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>NA20520</td>
<td>rs2148582</td>
<td>230849799</td>
<td>G→A</td>
<td>60</td>
<td></td>
</tr>
</tbody>
</table>

### ZMAT5

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample</th>
<th>SNP ID</th>
<th>Position</th>
<th>Nucleotide Change</th>
<th>Local GC %</th>
</tr>
</thead>
<tbody>
<tr>
<td>HG00240</td>
<td>rs140114</td>
<td>30133881</td>
<td>T→C</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>HG00350</td>
<td>rs140120</td>
<td>30138548</td>
<td>T→C</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>HG01066</td>
<td>rs131267</td>
<td>30147861</td>
<td>G→A</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>HG00251</td>
<td>rs140106</td>
<td>30130313</td>
<td>G→A</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>HG01250</td>
<td>rs140122</td>
<td>30138908</td>
<td>G→C</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>HG01360</td>
<td>rs140104</td>
<td>30129156</td>
<td>T→C</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>NA12874</td>
<td>rs140120</td>
<td>30138548</td>
<td>T→C</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>NA18510</td>
<td>rs131270</td>
<td>30149896</td>
<td>T→C</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>NA18573</td>
<td>rs140108</td>
<td>30130776</td>
<td>C→T</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>NA19005</td>
<td>rs131260</td>
<td>30132964</td>
<td>A→C</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>NA19020</td>
<td>rs131264</td>
<td>30134763</td>
<td>C→T</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>NA19036</td>
<td>rs131282</td>
<td>30156608</td>
<td>C→A</td>
<td>30</td>
<td></td>
</tr>
</tbody>
</table>
Table 1.1 SNPs involved in short-scale recombination events due to heteroduplexes represented by their IDs, chromosome positions, direction of mutation, local GC% and the corresponding individual, analyzed for each of the four genes of HMOX1, HMOX2, AGT and ZMAT5. There are a total of 67 cases in which 38 represent conversion of G/C pairs to A/T pairs and 17 represent conversion of A/T pairs to G/C pairs. 12 of them represent conversions from A to T, G to C or vice-versa that are not relevant in the context of biased gene conversion theory.

Polymorphisms of HMOX1 and AGT genes have been widely reported with regard to disease associations and other pathophysiological conditions. rs2071749 (a member of the red haplotype of HMOX1, Figure 1A) was shown not to have significant association with essential hypertension (EH) and blood pressure [52]. rs9607267 (a member of the blue haplotype of HMOX1) was shown to exhibit significant association with EH [53]. The M235T polymorphism or rs699 (a member of the red haplotype of AGT gene, Figure 1B) is highly reported, with regard to its association with boosting up athletic
performance [54], myocardial infarction [55] and blood pressure lowering response to angiotensin converting enzyme inhibitor [56]. In all these and several similar studies, the genotyping studies were carried out through restriction fragment length polymorphism and allele specific polymerase chain reactions. Our computational results show statistical associations of mutations thereby indicating several other genetic variants as biomarkers that could also be phenotypically associated with the aforementioned disease outcomes.

Our methods of characterizing haplotypes are based on combinations of programming and manual approaches, which present difficulties in characterizing them as holistic methods for whole genome analyses. In addition, the task for finding cases of putative heteroduplexes within 1000 genomes genotype data was also done manually. The cases for a single gene took more than a month’s time to be vividly delineated that served as one of the bottlenecks of our approaches to be extrapolated on a larger scale. The span and intersections of haplotypes also pose challenges and limitations to the identification and characterization of putative heteroduplexes. For example, if a gene is characterized by two frequently occurring haplotypes that are localized towards 5’ and 3’ ends respectively, the possibilities for detection of short-scale recombinations of those two haplotype alleles could be completely ruled out. In such cases, the formation of heteroduplexes, even though possible, would not be manifested and visibly inferred from the genotype data. In fact, our study of human TMEM23 gene present in chromosome 10 revealed numerous haplotypes that were confined to specific regions within the gene. Thus, the genotype data also poses limitations to the extrapolation of our approach.

1.7 Additional files
Additional file 1 Supplementary table A1 (provided in A.1)
Additional file 2 Supplementary table A2 (provided in A.2)
Additional file 3 Supplementary table A3 (provided in A.3)
Additional file 4 Supplementary table A4 (provided in A.4)
Additional file 5 Supplementary figure A5 (provided in A.5)
Additional file 6 Supplementary figure A6 (provided in A.6)
Additional file 7 Supplementary figure A7 (provided in A.7)
Additional file 8 Supplementary figure A8 (provided in A.8)
Additional file 9 Supplementary figure A9 (will be available online)
Additional file 10 Supplementary table A10 (will be available online)
Additional file 11 Supplementary table A11 (will be available online)
Additional file 12 Supplementary table A12 (will be available online)
Additional file 13 Supplementary table A13 (provided in A.13)
Additional file 14 Supplementary table A14 (will be available online)
Additional file 15 Supplementary texts A15 (provided in A.15)
1.8 References


Chapter 2.

Critical associations and non-randomness of inhomogeneous genomic compositions at middle-range distance

Arnab Saha-Mandal\textsuperscript{1}, Akkuratov Evgeny\textsuperscript{2}, Shuhao Qiu\textsuperscript{3}, Samer Gozem\textsuperscript{4}, Alexei Fedorov\textsuperscript{1,5*}

\textsuperscript{1}Program in Bioinformatics and Proteomics/Genomics, University of Toledo, Health Science Campus, OH 43614, USA.

\textsuperscript{2}International Biotechnology Center M.V. Lomonosov Moscow State University, Moscow 119991, Russia.

\textsuperscript{3}Program in Biomedical Sciences, University of Toledo, Health Science Campus, OH, USA.

\textsuperscript{4}Department of Chemistry, Bowling Green State University, Bowling Green, OH 43403, USA.

\textsuperscript{5}Department of Medicine, University of Toledo, Health Science Campus, OH, USA.
2.1 Abstract

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%) (ref). Such large proportions of non-protein coding DNA were previously characterized as “junk” DNA. However the erstwhile “junk” DNA has been found to be the source of several thousands of non-protein coding RNAs, regulatory regions controlling gene expression, chromatin folding structures which include nucleosome positioning sites and scaffold/matrix attached regions (refs) and several other signals responsible for genome organization. We hypothesize that many of the unknown signals function in interdependent groups, and can only be understood properly as the result of a more integrated analysis. The nucleotide sequence composition of non-coding DNA, including intergenic regions and introns, is far from random. The non-randomness or inhomogeneity of base composition exist in varying layers of complexity and sequence length, illustrated by frequency biases ranging from dinucleotide scale all the way up to patterns spread over millions of bases. In this regard, we investigated the presence of specific nucleotide composition patterns (from 30 up to 10,000 nucleotides) called Mid-Range Inhomogeneity (or MRI) regions, which are associated with non-B-form DNA conformations. Here, for the first time, we defined the term MRI signatures that quantified the association of MRI regions with one another. The MRI signatures demonstrated non-random intricate patterns of nucleotide bases within mammalian genomes. Specific MRI regions exhibited strong positive or negative tendencies to cluster with one another. We also showed that the MRI signatures tended to
be unique to a particular species. We hypothesize that these MRI association patterns reflect signals essential for genome function and global coordination of gene expression.

2.2 Introduction

2.2.1 Unusual DNA conformations

DNA can exist in numerous conformations that are critical for genome function. A Genome is a self-operating dynamic system that has an important property to exist in multiple conformations. Transitions between various DNA structures occur in response to environment signals and also in accordance with imbedded programs for organism development. We classify DNA conformations into three groups based on their size. These include short-range DNA structures are formed by a particular nucleotide and its immediate neighboring bases. For instance, adenine and guanine bases in the B-form DNA could be rotated along the glycosidic bond from Watson-Crick A-T or G-C pairing into Hoogstein pairing as described by Nikolova et al. 2011. Highlights of this paper in the same issue of Nature by Honig and Rohs concluded that “each letter [nucleotide] potentially has two meanings that determine both hydrogen-bonding patterns and structural variations in the double helix”. The second group is represented by mid-range structures that usually occupy dozens or hundreds of nucleotides. Ten of the so-called non-B-form or unusual DNA structures existing under physiological conditions were recently reviewed, and among them the most known are Z-DNA, H-DNA, quadruplexes, and GC-rich regions. They are associated with particular nucleotide compositions such as: alternating purine/pyrimidine bases (Z-DNA), homo-purine and pyrimidine stretches
(H-DNA), and particular short repetitive sequences enriched by guanines (quartets). Finally, the third group comprises long-range DNA structures that are spread over thousands of nucleotides. Their conformations are determined by interaction with specific proteins and RNAs. The best-known example is euchromatin and heterochromatin that represent loose and compact forms of DNA organization respectively. Recent advances in genomic technologies have revealed complex spatial organization of genomes and have also given way to coining of new terms like topological domains and chromosomal territories corresponding to long-range genomic scales. Our focus in this paper is mainly on Mid-Range DNA conformations. These DNA regions have very specific nucleotide compositions, which we call Mid-Range Inhomogeneity (or MRI) regions.

### 2.2.2 Classification of MRI regions

Genomic patterns on short-range scales represent various “words” composed from nucleotide “letters”. Each of these words occurs many times within DNA sequences. The longest words, also known as “pyknons”, comprise sequences up to 17 nucleotides long that are overabundant in the exons and introns of humans and other mammals. Many of these pyknons have also been classified as minisatellites that are short repeating sequences present in thousands of numbers within the human genome. The vast majority of sequences, only a little bit longer than pyknons, are unique even for the large genomes of animals and plants. For example, the complete theoretical set of 20-nt-long sequences is comprised of $4^{20}$ different words of length 20, which is just over one trillion. More than 99% of these 20-mer oligonucleotides never occur in the entire human genome ($\sim 3 \times 10^9$
bp). Therefore, biologists frequently use 20-mer oligonucleotides as PCR primers or hybridization probes for experimental characterization of particular genomic segments. The genomic arrangement of short sequences (<20 bp) is covered in insightful papers. Here we consider genomic patterns longer than 30 and up to several thousands of nucleotides to be called the mid-range scale. 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+C richness, purine richness, etc. We characterize different types of MRI regions shown in Figure 2-1. We also distinguish mid-range genomic scales from the long-range scale represented by genomic isochores, reviewed elsewhere. Traditionally, G+C-rich and G+C-poor isochores are considered to be from 100 kb and longer. Recently, scientists have started to describe ultra-short isochores in the range of tens of thousands of nucleotides. In order not to interfere with isochores, we limit the length of mid-range patterns to 10,000 bases. The main focus of this paper 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. Some of these patterns are scarcely investigated and still wait for thorough exploration and recognition.

### 2.2.3 Properties of MRI regions

Non-coding DNA has a profound non-randomness in its sequence composition at the mid-range scale (from 30 up to 10,000 nucleotides). This non-randomness is manifested
Figure 2-1. Possible seven types of MRI regions that represent overall nucleotide composition characterized by an abundance of nucleotides of two different types.

in base compositional extremes detected for each of four nucleotides (A, G, T, or C) or any of their combinations. Examples of such compositional non-randomness are A-rich, purine-rich, or G+T-rich regions. Almost every combination of nucleotides has such enriched regions. We refer to these regions as being “inhomogeneous.” These regions are associated with unusual DNA conformations and/or particular DNA properties (e.g. bends, turns, electrostatic potential). Notably, mid-range inhomogeneous regions have complex arrangements relative to each other and to specific genomic sites such as centromeres, telomeres, and promoters, pointing to their important role in genomic functioning and organization. Prakash et. al. demonstrated that MRI regions have non-random distributions along genes. For instance, Z-DNA has a strong preference to be enriched at 3'-ends. The same trend was also observed for GT- and AC-rich regions. Predominantly, scientists concentrate on a particular type of non-B structures, while genomic arrangement of different types of DNA conformations is practically unknown. Often non-B-form DNA regions or their considerable parts are represented by simple repeats. In first approximation, half of non-canonical DNA structures are comprised by
simple repeats while the other half are unique sequences with particular nucleotide compositions specific to a particular conformation (e.g. alternation of purines and pyrimidines). Simple repeats occupy 2.2% of the human genome and 3% of mouse and rat genomes. They may experience rapid expansion/contraction during evolution. Recently, Prakash and co-authors studied the evolution of MRI regions amongst primate species and found a very strong fixation bias for mutations within MRI regions. In other words, mutations rapidly erode MRI regions, in the direction of bringing their nucleotide composition toward genome-average levels. However, those mutations that favor the maintenance of MRI properties have a much higher chance to spread through the entire population. One of the straightforward explanations of this fixation bias phenomenon is the presence of purifying selection that tries to preserve valuable functions associated with MRI regions.

MRI regions are known to form unusual DNA conformations that have intrigued scientists since decades. CpG islands are known to be active methylation sites and can adopt Z-DNA form that act as transcriptional coactivators. AT rich regions can form antiparallel double helical duplexes in which they base pair through Hoogsteen hydrogen bonding instead of expected Watson-Crick pairing. The complementary DNA strands, A+G (purine) rich and C+T (pyrimidine) rich strands can form H-DNA triplexes that are sites for S1 nuclease cleavage and recombination hotspots. G+T (keto) rich regions are sites for binding of transcription factors like sp1. Guanines form G-4 quadruplexes in eukaryotic telomeres. The quadruplexes are arranged in four-stranded structures with strands connected to each other via Hoogsteen hydrogen bonding between guanines.
Intriguingly, G+T-rich oligonucleotides possess antiviral activities. For example, the 
T₂(G₄T₂)₃ sequence is virucidal against the herpes simplex virus. A+C (imino) rich 
regions are known to exist as tandem repeats (CA)ₙ and also regulate alternative splicing. 
Z-DNA has been detected near transcription start sites and also forms recombination 
hotspots.

2.2.4 Genomic signatures

It is well established that a particular base (A, G, C, or T) that appears in a given position 
of a genomic sequence significantly depends upon the nearest bases surrounding its 
position. Consequently, the frequency per site (F) of a dinucleotide XY is often not equal 
to the product of the individual frequencies per site of nucleotides X and Y \( (F_{XY} \neq F_X \times F_Y) \). 
The highest interdependence of base frequencies is always observed for adjacent 
nucleotides. The ratio \( (F_{XY} / (F_X \times F_Y)) \) for adjacent bases X and Y is known as a "genomic 
signature" and shown in the Table 1. Genomic signatures as low as 0.22 (for the CG 
dinucleotide in mouse) and as high as 1.75 (for the GC dinucleotide in Campylobacter 
jejuni) have been recorded. The interdependence of base frequencies sharply drops with 
increasing distance. In addition, many researchers described non-random distribution of 
a vast spectrum of genomic “words” (usually 5-10 base long oligonucleotides) or longer 
stretches of DNA, also known as “pyknons” (17-25 base-long). Here, we refer to this 
type of interdependency between nucleotides separated from each other by less than 30 
positions as short-range inhomogeneity (SRI).

Long range inhomogeneity (LRI) 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. 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. 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.

In this study we consider genomic patterns longer than 30 nucleotides and upto several thousands of nucleotides to be called Mid-Range scale or Mid-Range Inhomogeneity (MRI / genomic MRI).

2.3 Materials and Methods

Chromosome files from the Genbank version of the Human genome build 37, p5 was downloaded from NCBI ftp site ([ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/]). Each chromosome file is divided into contigs containing feature tables followed by genomic sequences. We extracted all the contigs from all chromosomes using our perl program **Contigs.pl** and selected the largest 10 contigs for further experiment. The details of the contigs so chosen are provided in Table 2.1. We further created seven different perl scripts corresponding to seven different MRI rich regions from Figure 2-1. An example would be that of **Window_X.pl**, where X could represent any MRI region (K, M, R, S, Y, W or Z). We designed our algorithm by choosing the nucleotide length for examining an
MRI region to be 100 nucleotides and a minimum threshold as >= 75% enrichment using a stretchy window approach. For example, an examination for existence of W (Adenine + Thymine) would be carried out by counting the adenine and thymine nucleotides within a particular 100 nucleotide stretch. If the enrichment of A+T regions is found to be less than 75%, the program will continue the examination by shifting across one nucleotide and then re-examining the genomic composition of the resulting window containing 100 nucleotides. If enrichment is however observed, the window adds up one more nucleotide thereby stretching itself to 101 nucleotides. It is then again examined for enrichment, stretching by a single nucleotide count if enrichment is found and terminating if the enrichment does not meet our threshold (>=75%). The nucleotide positions corresponding to a particular enriched region are recorded and the output corresponding to starting and ending positions of such enriched MRI regions are directed to output files forming MRI tables. Subsequently, those tables for a particular contig containing different MRI regions are concatenated together and then sorted, to arrange different MRI regions in order of their appearance across the length of the genome. Subsequently, another program MRI_text.pl processes the MRI tables into MRI text, characterized by letters representing MRI regions and inserted spaces proportional to the distance between them. For MRI regions located at a distance of less than 1kb from each other, the MRI symbols are placed adjacent to each other in the corresponding text. One space character is inserted for every 1kb distance between two MRI regions. From the MRI text, MRI signatures defined by the formula

\[ \delta = \frac{F_{XY}}{F_X * F_Y} \]  

(1)
are calculated through our program \textit{MRI\_signatures.pl}. In our algorithm for calculating MRI signatures, single MRI words with spaces preceding and following them (i.e. MRI regions located at distances of $> 1$kb from their nearest MRI neighbours) are not counted. Only words having more than one MRI character (i.e. MRI regions having neighbouring MRI words at a distance of less than 1kb) are considered. Occurrences of MRI regions in pairs as well as individual ones are counted from a concatenation of the entire text. The frequencies are thereby obtained using the \textit{MRI\_signatures.pl} program and the MRI signatures calculated. Figure 2-2 represents an example of the workflow.

Similar computations were also performed for the mouse genome. For this purpose, all Genbank versions of mouse chromosomes were downloaded from the NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/Mus\_musculus). The largest 10 contigs were similarly isolated and subjected to the same computations as for the human genome. The details of those contigs are provided in Table 2. Finally we also obtained MRI signatures for mouse genome. The enrichment criteria for human genome investigation (stretchy window approach starting from 100 nucleotides and threshold $\geq 75\%$) were unaltered for investigation of mouse MRI regions. The MRI signatures of mouse are presented in Table 4 along with side-by side comparison of identical signature values for humans.

### 2.4 Results

MRI signatures calculated for the human genome from formula 1 in Materials and Methods are displayed in Table 2.3. Alongside, genomic signatures of dinucleotides are also presented for comparison. The MRI signatures obtained from Table 2.3 demonstrate
that different MRI regions have a non-random arrangement relative to each other, thereby providing us with several pieces of information regarding the non-random composition of the human genome. It can be seen that at a distance of less than 1kb, GC rich regions (S) like very strongly to be with each other with their measure of association approximately equal to 10 times over that expected from a random model. Also, GC rich and AT rich regions (WS or SW) very strongly avoid each other at a distance of less than 1kb. Their measure of avoidance is ~ 14 times less than that expected from a random model. The association of ZDNA with GC rich region (SZ and ZS) was also found to be less by approximately 6 times from what would be expected in a random distribution. All of the obtained MRI signatures were contrasted with formerly obtained genomic signatures from Kariin et. al. in order to determine which regions preferred or avoided each other. In order to ascertain the non-randomness manifested in the MRI signatures, we also used perl programming (Shuffle_XMRI.pl) to shuffle different MRI regions one at a time, keeping other MRI regions constant. Figures 3, 4 and 5 show the results of genomic signatures obtained after randomization of K, W and Z respectively. MRI signatures from Table 1 also revealed other interesting aspects pertaining to the non-random genomic composition of humans. Some of the MRI regions displayed non-uniform signatures when the relative order was flipped with respect to each other (eg. MR vs RM). Reverting back to the original genomic signatures, we tried to look at MRI pairs that had a considerable difference in their signatures when positions of the pair of letters were interchanged and found some cases viz: MR/RM = 2.21/1.37 = 1.6 times ; YK/KY = 2.05/1.39 = 1.47 times ; RY/YR = 1.18/0.77 = 1.53 times. These values indicated a preference in the order of MRI regions at distances of less than 1kb. Against this biased
phenomenon, there were neutral pairs like \( \text{KM/MK} = 0.3/0.31 = 1; \text{RZ/ZR} = 0.64/0.62 \approx 1 \) that further provided contrasts and hence clues to the non-random compositional

---

**Figure 2-2. MRI calculation, characterization and display**

A: A portion of original genomic sequence within Human chromosome 9 inside which MRI regions have been characterized by pipelines of programs in Perl and Shell-script.

B: MRI co-ordinate table produced from sequence (A) using Perl shows conversion of sequences into specific MRI regions bounded by co-ordinates under captions Start and End. C: Textual representation of table (B) for all possible seven types of MRI regions (S is gc rich, W is at rich, K is gt rich, M is ac rich, R is ag rich, Y is ct rich, Z is alternating purine \((a/g)\) and pyrimidine \((c/t)\) rich. The space between two MRI regions \((Y\) and \(K\) shown as an example) is proportional to the distance between them from table (B).

---
<table>
<thead>
<tr>
<th>Largest Contigs</th>
<th>Chromosome number</th>
<th>Accession ID</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>NT_030059</td>
<td>79420533 bp</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>NT_029419</td>
<td>71516776 bp</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>NT_026437</td>
<td>88289540 bp</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>NT_022184</td>
<td>68452323 bp</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>NT_005403</td>
<td>84213159 bp</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>NT_005612</td>
<td>100537107 bp</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>NT_032977</td>
<td>90908613 bp</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>NT_016354</td>
<td>115591997 bp</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>NT_025741</td>
<td>75224524 bp</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>NT_007933</td>
<td>77412220 bp</td>
</tr>
</tbody>
</table>

**Table 2.1.** The largest 10 contigs of the human genome are shown. The largest contig is 115 million base pairs in length. Together the ten contigs tantamount to 830 million base pairs in length which is roughly equal to 29% of the human genome (3 billion nucleotides)
<table>
<thead>
<tr>
<th>Largest Contigs</th>
<th>Chromosome number</th>
<th>Accession ID</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>NT_039500</td>
<td>72389128 bp</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>NT_096135</td>
<td>87300923 bp</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>NT_039551</td>
<td>78455223 bp</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>NT_039606</td>
<td>105332540 bp</td>
</tr>
<tr>
<td>5</td>
<td>17</td>
<td>NT_039649</td>
<td>81232659 bp</td>
</tr>
<tr>
<td>6</td>
<td>18</td>
<td>NT_039674</td>
<td>73480409 bp</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>NT_039207</td>
<td>116378660 bp</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>NT_039240</td>
<td>109125955 bp</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>NT_187032</td>
<td>86266799 bp</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>NT_039353</td>
<td>108860100 bp</td>
</tr>
</tbody>
</table>

**Table 2.2.** The largest 10 contigs of the mouse genome are shown. The largest contig is 116 million base pairs in length. Together the ten contigs tantamount to 918 million base pairs in length which is roughly equal to 33% of the mouse genome (2.8 billion nucleotides)
Table 2.3: (A) Human genomic signatures of dinucleotides (Kariin et. al.) compared with MRI signatures (B) (calculated from MRI_signatures.pl, a Perl in-house program). The letters associated with MRI signatures represent \( S = \) GC-rich region, \( W = \) AT-rich, \( K = \) GT-rich, \( M = \) AC-rich, \( R = \) AG-rich, \( Y = \) CT-rich, \( Z = \) alternating purine and pyrimidine (See also Figure 1 and Figure 2). Genomic and MRI signatures (\( \delta \)) provide a measure of non-random association between the letters, and are calculated by the formula \( \delta = \frac{f_{XY}}{f_X f_Y} \), where \( f_{XY} \) represents frequency of the pair of nucleotides or MRI letters and \( f_X f_Y \) is a measure of the product of individual frequencies of the nucleotides/letters constituting the pair. In comparison with Kariin et al. that established over or under-representation of dinucleotides due to preferential association or avoidance between them respectively, over a range of 1.27(CC) to 0.27(CG), the MRI signatures demonstrate much higher range and variability from 9.94(SS) to 0.07(WS). Their association varies from very strong (SS, YK highlighted in pink) to unbiased (SY highlighted in turquoise) to negative tendency of clustering (YM highlighted in green) to very strong and outright avoidance (WS highlighted in yellow).

<table>
<thead>
<tr>
<th>A. Genomic signatures</th>
<th>B. MRI signatures</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA 1.16</td>
<td>SS 9.94</td>
</tr>
<tr>
<td>AC 0.80</td>
<td>WZ 1.37</td>
</tr>
<tr>
<td>AG 1.19</td>
<td>SY 0.95</td>
</tr>
<tr>
<td>AT 0.85</td>
<td>YW 0.65</td>
</tr>
<tr>
<td>CA 1.15</td>
<td>KS 0.51</td>
</tr>
<tr>
<td>CC 1.27</td>
<td>MR 2.21</td>
</tr>
<tr>
<td>CG 0.27</td>
<td>SR 1.23</td>
</tr>
<tr>
<td>CT 1.19</td>
<td>WR 0.70</td>
</tr>
<tr>
<td>ZK 1.42</td>
<td>ZZ 0.61</td>
</tr>
<tr>
<td>KY 1.39</td>
<td>ZS 0.16</td>
</tr>
<tr>
<td>ZW 1.49</td>
<td>YM 0.59</td>
</tr>
<tr>
<td>ZY 0.66</td>
<td>WS 0.07</td>
</tr>
<tr>
<td>ZR 0.62</td>
<td>KM 0.31</td>
</tr>
<tr>
<td>ZR 0.62</td>
<td>MS 0.42</td>
</tr>
<tr>
<td>ZR 0.62</td>
<td>SK 0.45</td>
</tr>
<tr>
<td>ZR 0.62</td>
<td>MR 0.31</td>
</tr>
<tr>
<td>ZR 0.62</td>
<td>MS 0.42</td>
</tr>
<tr>
<td>ZR 0.62</td>
<td>SK 0.45</td>
</tr>
<tr>
<td>ZR 0.62</td>
<td>MR 0.31</td>
</tr>
<tr>
<td>ZR 0.62</td>
<td>MS 0.42</td>
</tr>
<tr>
<td>ZR 0.62</td>
<td>SK 0.45</td>
</tr>
</tbody>
</table>
patterns within the human genome. We picked up one particular case RY/YR and investigated the reason for bias in MRI signatures.

The computational definitions of the MRI regions were modified in our programs. To identify strongly enriched purine and pyrimidine regions, our original sequences from the largest ten contigs were subjected to a programming based on 50 nucleotide stretchy window approach and a greater than 95% of the nucleotide composition (A+G for R rich and C+T for Y rich). We obtained MRI tables with coordinates of start and end similar to those explained in materials and methods, with the exception that these MRI tables specified MRI enrichments pertaining to R and Y only. The region in between the end of an MRI region and the start of the next MRI region was divided into bins of 10, 100 and 1000 nucleotide bases through different computational programs and numbers of occurrences of alternating purine/pyrimidine nucleotide stretches were obtained. Across a span of 500 nucleotides apart, a ratio of \( \text{RY/YR} = \frac{88}{47} \) was obtained. We applied our own monte carlo perl simulator program to ascertain the statistical significance associated with the observation and found a strong significance (\( p\text{-value} = 3.7\times10^{-6} \)). This ratio of RY/YR was found to be specific only to humans validated by genomic investigations of mouse, cow, chicken and Drosophila. We were unable to find possible reasons for the occurrence of purine/pyrimidine compositional bias in humans, such as that exhibited through a bias in relative MRI signatures. The MRI signatures from the genome of Mus musculus are presented alongside human MRI signatures in Table 2.4.

While mouse MRI signatures represent non-random patterns specific to its genome, some common points of observation with respect to both human and mouse genomes are worth noting. The very strong clustering of GC rich regions in human genomes (SS) was further
Figure 2-3. MRI signatures calculated due to randomization of K, keeping other MRI regions constant. This revealed change of MRI signatures involving K towards random genomic compositions, shown by the trend of randomized MRI signatures to be closer to one. Since other MRI regions were not randomized for this particular experiment, only those MRI signatures are compared to that involving K. The other MRI signatures stayed constant and hence are not presented in this graphical analysis.
Figure 2-4. MRI signatures calculated due to randomization of $W$, keeping other MRI regions constant. This revealed change of MRI signatures involving $W$ towards random genomic compositions, shown by the trend of MRI signatures to be closer to one. Since other MRI regions were not randomized for this particular experiment, only those MRI signatures are compared to that involving $W$. The other MRI signatures stayed constant and hence are not presented in this graphical analysis.
Figure 2-5. MRI signatures calculated due to randomization of Z, keeping other MRI regions constant. This revealed change of MRI signatures involving Z towards random genomic compositions, shown by the trend of MRI signatures to be closer to one. Since other MRI regions were not randomized for this particular experiment, only those MRI signatures are compared to that involving Z. The other MRI signatures stayed constant and hence are not presented in this graphical analysis.
<table>
<thead>
<tr>
<th>MRI types</th>
<th>MRI signatures (Homo Sapiens)</th>
<th>MRI signatures (Mus musculus)</th>
</tr>
</thead>
<tbody>
<tr>
<td>'KK'</td>
<td>2.095</td>
<td>0.861</td>
</tr>
<tr>
<td>'KM'</td>
<td>0.318</td>
<td>0.433</td>
</tr>
<tr>
<td>'KR'</td>
<td>0.645</td>
<td>0.833</td>
</tr>
<tr>
<td>'KS'</td>
<td>0.516</td>
<td>0.214</td>
</tr>
<tr>
<td>'KW'</td>
<td>0.595</td>
<td>0.724</td>
</tr>
<tr>
<td>'KY'</td>
<td>1.393</td>
<td>1.046</td>
</tr>
<tr>
<td>'KZ'</td>
<td>1.076</td>
<td>2.074</td>
</tr>
<tr>
<td>'MK'</td>
<td>0.31</td>
<td>0.522</td>
</tr>
<tr>
<td>'MM'</td>
<td>2.351</td>
<td>0.844</td>
</tr>
<tr>
<td>'MR'</td>
<td>2.218</td>
<td>1.386</td>
</tr>
<tr>
<td>'MS'</td>
<td>0.42</td>
<td>0.31</td>
</tr>
<tr>
<td>'MW'</td>
<td>0.697</td>
<td>0.704</td>
</tr>
<tr>
<td>'MY'</td>
<td>0.687</td>
<td>0.816</td>
</tr>
<tr>
<td>'MZ'</td>
<td>1.223</td>
<td>1.84</td>
</tr>
<tr>
<td>'RK'</td>
<td>0.658</td>
<td>0.834</td>
</tr>
<tr>
<td>'RM'</td>
<td>1.374</td>
<td>1.054</td>
</tr>
<tr>
<td>'RR'</td>
<td>2.58</td>
<td>1.532</td>
</tr>
<tr>
<td>'RS'</td>
<td>0.962</td>
<td>0.51</td>
</tr>
<tr>
<td>'RW'</td>
<td>0.717</td>
<td>0.902</td>
</tr>
<tr>
<td>'RY'</td>
<td>1.18</td>
<td>1.153</td>
</tr>
<tr>
<td>'RZ'</td>
<td>0.649</td>
<td>0.612</td>
</tr>
<tr>
<td>'SK'</td>
<td>0.45</td>
<td>0.322</td>
</tr>
<tr>
<td>'SM'</td>
<td>0.596</td>
<td>0.24</td>
</tr>
<tr>
<td>'SR'</td>
<td>1.232</td>
<td>0.431</td>
</tr>
<tr>
<td>'SS'</td>
<td>9.94</td>
<td>14.112</td>
</tr>
<tr>
<td>'SW'</td>
<td>0.075</td>
<td>0.062</td>
</tr>
<tr>
<td>'SY'</td>
<td>0.958</td>
<td>0.507</td>
</tr>
<tr>
<td>'SZ'</td>
<td>0.157</td>
<td>0.101</td>
</tr>
<tr>
<td>'WK'</td>
<td>0.698</td>
<td>0.696</td>
</tr>
<tr>
<td>'WM'</td>
<td>0.569</td>
<td>0.67</td>
</tr>
<tr>
<td>'WR'</td>
<td>0.702</td>
<td>0.846</td>
</tr>
<tr>
<td>'WS'</td>
<td>0.074</td>
<td>0.107</td>
</tr>
<tr>
<td>'WW'</td>
<td>1.269</td>
<td>1.475</td>
</tr>
<tr>
<td>'WY'</td>
<td>0.713</td>
<td>0.894</td>
</tr>
<tr>
<td>'WZ'</td>
<td>1.376</td>
<td>1.63</td>
</tr>
<tr>
<td>'YK'</td>
<td>2.053</td>
<td>1.29</td>
</tr>
<tr>
<td>'YM'</td>
<td>0.599</td>
<td>0.876</td>
</tr>
<tr>
<td>'YR'</td>
<td>0.776</td>
<td>1.012</td>
</tr>
</tbody>
</table>
Table 2.4. MRI signatures for human and mouse genomes. The highlighted values show a contrast between some of the relative associations of identical MRI regions between the two species. While identical MRI regions have a strong tendency to cluster with each other in human genomes, they have a preferentially lower tendency to do so in mouse genome as reflected across the signature (with the exception of association of mouse GC rich regions). The highlighted MRI signatures indicate mouse genomic compositions towards genome average levels.

intensified in mouse, with a measure of association reflected by ~ 16 times over that expected from a random genomic composition. The extremely strong avoidance of GC rich regions with ZDNA (ZS, SZ) as well as those of GC rich regions with AT rich regions (SW, WS) was illustrated across both organisms, suggesting evolutionary conservation within specific patterns of MRI.

2.5 Conclusions
Using the concept of MRI signatures for the very first time, we identified intricate non-random patterns within human and mouse genomic compositions. Also a comparison of MRI signatures between humans and mouse revealed the uniqueness and specificity of mid-range inhomogeneity. Among relative associations of nucleotide stretches of specific compositions, the very strong associations of GC rich regions with each other existed in both humans and mouse. A very strong avoidance of GC rich regions with AT rich regions as well as avoidance of GC rich regions with ZDNA was also manifested through our computations. The arrangements of MRI regions are indicative of the complexity of the genome machinery, which still continues to baffle scientists.

2.6 Discussions

We made two simplified approximations for the computation of MRI regions. Firstly, nucleotide stretches containing the same MRI regions separated with a small gap (less than 20 nucleotides) were considered as an uninterrupted stretch of that particular MRI. This is because, at a distance of less than 20 bp, the overall nucleotide composition may drop from our criteria of specified threshold (≥ 75%) but the possibility of nucleotides enriched with a different kind of MRI can be ruled out. Even if those nucleotides in between are enriched with different nucleotide compositions, such phenomenon would be more appropriate to be considered in the context of short-range inhomogeneity (SRI). For our purpose of investigation of Mid Range Inhomogeneity, since our definitions of MRI regions hold true for nucleotide stretches more than 30bp in length, a region of less than 20 nucleotides surrounded by identical MRI regions can be assumed to be part of the
same MRI. Secondly, since our MRI regions are based on identifications of overall nucleotide compositions, specific nucleotide stretches could be identifiable by more than one particular MRI type. A simple example would be that of stretches of GT\textsubscript{n} repeats, which could be classified as both ZDNA (alternating purines and pyrimidines) and K (gt rich region). Overlaps between different MRI regions can take place in several areas. However, this can also lead to the presence of an MRI region completely within another. For example, few hundred nucleotides that are overall GT rich (K region) could have one hundred nucleotides in between specifically containing stretches of GT repeats and thus identifiable as ZDNA. In this case the starting and ending position of the particular ZDNA region would be well within the encompassing K region. For all those cases of MRI, where the start and end of a particular MRI region fall inside the coordinates specified by a different MRI type, the nucleotide stretches are defined by the MRI containing the larger span of nucleotides.

The MRI signatures obtained from our computations are the first steps towards appreciation of human genome as a complex set of texts or instructions. Indeed every genome can be hypothesized to be a complex programming language comprising several characters represented by the nucleotide bases, replete with components driving the machinery of an organism. Aspects of such genomic behavior are captured in the review paper by Fedorov and Fedorova. Exhibiting various patterns and must be of crucial importance for proper functioning of multicellular eukaryotes. Our data presents an example of patterns embedded within the genome.
2.7 References


3. Isidore Rigoutsos: Short blocks from the noncoding parts of the human genome have instances within nearly all known genes and relate to biological processes. PNAS 2006, 103: 6605-6610


Chapter 3

Bioinformatics analysis of plant orthologous introns: identification of an intronic tRNA-like sequence

Evgeny E. Akkuratov\textsuperscript{1,2}, Lorraine Walters\textsuperscript{3,2}, Arnab Saha-Mandal\textsuperscript{2}, Sushant Khandekar\textsuperscript{4}, Erin Crawford\textsuperscript{5}, Craig L. Zirbel\textsuperscript{6}, Scott Leisner\textsuperscript{4}, Ashwin Prakash\textsuperscript{5,7}, Larisa Fedorova\textsuperscript{5}, Alexei Fedorov\textsuperscript{2,5*}

\textsuperscript{1}International Biotechnology Center M. V. Lomonosov Moscow State University, Moscow 119991, Russia.

\textsuperscript{2}Program in Bioinformatics and Proteomics/Genomics, University of Toledo, Health Science Campus, Toledo, OH 43614, USA.

\textsuperscript{3}Department of Bioengineering, University of Toledo, Main Campus, Toledo, OH 43606, USA.

\textsuperscript{4}Department of Biological Sciences, University of Toledo, Main Campus, Toledo, OH 43606, USA.
5Department of Medicine, University of Toledo, Health Science Campus, Toledo, OH 43614, USA.
6Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403, USA.
7Current address: Department of Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA.

*To whom correspondence should be addressed. Tel +419-383-5270; Fax +419-383-3102; e-mail: Alexei.fedorov@utoledo.edu

akkuratov.evgeny@gmail.com; lorraine.walters@rockets.utoledo.edu;
arhab.sahamandal@rockets.utoledo.edu; sushant43606@yahoo.com;
erin.crawford@utoledo.edu; zirbel@bgsu.edu; scott.leisner@utoledo.edu;
ashwin.prakash@jhmi.edu; larisa.fedorova@utoledo.edu; alexei.fedorov@utoledo.edu

Keywords: genomics, computational biology, MALAT1, mascRNA

My Contributions

Computational analysis of RNA secondary structures (including Figure 3-3); BLAST analysis of conserved plant intronic regions (from Figures S1-S3); Multiple alignment and phylogenetic analysis of our conserved intronic sequence against Arabidopsis thaliana tRNA database (including Figure S6); Mapping gene functions from Gene Ontology database and investigating enrichment of functions using Monte-Carlo simulator Perl program (including Figure S5); Computation of intronic tRNAs in vertebrate genes (including Tables S1 and S2); Overall writing of manuscript.
3.1 Abstract

**Background:** Orthologous introns have an identical position relative to the coding sequence in orthologous genes of different species. Despite orthologous introns from evolutionarily distant species often share a limited sequence similarity, they still have a common origin because introns are essentially unable to change their positions. Therefore, orthologous introns present a distinctive opportunity to study evolution of non-coding DNA regions. The goal of our study is to search for conserved sequence elements in plant introns, to compare this phenomenon to animals and to characterize the found conserved elements.

**Results:** By analyzing the complete genomes of five plants we generated a database of 40,512 orthologous intron groups of dicotyledonous plants, 28,519 orthologous intron groups of angiosperms, and 15,726 of land plants (moss and angiosperms). Multiple sequence alignments of each orthologous intron group were obtained using a Mafft algorithm. The number of conserved regions in plant introns appeared to be hundreds of times less than in mammals or vertebrates. Approximately three quarters of conserved intronic regions among angiosperms and dicots in particular, correspond to alternatively-spliced exonic sequences. We registered only a handful of conserved intronic ncRNAs of flowering plants. However, the most evolutionarily conserved intronic region, that is ubiquitous for all plants examined in this study, including moss, possessed multiple structural features of tRNAs, which caused us to classify it as putative tRNA-like ncRNA. Intrinsic sequences encoding tRNA-like structures are not unique to plants. Bioinformatics examination of the presence of tRNA inside introns revealed an unusually
long-term association of four glycine tRNAs inside the \textit{Vac14} gene of fish, amniotes, and mammals.

**Conclusions:** Plant introns contain hundreds of times fewer conserved ncRNAs than do the introns of animals. We assessed that a specific genomic position of a structural RNA may provide it an additional function in regulation of neighboring genes expression. Such a new role might evolve into a new type of ncRNA that preserves its acquired regulatory role and may abandon its original function.

### 3.2 Background

Introns are ubiquitous elements of eukaryotic genomes that perform several essential cellular functions [1-3]. The relative abundance and the length of introns significantly varying diverse branches of eukaryotes indicating that ranks of intron functions may be different for evolutionarily distant species. This paper is focused on comparison of intron roles in plants and animals.

Usually, functional regions of introns have evolutionarily conserved nucleotide sequences that have been preserved over millions of years [4, 5]. These functional intronic sequences may be one of the following types: 1) protein binding sites (enhancers or silencers for transcription or splicing); 2) alternatively-spliced exonic sequences that are incorporated into mRNAs for selective transcripts of the gene; and 3) non-coding RNA molecules including small ncRNAs (snoRNAs, microRNAs, endogenous siRNA, and piwiRNA) and long ncRNA (analogous to lincRNA). Recently, our team generated a database of orthologous introns for five mammalian species [4]. Within this set of 63,000
groups of orthologous introns, thousands of evolutionarily conserved sequence segments were characterized and associated with various types of non-coding RNAs. In the current study, we took advantage of the availability of recently sequenced genomes of moss, poplar, as well as grape along with other plants and performed the same bioinformatics analysis of orthologous introns within kingdom Plantae. We generated Exon-Intron Databases for moss \( (\text{Physcomitrella patens}; \text{Pp}) \), rice \( (\text{Oryza sativa}; \text{Os}) \), poplar \( (\text{Populustrichocarpa}; \text{Pt}) \), grape \( (\text{Vitus vinifera}; \text{Vv}) \), and mouse ear cress \( (\text{Arabidopsis thaliana}; \text{At}) \). Then, we created databases of orthologous introns of dicots, angiosperms, and land plants. Sequences of each orthologous group were aligned and conserved regions were characterized. This large-scale computational investigation revealed a prominent difference in the number, length and location of conserved intronic regions of plants and animals.

To compare the evolution of plant and animal intron sequences, different branches of vertebrates were used that evolved during approximately the same period of time as the plant taxa. Vertebrates were chosen because a number of their genomes have been completely sequenced and characterized in great detail. The entire group of our five land plant species (in which moss separated from the flowering plants approximately 420 million years ago, mya) may be compared with \( \text{Osteichthyes} \) taxon of bony fish and tetrapods that has ~ 500 mya history. The ancestors of angiosperms diverged from gymnosperms around 245–202 mya, and the first angiosperms known to exist were from 140 mya[6]. Hence, flowering plants may be likened to eutherian and marsupial mammals that diverged about 190-180 mya. Finally, dicotyledonous plants, which appeared in the fossil records as early as 110 mya
(http://lifeofplant.blogspot.com/2011/04/eudicots.html), may correspond to the eutherian taxon of placental mammals, that originated about 100 mya [7].

We acknowledge the availability of alternative public resources that present plant exon-intron datasets and in particular, the Common Introns Within Orthologous Genes (CIWOG) database described by Wilkerson et al. 2009[8]. The advantage of this database is that it presents genes for eight plant species and has user-specified options to control common intron detection in regions of poor alignment quality and putative intron sliding. Another recent database built with the same objective is Plant Intron and Exon Comparison and Evolution (PIECE)[9]. The orthologous intron groups characterized in this paper are highly consistent with those obtained in CIWOG database.

3.3 Results

3.3.1 Databases of plant orthologous introns

A database of orthologous introns of five plant species was generated using the same algorithms and pipelines of programs as described previously for mammalian orthologous introns [4]. Computation of plant intronic sequences for five species of land plants (arabidopsis, poplar, grape, rice, and moss) produced 15,726 orthologous intron groups, while four species of angiosperms (arabidopsis, poplar, grape, and rice)–resulted in 28,519 groups. When the computation was confined to three species of dicotyledonous plants (arabidopsis, poplar, and grape), 40,512 groups were obtained. Each group of orthologous introns was aligned with the Mafft program[10] and the entire set of multiple alignments is available from our Exon-Intron Database web site.
An evolutionarily conserved intronic segment may represent alternatively spliced exon(s), non-coding RNA, or functional regions involved in the expression regulation [1]. For computational finding of short ncRNAs (miRNA and endogenous siRNA), transcription factor binding sites, and alternative donor and acceptor splicing sites, we examined multiple alignments of plant orthologous introns for short regions (10-20 nucleotides) with strong sequence identity (≥80%). For computer characterization of alternatively skipped exons and snoRNAs, we searched for longer evolutionary conserved regions (120 nucleotides) with less percentage of identity (≥60%). Finally, for characterization of putative long ncRNAs, we increased the length to 400 nucleotides and relaxed the strength of identity to 50%.

3.3.2 Conserved intronic regions in dicotyledonous plants

The number of conserved intronic sequences in dicots is three orders of magnitude less than those in placental mammals. Indeed, previously within 63,077 orthologous intron groups of five eutherian species (mouse, rat, dog, cow, and human) we characterized thousands of evolutionarily conserved intronic sequences including 9,833 cases with long (>400 nts) conserved regions showing at least 50% identity[4]. In contrast, among 40,512 groups of orthologous introns of grape, poplar, and Arabidopsis, we found only two orthologous introns containing long (>400 nucleotides) evolutionarily conserved (>=50% identity) regions (Fig. 1). These two conserved intronic regions represent alternative exons described in the figure. We also examined the distribution of short ultra-conserved regions (20 invariable sequential nucleotides) within orthologous introns.
Only seven dicot introns share such short ultra-conserved sequences (Fig.S1). In contrast, among 63,077 orthologous intron groups of 5 placental mammals, 3,211 of them contain 20 nucleotide-long stretches of invariable nucleotides for all 5 species.

In order to estimate the impact of alternative splicing and ncRNAs on the conserved intronic regions in plants, we identified and examined the entire set of 25 groups of orthologous introns of arabidopsis, grape and poplar, which contain conserved regions with at least 60% identical bases within a sequence window spanning over 120 nucleotides. This set of 25 orthologous groups is presented in Figure S2. Sequences of these introns have been compared with plant EST and ncRNA databases using the BLAST program [11] in order to decipher their possible functions. This examination, explained in Figure S2, revealed the following: 1) Conserved regions within seventeen introns correspond to alternatively spliced exons. 2) Two correspond to ncRNAs (snoRNAs). 3) Three cases are likely not alternatively spliced exons because of numerous EST blast-hits from corresponding genes with no matches against intronic regions. These three intronic sequences likely represent unknown ncRNAs or DNA/RNA functional regions (e.g. enhancers or silencers). 4) For the remaining three introns, we were unable to identify possible functions because of their poor representation in EST database and no hits with ncRNA databases.

3.3.3 Conserved intronic regions in flowering plants

The number of conserved intronic regions even in a recent branch of plants (dicots) is relatively small. Therefore, in order to select a majority of these regions in broader groups of plants we applied computational algorithms with relaxed parameters for
characterization of conserved regions. Specifically, we consider that an intron contains an evolutionarily conserved region when it has a stretch of ten bases containing at least nine invariable nucleotides for all studied species. Examination of orthologous introns of four flowering plants (including three dicots and one monocotyledonous plant -- rice) revealed 52 groups that obey this sequence conservation criterion (invariant nts=9; window size =10nts). Among these 52 cases, 18 correspond to internal intronic regions; 15 are short regions at the intron 5′ terminus; 18 are short regions at the intron 3′ terminus; and one case (case #51 in Figure S3) contained a short conserved region at the 5′-terminus and in the middle of the same intron. Three cases #43, 48, and 50 from the Figure S3 with conserved 5′-intron termini contain the consensus for U12-spliceosomal introns RTATCCTTT, that are well known for their evolutionary conservation [12]. Among the aforementioned 34 cases of short conserved regions at the intron termini, 22 have additional evolutionarily conserved cryptic splicing sites within 10 nucleotides of the intron terminus (Highlighted in yellow on figure S3). Examination of the plant EST database confirmed that cryptic sites likely have participated in splicing in 18 out of these 22 cases indicated in Figure S3. The utilization of these cryptic sites for splicing results in short insertions in mRNAs that always cause a shift in the reading frame. Two of these most evolutionarily conserved cases with alternative donor and acceptor splicing sites are also described below (CASE 5 and 6 for land plants in the Figure S4). This observation is consistent with previously published results of frequent small insertions/deletions in mRNAs due to alternative donor or acceptor splicing sites in close vicinity to the major site [13, 14].
3.3.4 Conserved intronic regions common for moss and angiosperms

Only seven out of 15,726 orthologous intron groups representing moss, rice, arabidopsis, poplar, and grape have evolutionarily conserved regions characterized by the same computational filter -- 9 invariant nucleotides within 10-nucleotide long scanning window. The case with the longest and most stringently-conserved region, that appeared most impressive for these five species is shown in the Figure 2, while the remaining six cases with considerable nucleotide conservation (invariant nts =9; window size =10nts) are shown in the Figure S4. The conserved intronic sequence indicated in Figure 2 does not match any functional ncRNA from public databases through execution of online BLAST searching programs, nor corresponds to alternative splicing events based on EST examination using NCBI online EST dataset. Online BLAST examination did not detect the presence of this conserved sequence in any of the genomes outside land plants including Chlamydomonas reinhardtii. We termed this region CPIR-1 (conserved plant intronic region 1) and investigate it in detail below. Another of the remaining six conserved sequences indicated in Figure S4 (CASE 1) corresponds to a snoRNA (R104 of Arabidopsis). The next sequence in Fig. S4 (CASE 2) corresponds to an exonic sequence in paralogous gene (this particular case is also described in the Figure 1B for three dicot species). The next two sequences, CASE 3 and CASE 4, represent very short (10-14 nucleotides) conserved regions in the middle of introns that might be transcription factor binding sites (e.g. enhancers or silencers) or small ncRNAs. Finally, the last two sequences, CASE 5 and CASE 6 from Figure S4 represent short (10 nucleotides) conserved regions at the 5’- and 3’-intron termini, respectively. These two terminal conserved intronic sequences contain cryptic splice sites within the fourth and seventh
nucleotides from the intron ends (shown in red in the figure S4). Plant EST examination confirmed that, indeed, for the CASE 5 the cryptic site is used for alternative splicing in 31% of transcripts for a variety of flowering plant species. ESTs representing this cryptic splice site have a 4-nucleotide insertion that causes a shift in the reading frame. Among the 16 EST sequences representing the tetratricopeptide repeat-containing gene of the sequence for CASE 6, only a single EST (HO807763.1) utilizes the described cryptic acceptor splicing site. Activation of this cryptic site causes a 7-nucleotide-long insertion and thus, a shift in the reading frame. All in all, plants as diverse as mosses and angiosperms share only a few conserved intronic regions. In contrast, among vertebrates of the Osteichthyes taxon, the number of conserved intronic regions of similar sequence conservation is estimated to be in the hundreds.

Finally, no statistically significant sequence similarity was found between plant and mammalian intronic conserved regions except a single case with a short 5’-terminal region of U12-type Arabidopsis intron (case #43 from Figure S3, Arabidopsis INTRON_2_15034_NC_003075 that is identical to its human orthologous intron #4 of BRCC3 gene, UniGene: HS.558537). Evolutionarily conserved structures may be more prominent than conserved sequences in ncRNAs. However, the examination of RNA structure conservation requires manual individual approaches and has not been broadly conducted in this study.

Functions of plant genes with intronic evolutionary conserved sequences from Figures S1-S3 have been identified via NCBI and Gene Ontology databases and described in the Supplementary Figure S5. Statistically significant enrichment (p<0.001) of genes involved in nucleic acid binding (17 genes with GO:0003676), RNA binding (15 genes
with GO:0003723), nucleotide binding (14 genes with GO:0000166), positive regulation of transcription (10 genes with GO:0045893), mRNA splicing (7 genes with GO:0000398), RNA processing (7 genes with GO:006396) and RNA splicing (5 genes with GO:0008380) were found using Monte-Carlo simulations.

3.3.5 Characterization of tRNA-like plant intronic putative ncRNA

The most evolutionarily conserved plant intronic region (named CPIR-1; Fig. 2) described above, contains two blocks of conserved regions (highlighted in blue and yellow on the figure) that are separated by a short 10-20 nucleotide-long linker. The sequence of CPIR-1 has been preserved inside plant introns of an RNA-polymerase III inhibitor gene (GenBank accession number AT5G13240.1 from Arabidopsis thaliana) for at least 420 million years (the divergence time between moss and flowering plants). This gene is represented by > 500 copies in the EST database with BLAST p-values <10^-8. Therefore, the absence of hits of CPIR-1 sequence with plant ESTs indicates that this conserved region is unlikely an alternative splicing exon but rather, an unknown ncRNA. To decipher possible functions of this conserved region, we applied an online RNAfold program to find stable local 2D stem-loop structures inside CPIR-1 sequence. Then we analyzed possible spatial conformations of these local RNA structures the JAR3D web server (http://rna.bgsu.edu/jar3d) that compares the input sequence with all known RNA 3D motifs available in the RNA 3D Motif Atlas (http://rna.bgsu.edu/rna3dhub/motifs). This examination revealed that two regions (shown in magenta and red on Figure 2) inside the conserved yellow block showed significant 3D-structural matches with the D-loop and T-loop of tRNAs correspondingly. This result led to analyze the folding of the
yellow block alone, which revealed a secondary structure closely resembling that of tRNAs. Computer-predicted 2D structures of this RNA segment are shown in Figure 3A-E for the five plants examined. In addition, we co-folded these five sequences together using RNAalifold online program from Vienna RNA web-servers in order to characterize evolutionarily conserved RNA 2D structures. The output of this analysis is present on the Figure 3G. This figure demonstrates that the evolutionarily conserved 2D-structure of the yellow CPIR-1 region is remarkably similar to tRNAs. Table S1 (provided in Appendix B.1) depicts characteristics common for all transfer RNAs and compares their presence in plant CPIR-1 sequences. Importantly, we observed eight compensatory substitutions in the putative amino acid and anticodon stems that maintain secondary structure of base paring, but disrupt nucleotide sequence conservation. At the same time, one of the most important functional elements of tRNA – the anticodon loop, is not conserved and drastically different among the five analyzed species. Due to this reason, the tRNAscan-SE online tool does not recognize our CPIR-1 sequence or score it as a tRNA. Thus, the CPIR-1 sequence is not a functional tRNA, but rather, might be a putative tRNA-like ncRNA.

BLAST comparison of Arabidopsis CPIR-1 sequence against entire set of 639 Arabidopsis tRNAs did not reveal any significant match. However, Clustal-Omega alignment of Arabidopsis CPIR-1 with Arabidopsis tRNAs revealed 56% identity of CPIR-1 to Proline tRNAs (Figure S6). Thus, CPIR-1 might have originated from a Pro-tRNA sequence.

Intriguingly, this tRNA-like sequence is inside an intron of RNA-polymerase III inhibitor gene. Since many tRNA molecules are transcribed by RNA-polymerase III using internal
promoters inside tRNA sequences, we hypothesized that CPIR-1 may also be transcribed and this transcription may influence the expression of the RNA-polymerase III inhibitor gene. In order to test this hypothesis, we performed Northern analysis of total RNA extracts obtained from various Arabidopsis tissues and from whole plants grown under different propagation conditions (see Figure 4). As a probe, we used the entire sequence of the Arabidopsis intron containing CPIR-1 and the adjacent 5’- and 3’-exons. Thus, this probe should hybridize with RNA-polII inhibitor mRNA and our putative tRNA-like ncRNA. The results on Figure 4 clearly demonstrate that the RNA-polIII inhibitor gene is indeed expressed in each Arabidopsis tissue (roots, rosette leaves and inflorescence; A). Likewise, the RNA polIII transcript can be detected in plants propagated under a variety of growth conditions (dark treatment, cold treatment, cold treatment and then recovery from cold; B). However, no hybridization signal with CPIR-1 in the low-molecular-weight region (20-200 nucleotides) was detected in any Arabidopsis tissue or in whole plants propagated under different growth conditions.

3.3.6 Intrinsic tRNAs of vertebrates

Since the association of tRNA with introns is not well reported in the literature, we performed large-scale bioinformatics examination of this issue in vertebrates. Computer analysis of GenBank feature-tables for the latest release of the human genome demonstrated that 24 tRNAs are annotated inside introns of 14 human protein-coding genes that have well-defined functions (the data presented in Supplementary Table S2 or Appendix B.2). If all 631 human tRNAs annotated in the GenBank were randomly distributed inside the human genome, on average 157 tRNAs should be inside introns that
occupy ~25% of the genome. Hence, there is a strong avoidance of tRNAs to be within human introns. We also explored whether human intronic tRNAs are also present inside orthologous introns of mouse and other species. This investigation showed that these tRNAs are absent in corresponding orthologous mouse introns for all genes except one, named Vac14. Reciprocally, several mouse genes have intronic tRNAs that are not present in corresponding human orthologous genes. Further examination demonstrated that Vac14 has two Gly-tRNAs in intron 1 and two Gly-tRNAs in intron 9 in both human and mouse. These four intronic Gly-tRNAs exist in other mammalian species and in chicken, green anole, and zebrafish. However, no tRNA was found in Vac14 introns of more evolutionarily distant species such as jawless fish -- lancelet (Branchiostoma floridae), sea squirt (Ciona intestinalis), and sea urchin (Strongylocentrotus purpuratus). All in all, a unique evolutionarily conserved association of tRNA with introns exists only in the Vac14 gene of vertebrates and it has persisted for more than 500 million years (the time of separation of bony fish from tetrapods).

### 3.4 Discussion

Evolutionarily conserved intronic regions usually represent important functional elements such as: non-coding RNAs, alternatively spiced exonic regions, and transcription regulatory elements[1]. In this paper we demonstrated that the number of conserved regions in plant introns is less by two-to-three orders of magnitude than in vertebrates. Plant and vertebrate taxa for this comparison were chosen in such a way that they evolved during approximately the same period of time. However, the rate of nucleotide substitutions varies from species to species [15, 16]. Recent deep-sequencing
investigation demonstrated that \textit{Arabidopsis} has spontaneous mutation rate of $7 \times 10^{-9}$ base per site per generation \cite{17}. On the other hand, Smith and Donoghue \cite{18} demonstrated that trees and shrubs were evolving approximately 2.7-10 times more slowly than related herbaceous plants. According to these authors, the majority of trees/shrubs have 0.5-1.5 x $10^{-9}$ substitutions per site per year, while this parameter for herbs usually is in the range of 1-4 x $10^{-9}$. Among the four flowering plants examined in this paper, two belong to trees/shrubs (poplar and grape) and two to herbs (\textit{Arabidopsis} and rice). On the other hand, the mutation rate in mammalian genomes is approximately $2.2 \times 10^{-9}$ per base per year according to Kumar & Subramanian \cite{19}. Taken together, even if we assume that plants have slightly higher frequency of mutations in their genomes than mammals, it cannot explain drastic difference in the number of conserved intronic regions between these two taxa. Therefore, our results indicate that either plants have much lower numbers of alternatively splicing isoforms and intronic ncRNAs, or these sequences in plants are not under strong purifying selection pressure and, thus, evolve much faster than in vertebrates.

We demonstrated that at least three quarters of conserved intronic regions of dicots correspond to alternatively-spliced exonic sequences. In the majority of cases, these alternative exons have small insertions/deletions in different species causing shifts in the reading frame. This observation is in line with the data provided by Severing and co-authors \cite{20} regarding minimum involvement of evolutionarily conserved alternative splicing for the increase of proteome diversity in flowering plants. In addition, our results support the findings of Filichkin and others \cite{21} that 78\% of alternative splicing events in \textit{Arabidopsis} create premature termination codons (PTC) and are likely involved in the
regulation of expression via nonsense-mediated decay (NMD) and regulated unproductive splicing and translation (RUST) mechanisms.

Hundreds of thousands of small ncRNAs have been described for *Arabidopsis* [22]. So, it was a surprise for us to find only a few evolutionarily conserved intronic ncRNAs in this bioinformatics investigation. Nonetheless, we revealed the most evolutionarily conserved intronic sequence of higher plants (CPIR-1) and classified it as a putative tRNA-like ncRNA. This classification was based on eight compensatory mutations that preserve the predicted secondary structure of putative anticodon and acceptor stems in the CPIR-1 sequence and due to remarkable overall resemblance of multiple tRNA features including: conserved nucleotides, base-pairs, sizes of stems and loops. We conjecture that this CPIR-1 sequence was originally a functional tRNA molecule that was located inside an intron. This location gave the tRNA an additional role in the regulation of the host gene. Eventually, the tRNA became inactive due to mutations in the anticodon loop, yet the other function related to its intronic position has been preserved much longer.

Unfortunately, in our first experimental attempt, we were unable to prove that CPIR-1 is non-coding RNA. Interestingly, Wilusz, Sunwoo, and Spector have described MALAT-1, a long ncRNA (~ 7kb) of mammals, which includes 61-nt long tRNA-like small RNA, named mascRNA [23]. The authors demonstrated that mascRNA is produced via processing of the MALAT-1 nascent transcript and is rapidly degraded. Further experimental investigation of possible expression of CPIR-1 may be complicated and is beyond the scope of this manuscript.

In mammals, a fraction of tRNA genes might be present inside spliceosomal introns. For humans it happens with 3.8% of tRNAs. Usually, this tRNA-intron association is not
evolutionarily maintained. However, there is an exception with four Gly-tRNAs inside introns of *Vac14* gene that have been present there for at least 500 million years. *Vac14* encodes a major scaffold protein within a complex that regulates phosphatidyl inositol 3,5-biphosphate [PtdIns(3,5)P₂] levels across the animal kingdom[24, 25]. Recently discovered, PtdIns(3,5)P₂ is a ubiquitous eukaryote phosphor inositol of very low abundance. PtdIns(3,5)P₂ has been proposed to have at least five independent functions including: recruitment of cytosolic proteins to define organelle specificity; functional regulation of endolysosomal membrane proteins; determination of physical properties and fusogenic potential of endolysosomal membranes; serving as precursor for PI(3)P or PI(5)P; and modulation of endolysosomal pH[26]. In mammals, PtdIns(3,5)P₂ supports and probably regulates the retrograde membrane trafficking from lysosomal and late endosomal compartments to the Golgi complex. PtdIns(3,5)P₂ controls vesicle formation through highly specific activation of vesicle specific Ca²⁺ channels [27]. *Vac14* is essential for mouse nervous system development [28]. Therefore, the exceptionally long-term association of Gly-tRNAs with *Vac14* introns suggests that this tRNA-intron association might be fruitful and provide an additional source of regulation for the biologically important PtdIns(3,5)P₂ pathway.

### 3.5 Conclusions

We have shown that number of conserved regions in plant introns is fewer by hundreds of times than in animal introns. A majority of plant intronic conserved regions with
characterized functions represent alternatively-spliced sequences. We registered only a handful of conserved intronic ncRNAs of flowering plants.

### 3.6 Methods

#### 3.6.1 Databases of plant orthologous introns

Genomic Exon-Intron Databases (EID) of five plant species were generated with our EID pipeline of programs described in Shepelev & Fedorov [29]. These plant EIDs are available from our web page (http://bpg.utoledo.edu/~afedorov/lab/eid.html). Using these exon-intron databases we generated a database of orthologous introns of five plant species exploiting the same algorithms and pipelines of programs as described previously for mammalian orthologous introns [4]. All these programs are available from our web page (http://bpg.utoledo.edu/~afedorov/lab/prog.html). The corresponding protocol for the programs’ execution is presented in the Supplementary Figure S7. Computation of plant intronic sequences for five species (arabidopsis, poplar, grape, rice, and moss) produced 15,726 orthologous intron groups, while four species of flowering plants (arabidopsis, poplar, grape, and rice) – resulted in 28,519 groups. When the computation was confined to three species of dicotyledonous plants (arabidopsis, poplar, and grape), 40,512 groups were obtained. Each group of orthologous introns was aligned with a *Mafft* program [10] and the entire set of multiple alignments is available from our EID web site (http://bpg.utoledo.edu/~afedorov/lab/eid_plantOl.html). This web site contains 3pln_mafft.gz (25.8 MB file for Grape_Poplar_Arabidopsisalignments); 4pln_mafft.gz
(22.7 MB file for Grape_Poplar_Arabidopsis_Ricealignments); and 5pln_mafft.gz (14.5 MB file for Grape_Poplar_Arabidopsis_Rice_Moss).

### 3.6.2 Computational analysis of RNA secondary structures

CPIR-1 (the most evolutionarily Conserved Plant IntrinsicRegion) sequences for each of the five plant species were examined through online RNAfold program from the Vienna RNA web server (http://rna.tbi.univie.ac.at/)[30]. The program predicts the minimum free energy of the secondary structure using dynamic programming algorithm of Zuker and Stiegler [31] and also calculates the equilibrium base pairing probabilities via John McCaskill's partition function algorithm [32]. The output consists of a graphical display of secondary structure, colored by base pairing probabilities.

For studying evolutionary conservation of RNA secondary structures of plant CPIR-1 sequences we used the RNAalifold online program, also available at the Vienna RNA web server, which predicts the consensus structure of a set of aligned RNA sequences [33]. It averages energy contributions from all sequences in addition to the dynamic programming algorithm. The input was CLUSTALW multiple alignment of CPIR-1 sequences of five plant species. The consensus secondary structure of aligned CPIR-1 sequence is shown in Figure 3G.

The JAR3D server (http://rna.bgsu.edu/jar3d) was used to identify possible recurrent 3D structural motifs from the sequences of conserved internal and hairpin loops in predicted secondary structures (Zirbel CL, Petrov AI, Roll J, Leontis NB, In preparation). JAR3D is a novel toolkit developed from WebFR3D server (http://rna.bgsu.edu/FR3D) [34].
Possible tRNA structures were examined using online tRNA\textit{scan-SE} web server under default parameters (http://lowelab.ucsc.edu/tRNA\textit{scan-SE}/)[35].

### 3.6.3 Blast analysis of conserved plant intronic regions

Evolutionarily conserved intronic sequences were examined for their possible involvement in alternative splicing events or being non-coding RNAs using online BLASTN searching of their presence inside RNA databases. Specifically, intronic sequences were searched against the expressed sequence tags (ESTs, excluding human and mouse sequences) database at NCBI using the standard nucleotide BLASTN program [11] with filters turned off. The EST sequences with reliable matches (blast e-value < 0.001) against the conserved portion of intronic sequence were further analyzed for the presence of neighboring exons from the same gene. If the corresponding ESTs showed a fair representation of at least two properly spliced exons, and the borders of conserved portion of the intronic sequence also possessed consensus of splicing junctions, they were regarded as an alternatively spliced exon. In addition, the conserved intronic sequences were searched against the functional RNA database (fRNAdb) (http://www.ncrna.org/frnadb/) [36] and non-coding RNA database (NONCODE) (http://www.noncode.org/NONCODERv3/)[37] to decipher possible functional RNAs.

### 3.6.4 tRNA phylogeny

The entire pool of Arabidopsis tRNA sequences (GtRNA\textit{db-all-tRNAs.fa.gz}) were downloaded from the http://lowelab.ucsc.edu/GtRNA\textit{db/download.html} that comprised a
total of 639 sequences[38]. Using perl, these sequences were classified into groups of tRNAs bearing the same codon specificity that gave 47 different groups. tRNAs with highest cove-scores [39] (A higher cove-score implying true characterization of a tRNA) from each group were extracted and together with CPIR-1 sequence, were aligned using ClustalOmega [40, 41] web-resource. ClustalW2-phylogeny [41, 42] was used for phylogenetic computations of the resulting alignment, using Neighbor joining clustering algorithm [43]. The obtained sequence phylogeny with explanations is presented in the Figure S6.

3.6.5 Mapping gene functions from Gene Ontology database

The gene association files for Arabidopsis were downloaded from the Gene Ontology (GO) database (http://www.geneontology.org/GO.downloads.annotations.shtml) [44]. The file “gene_association.tair.gz“ provided a comprehensive source for Arabidopsis thaliana GO annotations composed of gene associations made by The Arabidopsis Information Resource (TAIR) and The Institute for Genomic Research (TIGR) [45]. The GO functions for our set of 71 Arabidopsis genes were mapped and further identified for functional enrichment. Statistical evaluation was performed using a Monte-Carlo simulation with in-house Perl scripts. For this purpose, 1000 random samples of 71 Arabidopsis genes (the same number as our gene samples containing intronic conserved regions) were created. For each random gene sample GO functions were subsequently analyzed and compared with the real set of GO functions for our 71 genes hosting intronic conserved regions.
3.6.6 Computation of intronic tRNAs in humans and mouse

Genbank feature-tables of each human chromosome (Build 37.3) were downloaded from NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/ARCHIVE/BUILD.37.3/) onto the local Linux workstation. In-house Perl programs encompassing coordinates of tRNAs and introns from the feature-tables were created. A total of 53 cases of tRNAs inside human introns and exons were characterized. Then, cases with tRNAs inside pseudogenes and non-protein coding genes were discarded, resulting in a total of 24 tRNAs inside 14 experimentally confirmed protein-coding genes (Table S2 or Appendix B.2). For each gene, the corresponding ortholog in mouse was identified by reciprocal best-match of protein sequences. Orthologous introns for these 14 human-mouse gene pairs were characterized and the presence of human tRNAs were checked in orthologous introns of mouse using BLASTN program with default parameters with all filters on repetitive sequences turn off. This search revealed that identified human tRNAs were absent in orthologous mouse introns except for 4 glycine tRNAs, all of which were present inside the gene VAC14. The secondary structures of those tRNAs were obtained from tRNAscan-SE (http://lowelab.ucsc.edu/tRNAscan-SE/) [35] and one of them is shown in Figure 3F.

3.6.7 Plant growth conditions and RNA analysis

Arabidopsis thaliana Col-0 ecotype plants were grown hydroponically as described in [46]. Plants grown under these conditions were labeled as controls and were harvested in the middle of the 16 hr light period of the light/dark cycle. Some plants were placed in
the dark for 12 hr instead of 8 hr, were then harvested and labeled as Dark-treated plants. Some plants were placed in a cold room (4° C) for 8 hr under 100 μE of light, whereas others were placed in the cold for 8 hr under the given light conditions and then moved back to the normal growth chamber for 24 hr to recover from the cold stress. The above ground tissue from three plants for each treatment were individually harvested and RNA was isolated from each individual plant using the QiagenRNeasy Plant Mini Kit. RNA quality was assessed by formaldehyde gel electrophoresis and quantified by Nanodrop spectrophotometry. A total of 5 μg of RNA was then loaded in a well of a formaldehyde 2% agarose gel.

To examine the expression of the putative CPIR-1 RNA in various tissues, plants were grown as above under control conditions and harvested at 35 days. By this time the plants had flowered. Plants were separated into rosette leaves, inflorescence, and roots. The tissues from 6 plants each was pooled together and RNA was isolated as above.

3.6.8 Northern blot experiments

A probe for use in Northern blots against a tRNA-like ncRNA in *Arabidopsis thaliana* gene AT5G13240 was developed by designing PCR primers that flank the targeted region using Oligo primer analysis software v. 6.71 (Molecular Biology Insights, Inc., Cascade CO). The following primers were selected: forward- 5’ GAAAGGCTTTTGATCTACT TG 3’, reverse- 5’ TGTCCCAGCTTTTCCCTCCG AG 3’. Primer sequences were checked for specificity using NCBI/Primer-BLAST against the reference assembly for *Arabidopsis thaliana*. Primers were validated by PCR using genomic DNA from *Arabidopsis thaliana* ecotype Col-0. PCR conditions were as follows: 10μl reactions
containing 50 ng of gDNA, 0.05 μg of each primer, 1 X PCR buffer containing 3 mM MgCl₂ (cat# 1778, Idaho Technology, Inc., Salt Lake City, UT), 0.2 mM dNTPs, and 0.5 units Taq polymerase (Promega, Madison, WI) cycled 35 times in Rapidcycler 2 thermal cycler (Idaho Technology), 5 sec. at 94°C, 10 sec. at 58°C, 15 sec. at 72°C, slope=9.9. A single product of the expected size was detected by electrophoresis using the DNA 1000 assay on the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA) and the control reaction (containing no template) generated no PCR products, so primers and gDNA were sent to Lofstrand Labs (Gaithersburg MD) for probe creation and Northern blotting.

Total RNA extracts were sent in dry ice to Lofstrand Labs company (Gaithersburg, Maryland, USA) to perform Northern blot experiments. The protocol for this procedure and experimental details are attached in the Supplementary Figure S8.

3.7 Competing interests

The authors declare that they have no competing interests.

3.8 Authors’ contributions

EA, LW, ASM, and AP performed computational experiments and analyzed the data. SK, EC, and SL performed wet experiments with Arabidopsis RNA and DNA. CLZ performed 2D and 3D computational analysis of putative ncRNAs. AF and LF designed and supervised the study and paper writing.
3.9 Acknowledgments

This work was supported by National Science Foundation Career award “Investigation of intron cellular roles” (grant number MCB-0643542) to A.F.; by USDA-ARS Specific Cooperative Agreement: 58-3607-1-193 to S.L.; and National Institutes of Health (grant number 1R01GM085328-01A1) to C.L.Z.

3.10 Figures
Figure 3-1. **Multiple sequence alignment of plant orthologous introns that have long evolutionarily conserved regions.** Vv stands for *Vitus vinifera*; At – *Arabidopsis thaliana*; Pt – *Populus trichocarpa*. Identical nucleotides in the alignment are marked by stars beneath them. **A.** Sequences represent intron #6 of grape (exon/intron database id: 5548_NW_002238224); intron #6 of *Arabidopsis* (250A_NC_003070); and intron #7 of poplar (4186_NC_008468). Two regions highlighted in yellow correspond to two optionally-skipped exons present inside the reference sequence NM_001197976.1 of grape.
Arabidopsis and EST sequence EV021836.1 of Brassica napus. The first optional exon of Brassica napus contains a frameshift compared to the Arabidopsis counterpart (insertion of “A” at position 180 of Arabidopsis intron). Additionally, the first optional intron (shown in red font) is absent in the following ESTs: (EY747583.1, orange; JG606165.1, bleeding heart; EY779942.1, mandarin; EL387719.1 safflower; CK253964.1 potato; GW451284.1 coffee). Two optional exons and the intron between them are present as a single optional exon in bean (EG698013.1). Multiple indels inside optional exons result in frameshifts in different species and indicate that these alternative exons serve as nonsense-mediated decay (NMD) signals [20].

B. Sequences represent intron #4 of grape (exon/intron database id: 11134_NW_002238109); intron #4 of Arabidopsis (8662D_NC_003074); and intron #3 of poplar (1455_NC_008467). Regions highlighted in yellow correspond to exonic sequences in a highly homologous (85% nucleotide identity) gene DEAD-box ATP-dependent RNA helicase 46 (NM_121465.1). No EST clones that have highlighted optional exons have been discovered.
Figure 3-2. Multiple alignment of plant orthologous intron sequences within an RNA-polymerase III inhibitor gene containing the most evolutionarily conserved sequence (CPIR-1). Vv stands for Vitus vinifera (intron EID identifier: INTRON_6-640_NW_002238155); At – Arabidopsis thaliana (INTRON_6-16795_NC_003076); Pp – Physcomitrella patens (INTRON_5-5711_NW_001865287); Pt – Populus trichocarpa (INTRON_6-13465_NC_008476); Os – Oryza sativa (INTRON_6-9449_NC_008397).

Two conserved regions are highlighted in blue and yellow. Three stable stem-loop structures inside the yellow region are shown in magenta, green, and red fonts. These regions match the D-loop, anticodon loop, and T-loop spatial motifs of tRNAs, respectively.
Figure 3-3. The predicted structures of the evolutionarily conserved yellow region of the CPIR-1 sequence given in figure 3-2 for the five different plant species: A) Vitus vinifera, B) Arabidopsis thaliana, C) Physcomitrella patens, D) Populustrichocarpa, and E) Oryza sativa. Structures A-E have been obtained from RNAfold server.
(http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi) that scores RNA structures based on their thermodynamic stability values ($\Delta G$) and colors them according to their base pairing probability, the scale of which varies from 0 to 1 shown in top left of figure A. The structures in A-E figures have tRNA-like features with respect to conserved loops that have been labeled with ~. Figure F represents 2D structure of the human Gly-tRNA from intron 1 of VAC14 gene. This structure has been generated by tRNAscan-SE online program (http://lowelab.ucsc.edu/tRNAscan-SE/). A cove score (produced by tRNAscanSE) greater than 20 is assigned to a sequence that the program predicts to be a real tRNA. The labels and arrows represent regions that are consonant with the features of real tRNAs described in the Results section. G represents a consensus of secondary structure of tRNA-like motif of the five plant sequences (A-E) obtained by RNAalifold program with default parameters. Characteristics common to real tRNAs are shown by arrows.
Figure 3-4. Gel blot analysis of Arabidopsis thaliana total RNA from various plant organs A, or under different growth conditions B, probed with the exon6-intron7-exon7 region from the RNA polymerase III inhibitor gene. Size in kb of the markers is indicated on the left. Inflo, inflorescence tissue; Lf, rosette leaf tissue; Rt, root tissue; S or H, Invitrogen 0.5-10-Kb RNA ladder; L, NEB Low Rangeess RNA ladder; 0, no RNA control. Numbers for A indicate the plant pool of tissue from which the RNA was derived, organs with the same numbers are from the same pool of plants. Numbers for B indicate individual plants for each treatment. Treatments [e.g. cold, dark, and cold + recov. (cold treatment followed by a 24 hr recovery period under control conditions)] are described in the Materials and Methods. Arrow indicates size of normal transcript, dashed box indicates where the tRNA-like structure should be located if abundant.
3.11 References


3.12 Additional files

**Additional file 1** Supplementary table S1 (provided in B.1)

**Additional file 2** Supplementary table S2 (provided in B.2)

**Additional file 3** Supplementary figure S1 (will be available upon publication)

**Additional file 4** Supplementary figure S2 (will be available upon publication)

**Additional file 5** Supplementary figure S3 (will be available upon publication)

**Additional file 6** Supplementary figure S4 (will be available upon publication)

**Additional file 7** Supplementary figure S5 (will be available upon publication)

**Additional file 8** Supplementary figure S6 (provided in B.3)

**Additional file 9** Supplementary figure S7 (will be available upon publication)

**Additional file 10** Supplementary figure S8 (will be available upon publication)
APPENDIX A

This Appendix contains supplementary information from Chapter 1:

What do 1000 genomes tell us about Biased Gene Conversion Theory

A.I. Annotated genotypes for population samples from 1000 genomes consortium mapped corresponding to chromosomal region of HMOX1 gene. The rows represent individuals and columns denote the genotype data corresponding to a particular SNP position, exemplified in values of 2, 1 and 0. These values denote homozygous alleles, heterozygous alleles and homozygous presence of reference alleles respectively. Columns with the same color illustrate the presence of SNPs in groups of haplotypes. Only a few of the population samples out of a total of 1092 are displayed. The entire table is available from bpg.utoledo.edu
A.2. Annotated genotypes for population samples from 1000 genomes consortium mapped corresponding to chromosomal region of HMOX2 gene. The rows represent individuals and columns denote the genotype data corresponding to a particular SNP position, exemplified in values of 2, 1 and 0. These values denote homozygous alleles, heterozygous alleles and homozygous presence of reference alleles respectively. Columns with the same color illustrate the presence of SNPs in groups of haplotypes. Only a few of the population samples out of a total of 1092 are displayed. The entire table is available from bpg.utoledo.edu.
A.3. Annotated genotypes for population samples from 1000 genomes consortium mapped corresponding to chromosomal region of AGT gene. The rows represent individuals and columns denote the genotype data corresponding to a particular SNP position, exemplified in values of 2, 1 and 0. These values denote homozygous alleles, heterozygous alleles and homozygous presence of reference alleles respectively. Columns with the same color illustrate the presence of SNPs in groups of haplotypes. Only a few of the population samples out of a total of 1092 are displayed. The entire table is available from bpg.utoledo.edu

A.4. Annotated genotypes for population samples from 1000 genomes consortium mapped corresponding to chromosomal region of ZMAT5 gene. The rows represent individuals and columns denote the genotype data corresponding to a particular SNP position, exemplified in values of 2, 1 and 0. These values denote homozygous alleles, heterozygous alleles and homozygous presence of reference alleles respectively. Columns with the same color illustrate the presence of SNPs in groups of haplotypes. Only a few of the population samples out of a total of 1092 are displayed. The entire table is available from bpg.utoledo.edu
A.5. Pairwise correlation coefficients of linkage disequilibrium obtained from formula 1 in chapter 1 and computed for the most frequently occurring SNPs in *HMOX1* gene. The SNPs are displayed across rows as well as columns. Colors are indicative of SNPs belonging to the same haplotype. The black diagonal line represents the correlation of a SNP with itself which ought to be perfectly one in magnitude. The figure is displayed in short scale. The full table is available from bpg.utoledo.edu

<table>
<thead>
<tr>
<th>CORR</th>
<th>35777618</th>
<th>35778505</th>
<th>35785666</th>
<th>35781207</th>
<th>35784317</th>
<th>35786226</th>
<th>35786706</th>
<th>35777856</th>
</tr>
</thead>
<tbody>
<tr>
<td>35777618</td>
<td>1</td>
<td>0.9507</td>
<td>0.8724</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0054</td>
</tr>
<tr>
<td>35778505</td>
<td>0.9507</td>
<td>1</td>
<td>0.8424</td>
<td>0.0003</td>
<td>0.0003</td>
<td>0.0005</td>
<td>0.0005</td>
<td>0.0038</td>
</tr>
<tr>
<td>35785666</td>
<td>0.8724</td>
<td>0.8424</td>
<td>1</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.0002</td>
<td>0.0002</td>
<td>0.0052</td>
</tr>
<tr>
<td>35781207</td>
<td>0</td>
<td>0</td>
<td>0.0003</td>
<td>1</td>
<td>0.9956</td>
<td>0.976</td>
<td>0.976</td>
<td>0.0064</td>
</tr>
<tr>
<td>35784317</td>
<td>0</td>
<td>0</td>
<td>0.0003</td>
<td>0.9956</td>
<td>1</td>
<td>0.976</td>
<td>0.976</td>
<td>0.0065</td>
</tr>
<tr>
<td>35786226</td>
<td>0</td>
<td>0</td>
<td>0.0005</td>
<td>0.976</td>
<td>0.976</td>
<td>1</td>
<td>1</td>
<td>0.0061</td>
</tr>
<tr>
<td>35786706</td>
<td>0</td>
<td>0</td>
<td>0.0005</td>
<td>0.976</td>
<td>0.976</td>
<td>1</td>
<td>1</td>
<td>0.0061</td>
</tr>
<tr>
<td>35777856</td>
<td>0.0054</td>
<td>0.0038</td>
<td>0.0052</td>
<td>0.0064</td>
<td>0.0065</td>
<td>0.0061</td>
<td>1</td>
<td>0.9929</td>
</tr>
<tr>
<td>35778278</td>
<td>0.0054</td>
<td>0.0038</td>
<td>0.0052</td>
<td>0.0066</td>
<td>0.0067</td>
<td>0.0063</td>
<td>0.0063</td>
<td>0.9929</td>
</tr>
</tbody>
</table>
A.6. Pairwise correlation coefficients of linkage disequilibrium obtained from formula 1 in chapter 1 and computed for the most frequently occurring SNPs in HMOX2 gene. The SNPs are displayed across rows as well as columns. Colors are indicative of SNPs belonging to the same haplotype. The black diagonal line represents the correlation of a SNP with itself which ought to be perfectly one in magnitude. The figure is displayed in short scale. The full table is available from bpg.utoledo.edu

<table>
<thead>
<tr>
<th>CORR</th>
<th>4528634(T-C)</th>
<th>4530862(T-C)</th>
<th>4530919(C-T)</th>
<th>4536165(A-C)</th>
<th>4537104(T-C)</th>
<th>4537893(G-A)</th>
<th>4538114(G-A)</th>
<th>4536555(C-T)</th>
<th>4536634(A-C)</th>
<th>4537808(C-T)</th>
<th>4538008(T-C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4528634(T-C)</td>
<td>0.8471</td>
<td>0.8455</td>
<td>0.0047</td>
<td>0.0039</td>
<td>0.0036</td>
<td>0.0036</td>
<td>0.0077</td>
<td>0.0077</td>
<td>0.0084</td>
<td>0.0088</td>
<td>0.0077</td>
</tr>
<tr>
<td>4530862(T-C)</td>
<td>0.8471</td>
<td>1</td>
<td>0.0077</td>
<td>0.0077</td>
<td>0.0084</td>
<td>0.0088</td>
<td>0.009</td>
<td>0.009</td>
<td>0.009</td>
<td>0.009</td>
<td>0.009</td>
</tr>
<tr>
<td>4530919(C-T)</td>
<td>0.8455</td>
<td>0.8729</td>
<td>0.0057</td>
<td>0.0051</td>
<td>0.0045</td>
<td>0.0045</td>
<td>0.0072</td>
<td>0.0072</td>
<td>0.009</td>
<td>0.009</td>
<td>0.009</td>
</tr>
<tr>
<td>4536165(A-C)</td>
<td>0.0047</td>
<td>0.0057</td>
<td>0.0034</td>
<td>1</td>
<td>0.9853</td>
<td>0.9298</td>
<td>0.938</td>
<td>0.0477</td>
<td>0.0477</td>
<td>0.0405</td>
<td>0.0398</td>
</tr>
<tr>
<td>4537104(T-C)</td>
<td>0.0039</td>
<td>0.0051</td>
<td>0.0029</td>
<td>0.9853</td>
<td>1</td>
<td>0.9266</td>
<td>0.9524</td>
<td>0.0464</td>
<td>0.0464</td>
<td>0.0391</td>
<td>0.0384</td>
</tr>
<tr>
<td>4537893(G-A)</td>
<td>0.0036</td>
<td>0.0045</td>
<td>0.0025</td>
<td>0.9298</td>
<td>0.9266</td>
<td>1</td>
<td>0.9371</td>
<td>0.0413</td>
<td>0.0413</td>
<td>0.0347</td>
<td>0.0347</td>
</tr>
<tr>
<td>4538114(G-A)</td>
<td>0.0036</td>
<td>0.0045</td>
<td>0.0024</td>
<td>0.938</td>
<td>0.9524</td>
<td>0.9371</td>
<td>1</td>
<td>0.0422</td>
<td>0.0422</td>
<td>0.0351</td>
<td>0.0344</td>
</tr>
<tr>
<td>4536555(C-T)</td>
<td>0.0077</td>
<td>0.0072</td>
<td>0.0077</td>
<td>0.0477</td>
<td>0.0464</td>
<td>0.0413</td>
<td>0.0422</td>
<td>1</td>
<td>1</td>
<td>0.8664</td>
<td>0.8797</td>
</tr>
<tr>
<td>4536634(A-C)</td>
<td>0.0077</td>
<td>0.0072</td>
<td>0.0077</td>
<td>0.0477</td>
<td>0.0464</td>
<td>0.0413</td>
<td>0.0422</td>
<td>1</td>
<td>1</td>
<td>0.8664</td>
<td>0.8797</td>
</tr>
<tr>
<td>4537808(C-T)</td>
<td>0.0084</td>
<td>0.009</td>
<td>0.0096</td>
<td>0.0405</td>
<td>0.0391</td>
<td>0.0347</td>
<td>0.0351</td>
<td>0.8664</td>
<td>0.8664</td>
<td>1</td>
<td>0.985</td>
</tr>
<tr>
<td>4538008(T-C)</td>
<td>0.0088</td>
<td>0.0087</td>
<td>0.0093</td>
<td>0.0398</td>
<td>0.0384</td>
<td>0.034</td>
<td>0.0344</td>
<td>0.8797</td>
<td>0.8797</td>
<td>0.985</td>
<td>1</td>
</tr>
</tbody>
</table>
A.7. Pairwise correlation coefficients of linkage disequilibrium obtained from formula 1 in chapter 1 and computed for the most frequently occurring SNPs in AGT gene. The SNPs are displayed across rows as well as columns. Colors are indicative of SNPs belonging to the same haplotype. The black diagonal line represents the correlation of a SNP with itself which ought to be perfectly one in magnitude. The figure is displayed in short scale. The full table is available from bpg.utoledo.edu
A.8. Pairwise correlation coefficients of linkage disequilibrium obtained from formula 1 in chapter 1 and computed for the most frequently occurring SNPs in ZMAT5 gene. The SNPs are displayed across rows as well as columns. Colors are indicative of SNPs belonging to the same haplotype. The black diagonal line represents the correlation of a SNP with itself which ought to be perfectly one in magnitude. The figure is displayed in short scale. The full table is available from bpg.utoledo.edu

<table>
<thead>
<tr>
<th>CORR</th>
<th>30157589(T-C)</th>
<th>30158463(C-T)</th>
<th>30130115(A-T)</th>
<th>30130776(T-C)</th>
<th>30134763(C-T)</th>
<th>30137262(G-A)</th>
<th>30138548(T-C)</th>
<th>30153721(T-C)</th>
<th>30154101(C-A)</th>
<th>30155251(G-C)</th>
<th>30149144(C-T)</th>
<th>30149961(C-T)</th>
</tr>
</thead>
<tbody>
<tr>
<td>30157589(T-C)</td>
<td>1</td>
<td>0.9816</td>
<td>0.2526</td>
<td>0.279</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>30158463(C-T)</td>
<td>0.9816</td>
<td>1</td>
<td>0.2526</td>
<td>0.279</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0006</td>
</tr>
<tr>
<td>30130115(A-T)</td>
<td>0.2526</td>
<td>0.2526</td>
<td>1</td>
<td>0.8978</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0004</td>
<td>0.0004</td>
<td>0.0005</td>
<td>0.0007</td>
</tr>
<tr>
<td>30130776(T-C)</td>
<td>0.279</td>
<td>0.279</td>
<td>0.8978</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td>30134763(C-T)</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.9909</td>
<td>0.9746</td>
<td>0.7938</td>
<td>0.7921</td>
<td>0.7868</td>
<td>0.0132</td>
<td>0.0132</td>
<td>0.0133</td>
<td>0.0133</td>
</tr>
<tr>
<td>30137262(G-A)</td>
<td>0.0001</td>
<td>0</td>
<td>0</td>
<td>0.9909</td>
<td>1</td>
<td>0.9801</td>
<td>0.7986</td>
<td>0.7968</td>
<td>0.7915</td>
<td>0.0133</td>
<td>0.0133</td>
<td>0.0133</td>
</tr>
<tr>
<td>30138548(T-C)</td>
<td>0.0001</td>
<td>0</td>
<td>0</td>
<td>0.9746</td>
<td>0.9801</td>
<td>1</td>
<td>0.7921</td>
<td>0.7903</td>
<td>0.7883</td>
<td>0.014</td>
<td>0.014</td>
<td>0.014</td>
</tr>
<tr>
<td>30153721(T-C)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.7938</td>
<td>0.7986</td>
<td>0.7921</td>
<td>1</td>
<td>0.9982</td>
<td>0.9926</td>
<td>0.0079</td>
<td>0.0079</td>
<td>0.0079</td>
</tr>
<tr>
<td>30154101(C-A)</td>
<td>0.0001</td>
<td>0</td>
<td>0</td>
<td>0.7921</td>
<td>0.7968</td>
<td>0.7903</td>
<td>0.9982</td>
<td>1</td>
<td>0.9945</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
</tr>
<tr>
<td>30155251(G-C)</td>
<td>0.0001</td>
<td>0</td>
<td>0</td>
<td>0.7868</td>
<td>0.7915</td>
<td>0.7883</td>
<td>0.9926</td>
<td>0.9945</td>
<td>1</td>
<td>0.007</td>
<td>0.007</td>
<td>0.007</td>
</tr>
<tr>
<td>30149144(C-T)</td>
<td>0.0003</td>
<td>0</td>
<td>0</td>
<td>0.0132</td>
<td>0.0133</td>
<td>0.014</td>
<td>0.0079</td>
<td>0.008</td>
<td>0.007</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>30149961(C-T)</td>
<td>0.0003</td>
<td>0</td>
<td>0</td>
<td>0.0132</td>
<td>0.0133</td>
<td>0.014</td>
<td>0.0079</td>
<td>0.008</td>
<td>0.007</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

110
A.13. Complex cases of recombinations from population samples of 1000 genomes corresponding to the genomic locus of AGT gene. The genotypes of corresponding individuals are unlikely to be explained by 1-5 kb of heteroduplex crossovers or simple recombination events between haplotypes.
A.15. Supplementary Explanations for Figures 2-5 indicating formations of putative heteroduplexes within the four genes HMOX1, HMOX2, AGT and ZMAT5 respectively are provided below. The explanations are provided on a case by case basis.

**Explanations for Figure 2 (HMOX1)**

**Case 1 (HG00593).** Insertion of fifth red SNP and reverse mutations of seventh and eighth blue SNPs within the allele belonging to blue haplotype. The eighth blue SNP has not been considered on account of its likelihood (-0.04) below our threshold

**Case 2 (HG01494).** Insertion of second red SNP and reverse mutations of fifth and sixth green SNPs within the allele belonging to green haplotype. Only the fifth green reverse mutation had a threshold above our cutoff likelihood and was consequently labeled.

**Case 3 (NA11930).** Insertion of ninth blue SNP within the red haplotype allele

**Case 4 (NA11994).** Insertion of third green SNP into either of the alleles belonging to red haplotype

**Case 5 (NA19355).** Insertion of a frequently occurring mutation associated with blue haplotype (but not a member of the blue haplogroup) into one of the alleles belonging to red haplotype

**Case 6 (NA12761).** Insertion of fifth red SNP and reverse mutations of fifth, sixth and seventh blue SNPs in a particular region and reverse mutation of the eleventh blue SNP downstream of the gene which is not considered since it is a base addition (C → CA). Of all the mutations/ reverse mutations in the region, only absence of fifth and seventh blue SNPs had likelihood values greater than our threshold.
Explanations for Figure 3 (HMOX2)

Case 1 (HG00129). Insertion of sixteenth green SNP into either of the two alleles belonging to red haplotype.

Case 2 (HG00130). Insertion of third red SNP into either of the two alleles belonging to green haplotype. In one of the alleles, the fifth and sixth green SNPs are missing probably due to extended region of heteroduplex formed by crossover of red allele. However, they are not considered due to poor likelihoods.

Case 3 (HG00134). Insertion of sixteenth green SNP into blue haplotype allele. There is also a presence of the third red SNP in either of the two alleles which does not cross our threshold.

Case 4 (HG00355). Insertion of first red SNP into blue haplotype allele. There are a couple of reverse mutations within the blue haplotype allele that cannot be explained by the length of span of a possible heteroduplex and are hence left out. The insertion of sixteenth green SNP into the blue haplotype has been discussed previously and is hence left out.

Case 5 (HG01149). Insertion of 32nd and 33rd blue SNPs into the red haplotype allele. The possibility of the SNPs within the green haplotype allele can be ruled out because of numerous observations (from 1000 genomes genotype data) of insertion of the last two blue SNPs within red haplotype. The last (33rd) blue SNP is a nucleotide insertion (A->AG) which does not fall in the context of our investigations.

Case 6 (HG01259). Insertion of second red SNP within green haplotype allele.

Case 7 (NA18505). Insertion of three green SNPs (13th, 14th and 15th) into red haplotype. The green allele shows insertion of the third red SNP which has been covered in case 2.
Case 8 (NA18516). Insertion of second red SNP into blue haplotype allele

Case 9 (NA18952). Insertion of second and third red SNPs within the green haplotype allele. The span of heteroduplex so formed also covers the fourth green SNP which gets deleted due to insertion of red haplotype allele at that position. The second red SNP in this case has a lower likelihood value and hence is not considered further.

Case 10 (NA19096). Insertion of 24th and 25th blue SNPs within the allele belonging to red haplotype. During the heteroduplex formation, the insertion of corresponding blue haplotype allele also causes deletion of 4th, 5th and 6th red SNPs within red haplotype. For the entire span of the heteroduplex, the 25th blue SNP has a likelihood value (-0.05) lower than our threshold and hence is not considered. Towards the 3’ end of either alleles, there is an insertion of the 21st green SNP that has a poor likelihood (-0.45).

Case 11 (NA19108). Insertion of first red SNP into blue haplotype. The insertion is also accompanied by the absence of the adjacent tenth blue SNP within the blue haplotype allele.

Case 12 (NA19657). Insertion of four green SNPs (5th, 6th, 8th and 9th) within the red haplotype allele. Out of them, only one of the green SNPs (9th) has a likelihood value greater than our threshold and hence is considered further.

Case 13 (NA19672). Insertion of first red SNP into green haplotype allele.

Case 14 (NA20336). Insertion of third blue SNP into the red haplotype allele. The insertion is accompanied by absence of adjacent blue SNPs. This can be explained by the formation of heteroduplex crossovers between red and blue alleles amongst which the third blue SNP was selected by the mismatch repair mechanism and other SNPs among blue weren’t.
Explanations for Figure 4 (AGT)

**Case 1 (HG00406).** Insertion of fifth red SNP into the allele comprising blue + orange haplotype.

**Case 2 (HG00418).** Insertion of fourth red SNP into either of the two alleles belonging to blue + yellow haplotype.

**Case 3 (NA18592).** Insertion of fourth yellow SNP into either of the two alleles belonging to blue + green haplotype or the exclusively green haplotype. The last two SNPs of red haplotype are equally likely to be present in either of the two alleles. Their existence can be explained by simple recombination events and hence not analyzed further.

**Case 4 (NA12006).** Insertion of sixth red SNP within the allele belonging to blue + orange haplotype.

**Case 5 (NA12717).** Insertion of three yellow SNPs (fifth, sixth and seventh) into either of the two alleles belonging to haplotypes red or blue + green

**Case 6 (NA18489).** Insertion of second and fourth orange SNPs within the allele belonging to blue + yellow haplotype. The third orange SNP is not selected by the mismatch repair mechanism acting after local recombination events have taken place and hence denoted by an empty symbol. Also, the group of orange mutations occur exclusively within the blue + yellow haplotype group as this has also been observed in other individuals eg. NA19307, NA19401, NA19445, NA19445, NA19453 among whom, NA19453 is homozygous with the above allele composition. Of the three orange SNPs involved in short-range recombination, the third and fourth had likelihood values greater than the threshold and hence are considered for further analysis.
Case 7 (NA18873). Insertion of third red SNP into either of the two alleles belonging to blue + green haplotype

Case 8 (NA18949). Insertion of first orange SNP into either of the two alleles belonging to blue + yellow haplotype or red haplotype

Case 9 (NA18990). Insertion of first green SNP into the allele belonging to blue + yellow haplotype

Case 10 (NA19028). Insertion of second red SNP into either of the two alleles belonging to yellow + blue haplotype or blue + orange haplotype

Case 11 (NA19074). Insertion of three SNPs of green haplotype (first, second and third) into either of the two alleles belonging to blue + yellow haplotype. The second allele contains the fourth yellow SNP (complete haplogroup) in lieu of the above three green SNPs

Case 12 (NA19445). Insertion of fourth red SNP into either of the two alleles belonging to blue + yellow or, blue + orange haplotypes

Case 13 (NA20520). Insertion of sixth red SNP (and missing of adjacent fifth and seventh red SNPs) within one of the alleles belonging to blue + green haplotype. The SNPs adjacent to the inserted one have low likelihood values and hence are not reported.

**Explanations for Figure 5 (ZMAT5)**

Case 1 (HG00106). Insertion of the last (12th) green SNP within the allele belonging to the orange haplotype. There are several individuals having this SNP within the orange haplotype that rules out the possibility of this SNP within red + blue haplotype allele
Case 2 (HG00240). Insertion of the sixth red SNP into either of the alleles belonging to green + purple haplotype

Case 3 (HG00350). Insertion of eighth green SNP within the allele belonging to red haplotype

Case 4 (HG01066). Insertion of eleventh red SNP within the green + purple haplotype allele

Case 5 (HG00251). Insertion of third red SNP within the green + purple haplotype allele

Case 6 (HG01250). Insertion of ninth red SNP into the allele belonging to green + purple haplotype

Case 7 (NA12874). Insertion of eighth green SNP within either of the alleles belonging to red haplotype

Case 8 (NA18573). Insertion of second blue SNP into either of the two alleles belonging to red haplotype

Case 9 (NA19005). Insertion of fourth and fifth green SNP into the red haplotype allele. As the fourth green SNP has a poor likelihood (-0.1), only the fifth green SNP is taken into consideration and consequently labeled

Case 10 (NA19020). Insertion of sixth green SNP within either of the alleles belonging to red haplotype. One of the red alleles has a reverse mutation which is not considered on account of poor likelihood

Case 11 (NA19319). Insertion of last (twelfth) green SNP into the allele belonging to red haplotype. Several alleles of green haplotype belonging to green + purple haplogroup are missing as shown in empty stars but the span of alleles is longer than what is expected
according to heteroduplex model. Thus those alleles are not considered in the present context.

**Case 12 (NA19383).** Insertion of second green SNP within the red haplotype allele

**Case 13 (NA19712).** While insertion of seventh and ninth red SNPs take place within the green + orange haplotype allele, the eighth red SNP is not selected during the putative heteroduplex formation. This SNP also exhibits poor likelihood and is thus not analyzed further

**Case 14 (NA19046).** Insertion of seventh and fifteenth red SNP within either of the two alleles belonging to green + orange or green + purple haplogroups. Also accompanying these events are absence of twelfth and thirteenth purple SNPs within green + purple haplotype allele, most likely due to a longer span of heteroduplex formation, resulting from insertion of fifteenth red SNP. Only one of the aforementioned SNPs had a strong likelihood that has been considered further

**Case 15 (NA19117).** The second allele is formed from simple recombination events between orange and green + purple haplotype alleles that cause the absence of several green SNPs towards its’ 5’end. The presence of only one green (4th) towards 5’end of the second allele does not allow it to be attributable to a heteroduplex formation. It is hence left out of the present context. The sixth and seventh green SNPs are absent within the allele belonging to green + orange haplotype. The eighth red SNP is present in either of the two alleles. The purple SNPs neighboring/flanking the last green SNP within the second allele are absent. Of all the allelic recombinations and displacement of mutations taking place, only one of the SNPs exhibits a strong likelihood that has been considered further.
APPENDIX B

This Appendix contains supplementary information from Chapter 3: Bioinformatics of orthologous introns, Identification of an intronic tRNA-like sequence.


<table>
<thead>
<tr>
<th>tRNA characteristics</th>
<th>Observations for plant CPIR-1 sequences</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stable 7-bp-long acceptor stem</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>3-4-bp D-loop stem</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td></td>
<td>We observe 5-bp in plants, however, the terminal pair is unstable</td>
<td></td>
</tr>
<tr>
<td>5-bp stem of T-loop</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>5-bp in the anticodon stem</td>
<td>×</td>
<td>Majority of plants have longer stem</td>
</tr>
<tr>
<td>Two nucleotides between amino acid stem and D-stem</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>One nucleotide between D and anticodon stem</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>Statement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------------------------------------------------</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>There is no nucleotide between amino acid stem and T-stem</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>There are always 7 nucleotides in the T-loop</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>There are always 7 nucleotides in the anticodon loop</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>The length of anticodon loop is variable for plant tRNA-like structures</td>
<td></td>
<td></td>
</tr>
<tr>
<td>There are 7-10 nucleotides in the D-loop</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>There are 4-21 nucleotides in the variable arm</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>U lies between amino acid and D-stem</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A lies at the beginning of D-loop</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Two G's are present at the D-loop</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>U is present in the anticodon loop</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>G-C base pairing occurs in the T-stem</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>TUC sequence and A are present in the T-loop</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>
**B.2. Identification of a total of 21 tRNAs inside 14 protein-coding functional genes of the human genome. The numbers of tRNAs found in each gene as well as the sequence is also shown.**

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th># of tRNA, Sequences of tRNA/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>FNBP1L</td>
<td>1 Aagggcaccgaggtttgagaaccgggggctctttgcgtaactaatgcttcaccccticgagctatccccc</td>
</tr>
<tr>
<td>ARPC2</td>
<td>1 ctttcaatagtttcagcgtgtagcagagagactataaggctcccttaggttgtcctgtgtctcagtcgtgaaggg</td>
</tr>
<tr>
<td>PPFIBP1</td>
<td>1 accccgagtagctcagtagtagagcagactagactatatcggggcataagggctagtgtgcctgtctccttttcgggga</td>
</tr>
<tr>
<td>SHF</td>
<td>3 tgcctgtgactgtgaatcagggagggttggtcgtggtgaaccgcataaacgcaggaatcaccatagctagcagtcagtcagtcagggg</td>
</tr>
<tr>
<td></td>
<td>gcctgtgactgtgaatcagggagggttggtcgtggtgaaccgcataaacgcaggaatcaccatagctagcagtcagtcagggg</td>
</tr>
<tr>
<td></td>
<td>gctgtgtgggtagtgaattaccaggtgctgtcttaataacctaccatcagctagcagtcagtcagggg</td>
</tr>
<tr>
<td>VAC14</td>
<td>4 tgcctggccggggaatccacccgggcctcccggcaggcgagaattctaccactgaaccaccaaatgcctgcctaccaagggc</td>
</tr>
<tr>
<td></td>
<td>gctgtgactgtgaatcagggagggttggtcgtggtgaaccgcataaacgcaggaatcaccatagctagcagtcagtcagggg</td>
</tr>
<tr>
<td></td>
<td>gctgtgactgtgaatcagggagggttggtcgtggtgaaccgcataaacgcaggaatcaccatagctagcagtcagtcagggg</td>
</tr>
<tr>
<td>FBXO31</td>
<td>1 tgcctggccggggaatccacccgggcctcccggcaggcgagaattctaccactgaaccaccaaatgcctgcctaccaagggc</td>
</tr>
<tr>
<td>HES7</td>
<td>1 ggcctctggcgaatggtagatgcaagctttcaatttaacatgggttgtgctcagtaatcccacccagagtcg</td>
</tr>
<tr>
<td>CTC1</td>
<td>3 aggccaccggtggattacgagggagctctgtttaactagacaggagcttttaaaccaactaagcgccagggc</td>
</tr>
<tr>
<td></td>
<td>gctgtgactgtgaatcagggagggttggtcgtggtgaaccgcataaacgcaggaatcaccatagctagcagtcagtcagggg</td>
</tr>
<tr>
<td></td>
<td>gctgtgactgtgaatcagggagggttggtcgtggtgaaccgcataaacgcaggaatcaccatagctagcagtcagtcagggg</td>
</tr>
<tr>
<td></td>
<td>tgcctggccggggaatccacccgggcctcccggcaggcgagaattctaccactgaaccaccaaatgcctgcctaccaagggc</td>
</tr>
<tr>
<td></td>
<td>gctgtgactgtgaatcagggagggttggtcgtggtgaaccgcataaacgcaggaatcaccatagctagcagtcagtcagggg</td>
</tr>
<tr>
<td></td>
<td>gctgtgactgtgaatcagggagggttggtcgtggtgaaccgcataaacgcaggaatcaccatagctagcagtcagtcagggg</td>
</tr>
<tr>
<td>ULK2</td>
<td>1 gcatttggtgtgtagttgttgtagtagatgaatattctgcctcacaacgagggagaccaagtttcgcatttcgctgaatggc</td>
</tr>
<tr>
<td>BCAS3</td>
<td>1 gcctggccggggaatccacccgggcctcccggcaggcgagaattctaccactgaaccaccaaatgcctgcctaccaagggc</td>
</tr>
<tr>
<td>CEP95</td>
<td>1 aggccaccggtggattacgagggagctctgtttaactagacaggagcttttaaaccaactaagcgccagggc</td>
</tr>
<tr>
<td>ATP5A1</td>
<td>1 tgcctggccggggaatccacccgggcctcccggcaggcgagaattctaccactgaaccaccaaatgcctgcctaccaagggc</td>
</tr>
<tr>
<td>PARVB</td>
<td>1 gcctggccggggaatccacccgggcctcccggcaggcgagaattctaccactgaaccaccaaatgcctgcctaccaagggc</td>
</tr>
</tbody>
</table>
| HN10        | 1 tgttataagaggggtagattgtaaatataacatgggtgtgctttactatgctattgtcagattgctagttggcactttaacac
B.3. Evolutionary analysis of the conserved CPIR-1 sequence

![CLUSTAL-Omega alignment of Arabidopsis tRNA-like CPIR sequence with best matched Pro-tRNAs from Arabidopsis](image)

**Figure S6.** **A)** Clustal-Omega multiple alignment of best matched proline tRNAs from Arabidopsis. The characteristic features from corresponding secondary structures of proline tRNAs that are common to all tRNAs have been labeled in color **B)** Rectangular Cladogram of CPIR-1 sequence with best scored tRNAs from Arabidopsis that have a unique codon specificity. The cladogram was generated from Clustal W2 phylogeny online tool using neighbor-joining algorithm and subsequently processed in treeview. The red ellipse represents the best matched tRNAs with CPIR-1 that happen to be proline tRNAs.