CONTEXT DEPENDENT MUTATION BIASES IN THE HUMAN GENOME

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

AHMET YETKIN ALICI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
BIOLOGY

AUGUST 2017

Approval of the thesis:

CONTEXT DEPENDENT MUTATION BIASES IN THE HUMAN GENOME

submitted by **AHMET YETKIN ALICI** in partial fulfillment of the requirements for the degree of **Master of Science in Biology Department, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Orhan Adalı _____
Head of Department, **Biology**

Assoc. Prof. Dr. Mehmet Somel _____
Supervisor, **Biology Department, METU**

Examining Committee Members:

Assoc. Prof. Dr. Ayşe Elif Erson Bensan
Biology Department, METU _____

Assoc. Prof. Dr. Mehmet Somel
Biology Department, METU _____

Assoc. Prof. Dr. Tolga Can
Dept. of Computer Engineering, METU _____

Assist. Prof. Dr. Emre Karakoç
Dept. of Computer Engineering, Medipol University _____

Assist. Prof. Dr. Can Alkan
Dept. of Computer Engineering, İ.D. Bilkent University _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: AHMET YETKIN ALICI

Signature :

ABSTRACT

CONTEXT DEPENDENT MUTATION BIASES IN THE HUMAN GENOME

Alıcı, Ahmet Yetkin
M.S., Department of Biology
Supervisor : Assoc. Prof. Dr. Mehmet Somel

August 2017, 69 pages

Different types of mutations occur and spread in the genome at varying rates. For instance, C->T transitions at CpG sites are the most frequent mutation in mammalian genomes. In contrast, GC-biased gene conversion causes A or T->G or C mutations to spread and fix rapidly in populations. Such fixation biases have not yet been investigated taking neighbouring sequence context into account. Using human population genomic data from the 1000 Genomes Project and comparative genomic data from other primates, possible fixation biases in the human genome at the level of quintuplets (5 bp sequences where the middle base is mutated) will be investigated.

Keywords: Mutation Bias, Fixation Bias, Comparative Genomics, Population Genomics

ÖZ

İNSAN GENOMUNDA ÇEVRE NÜKLEOTİTLERE BAĞLI MUTASYON EĞİLİMLERİ

Alıcı, Ahmet Yetkin
Yüksek Lisans, Biyoloji Bölümü
Tez Yöneticisi : Doç. Dr. Mehmet Somel

Ağustos 2017, 69 sayfa

Değişik mutasyon tipleri genomda farklı hızlarda ortaya çıkar ve yayılır. Örneğin CpG bölgelerindeki C->T dönüşümleri, memeli genomlarında en sık rastlanan mutasyon tipidir. Bunun aksine, GC eğilimli gen çevirimleri A ya da T bazlarının G ya da C bazlarına dönüştüğü mutasyonların, populasyonlarda hızla yayılmasına ve sabitlenmesine neden olmaktadır. Bu sabitlenme eğilimleri, şimdiye kadar çevre nükleotit dizilerinin etkisi dikkate alınarak incelenmemiştir. 1000 Genom Projesinden insan populasyon genomiği verileri ve diğer primatların karşılaştırmalı genomik verileri kullanılarak, insan genomundaki olası sabitlenme eğilimleri, ortadaki baz mutasyona uğramış olmak üzere 5erli baz çiftleri düzeyinde incelenecektir.

Anahtar Kelimeler: Mutasyon Eğilimi, Sabitlenme Eğilimi, Karşılaştırmalı Genomik, Popülasyon Genomiği

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

Here, to provide a background to the study, a brief overview of relevant fields is provided. Population genomics, genome evolution and related statistical analysis methods are dynamic fields of research, that bridge gaps between evolutionary biology and molecular biology, as well as evolutionary theory and its applications.

1.1 Population Genomics

Advances in DNA sequencing techniques in the last decade made it possible to generate whole genome sequences in a time- and cost-efficient way (Van Dijk et al., 2014), (Schuster, 2008). Large-scale projects like the 1000 Genomes (Auton et al., 2015), (McVean et al., 2012) undertaken by international consortia have been launched following these advances, leading to the generation of large amounts of genome sequence data from multiple human populations.

Comparative genome studies based on data produced by these projects consider very diverse research questions such as: what are the molecular basis of adaptive features (Huerta-Sánchez et al., 2014), how is the human population structured and how can we infer links between ancestral and present-day populations (Wall and Slatkin, 2012), how are disease phenotypes mapped to genotypes and how are they distributed within the population (Service et al., 2014), etc.

Although the questions are diverse, they are all more or less dependent on population genetic theory and models of molecular evolution, because only with such models

can we interpret DNA sequence data. Studies based on data, on the other hand, test predictions of population genetics theory, prevalence of different population genetic forces in nature, and extend the theory further from its classical domain where the objects of study are well defined, discrete, isolated loci (Casillas and Barbadilla, 2017), (Lynch, 2007).

1.2 Detecting Signals of Natural Selection

Detecting signatures of natural selection is one of the major aims of population genetic studies on genome-wide data.

In molecular evolution, natural selection is broadly classified as positive, negative or balancing selection (Nielsen and Slatkin, 2013). Natural selection that causes the spread of a new mutation through a population is classified as positive selection; and the opposite case, where selection decreases the survival chance of a new mutation, is classified as negative selection. If selection favours a new variant alongside with the pre-existing variant, it is called balancing selection.

The neutral theory of molecular evolution as described by (Kimura, 1991), (Kimura, 1968) is used as a null model of molecular evolution. The general approach to detect natural selection is to compare DNA sequences within and/or between populations/species, and check whether observed changes between sequences deviate from that expected by the neutral theory.

The classical method to detect deviation in coding sequence data from expectation of neutral theory is to check the d_N /d_S ratio (Kimura, 1983). The rate of non-synonymous substitution d_N is defined as the number of amino acid-changing fixed differences in the DNA sequence of interest with respect to some inferred ancestral sequence, divided by the number of positions in that sequence where a mutation would lead to an amino acid change. The rate of synonymous nucleotide substitution d_S is similarly defined as the number of fixed differences that do not change the coded amino acid, divided by number of positions, on which a possible mutation would not

result in an amino acid change.

Synonymous mutations are generally assumed not to be subject to selection because they do not change the phenotype (but see e.g. (Lawrie et al., 2013)), hence the synonymous substitution rate (d_S) is considered as an estimate of the rate of change in the absence of selection. Non-synonymous mutations, on the other hand, can be affected by selection, because they change structure of the coded protein. If the rate of non-synonymous substitutions (d_N) deviates from the expected rate of substitutions in the absence of selection (i.e. synonymous substitution rate, or d_S), such deviance could be attributed to selection. According to this reasoning, $d_N = d_S$ suggests neutral evolution. $d_N > d_S$, i.e. higher than expected rate of change, suggests recurrent positive selection. $d_N < d_S$, i.e. lower than expected rate of change, suggests negative selection.

This method is applicable to protein coding regions only. In addition, it is based on comparison of d_NA sequences between species, hence it may not be able to detect more recent selection events. Furthermore, to be detectable by the d_N / d_S ratio, positive selection has to have accumulated multiple substitutions. If a variant of the sequence with only one or few changed nucleotides is the allele that is selected for, d_N / d_S ratio would be still lower than 1, as the majority of the positions are under negative selection in this case. Only using multiple species alignments and more sophisticated models of selection (e.g. PAML) (Yang, 2007), may one detect selection events in this case.

Methods that use population genetic data overcome some or all of the above mentioned deficiencies.

The McDonald-Kreitman test (McDonald and Kreitman, 1991) is another method for comparing non-synonymous with synonymous mutations, which uses not only fixed mutations but also mutations that are polymorphic within a population. The classical McDonald-Kreitman table is a 2X2 contingency table, where each cell contains mutation counts with the corresponding properties (synonymous / non - synonymous and fixed / polymorphic) from the sequence (locus or loci) of interest.

A simple interpretation of the MK test follows the reasoning that the polymorphic position frequency in a sequence reflects the mutation rate. According to population genetic theory, the fate of new mutations are mostly determined by random events, unless they are strongly deleterious.

Thus, polymorphisms mainly represent the background mutation rate and genetic drift, but not selection (due to the time constraint). Fixations, on the other hand, are mutations that have accumulated in a sequence over larger time scales; their frequencies thus represent both the background mutation rate and genetic drift, but also natural selection.

If the proportion of fixed mutations within non-synonymous ones is higher than the same proportion within synonymous mutations (or, similarly, if the proportion of non-synonymous mutations within fixed ones is higher than this proportion within polymorphic mutations), that is:

$$\frac{F_N}{P_N} > \frac{F_S}{P_S} \text{ (likewise } \frac{F_N}{F_S} > \frac{P_N}{P_S})$$

this indicates positive selection. Here F denotes fixed mutations, P denotes polymorphisms, and subscripts N and S denote non-synonymous and synonymous mutations, respectively. If the ratios have the relation:

$$\frac{F_N}{P_N} < \frac{F_S}{P_S} \text{ (likewise } \frac{F_N}{F_S} < \frac{P_N}{P_S})$$

this can be attributed to negative selection. Meanwhile equal (or almost equal) ratios imply no effect of selection.

The MK test can thus be considered an improvement over the d_N /d_S ratio approach, because the former takes polymorphic mutations into account. This way, a possible bias due to difference in mutation rates between mutation types can be avoided. Moreover, a significance test like Chi-square or Fisher's exact test can be conducted.

Notably, the difference between polymorphism ratios between synonymous and non-

synonymous mutations may also be due to balancing selection. Also, a reduced synonymous polymorphism frequency may arise due to a selective sweep.

There are also versions of the MK-test that do not compare synonymous and non-synonymous mutations and are not restricted to protein coding regions. In these cases, fixed and polymorphic mutation counts of different loci are compared, and the results are interpreted by comparing loci relative to each other (Andolfatto, 2005). This analysis is also used for estimation of selection coefficients (Sawyer and Hartl, 1992),(Sethupathy and Hannenhalli, 2008).

Other methods focus on within-species sequence variation. They are applicable to coding and non-coding regions, and are able to detect recent, even ongoing selection events, and detect selection acting on a single nucleotide.

The site frequency spectrum (SFS) refers to the distribution of sites within a sequence according to their allele frequencies (Nielsen and Slatkin, 2013). The SFS carries useful information about evolutionary forces acting on a sequence. The SFS for derived allele frequencies is called the **derived allele frequency distribution** (DAF, also called the "unfolded" SFS).

Comparing the DAF of a population of sequences against a distribution expected from the neutral model can reveal selection. Figure 1.1 shows expected distribution of derived allele frequencies on a population with 10 diploid individuals (adapted from (Nielsen and Slatkin, 2013)). The DAF distribution under the neutral model is a negative exponential function. Negative selection increases the number of sites that have low derived allele frequency, leaving few sites having intermediate frequencies, and yet fewer sites having high frequencies. Positive selection acts on a sequence in the opposite way, increasing the number of sites with high derived allele frequency. Positive selection can also increase the number of intermediate frequency sites to a lesser extent.

Another type of detectable event from the DAF are selective sweeps. This occurs when a single mutation (or small number of mutations) undergoes strong positive selection, and polymorphic positions in the proximity either increase or decrease fre-

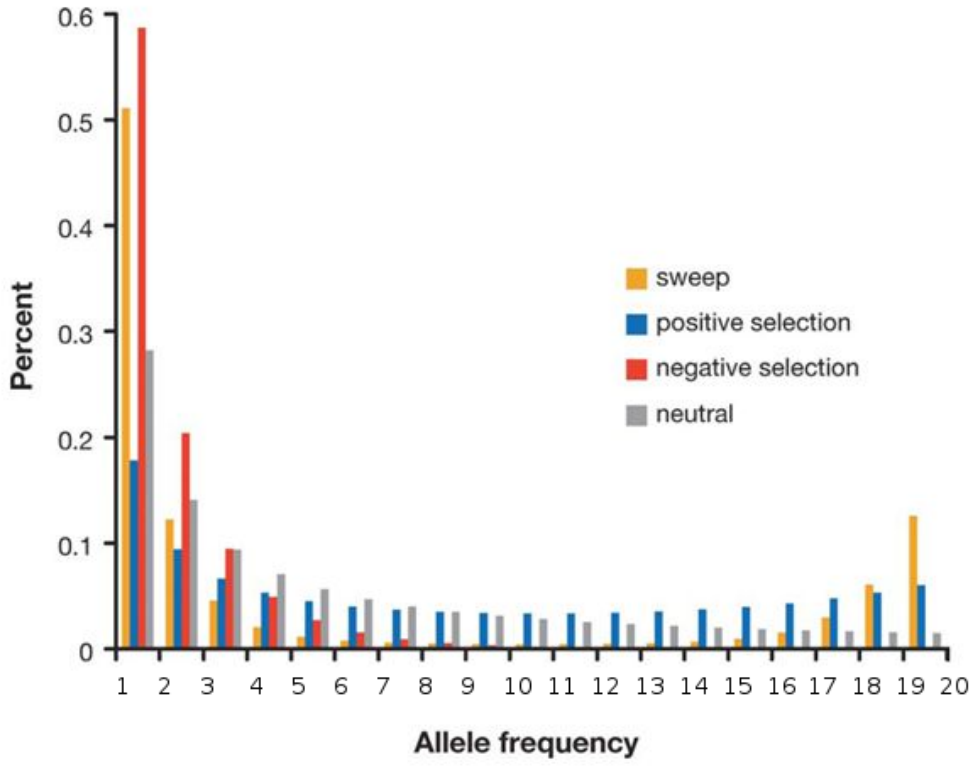

Figure 1.1: Derived allele frequency spectrum

quency of their derived alleles, depending on whether their derived allele is linked to the positively selected allele or not. A selective sweep results in increased number of positions of both low and high derived allele frequencies and decreased number of sites with intermediate derived allele frequency. The effect of selective sweeps is different than that of positive selection on a sequence, in that, when a sequence is under positive selection, many derived alleles on the sequence increase their frequency in the population due to selection acting on them. When a sequence is under the effect of selective sweep, few positions are under selection, and the allele frequencies of the rest of the positions change according to which of their alleles is linked to the selected allele on the selected position.

The **Tajima's D** test provides a numerical assessment of whether observed derived allele distribution fits to neutrally expected distribution (Tajima, 1989). This test can be applied to population genetic data where orthologous sequences from multiple individuals are aligned. Tajima's D uses the ratio between number of polymorphic sites in the alignment and mean of pairwise difference between all sequences, as test statistic.

When fewer than expected (i.e. than neutrally expected) polymorphisms have high derived allele frequencies, mean pairwise difference in the population is lower than expected. This can be due to negative selection or a recent selective sweep. In the opposite case, with many more intermediate and higher derived allele frequency polymorphisms, mean pairwise difference is expected to be more than that expected by neutral evolution. This observation can be due to recurrent positive selection or balancing selection. An observation conforming with the expected mean pair difference (given the number of polymorphic sites) is an indication of neutral evolution.

Fay and Wu's H is an improvement over Tajima's D. This test resolves the ambiguity about cause of the observed deviation, by incorporating a comparison with outgroup sequences and inferring the ancestral state (Fay and Wu, 2000). The difference between a selective sweep and negative selection will be more obvious using ancestral state information: the negative selection scenario will have more similar ancestral and derived sequences than the selective sweep scenario.

1.3 Patterns Formed by Non-selective Processes

Natural selection is one of the main mechanisms of evolution. Throughout the development of the field, however, other processes have been shown to be effective as well (Hamilton, 2009). Random genetic drift changes allele frequencies, through random variation around the expected value of each individual's contribution to the next generation. Even if these expected values (fitness values) are equal for competing alleles, eventually one of them will get fixed in the population, as biological populations are finite and this random fluctuation will hit the irreversible boundary of the frequency range.

Mutations are raw material of evolution that provide the variation upon which selection and drift act. The rate of new mutations affects genome evolution rate mostly but not exclusively in neutral parts. Heterogeneity in mutation rates among mutation types is a factor shaping genome evolution. Nearly without exception, the effect of mutation rate heterogeneity does not manifest itself as recurrent mutations creating the same allele in distinct individuals and increasing that allele's frequency. Mutation rates are too low to make such an impact. They are, in fact so low in many natural populations that most of the possible alleles are not present in the gene pool at any given time, and competition takes place among the present ones (Nei, 1987), (McCandlish and Stoltzfus, 2014). In these circumstances, alleles that show up more often have more chances of getting fixed.

In case of single nucleotide substitutions, a prevalent mutation rate bias is towards transitions (Fitch, 1967) (Vogel, 1972). Figure 1.2 shows types of mutations with respect to involved bases (by Petulda (Own work) [CC BY-SA 4.0 (http:// creativecommons.org/ licenses/by-sa/ 4.0)], via Wikimedia Commons). Transitions ($A \leftrightarrow G$ and $T \leftrightarrow C$) occur much more frequently than transversions ($A \leftrightarrow T$, $A \leftrightarrow C$, $G \leftrightarrow T$ and $G \leftrightarrow C$). On an A nucleotide background for example fixation of a G is more probable than a C or T.

Neighbouring base **context** can also influence mutation rates. In CpG sites in many eukaryotic genomes, where a C precedes G, methylation of the C increases the likeli-

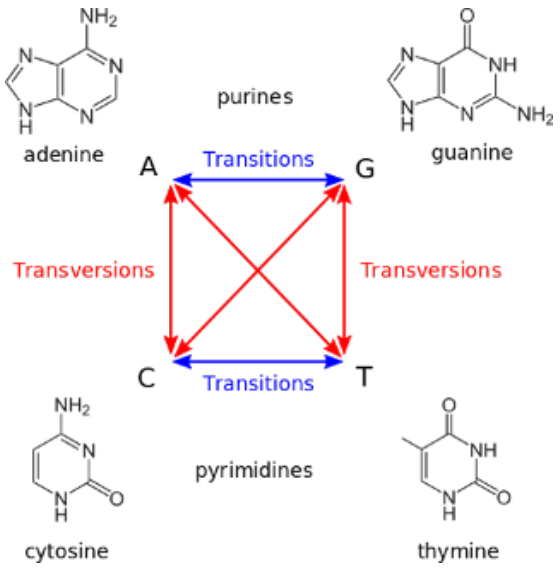

Figure 1.2: Transition and transversion mutations

hood of a $C \rightarrow T$ mutation, because the common cytosine deamination event results in a thymine in the case of methylated cytosine (instead of a detectable uracil in the case of unmethylated cytosine) (Ehrlich and Wang, 1981) (Krawczak et al., 1998). This is a prominent example of a mutation bias caused by context.

Recombination is another population genetic process shaping genome evolution. Its primary role in evolution is to increase genomic diversity. Recombination rate is heterogeneous throughout genome. This produces patterns of longer and shorter haplotypes and also clusters of recombination artifacts in genome. Gene conversion is one of the those artifacts with interesting consequences (Chen et al., 2007).

Recombination in meiotic cell division results with reciprocal exchange of chromosomal parts between homologous chromosomes. In each recombination event, however, there is a small (1kbp) sequence of non-reciprocal combination. This sequence is the part of the recombination complex that falls in between two Holliday junctions. DNA strands of two homologous chromosomes complement each other and mismatches due to heterozygous alleles are repaired using one of the alleles as the donor. Hence, this phenomenon, called gene conversion, causes a non-Mendelian segregation of the alleles in a single meiosis event. Figure 1.3 depicts mechanisms of recombination

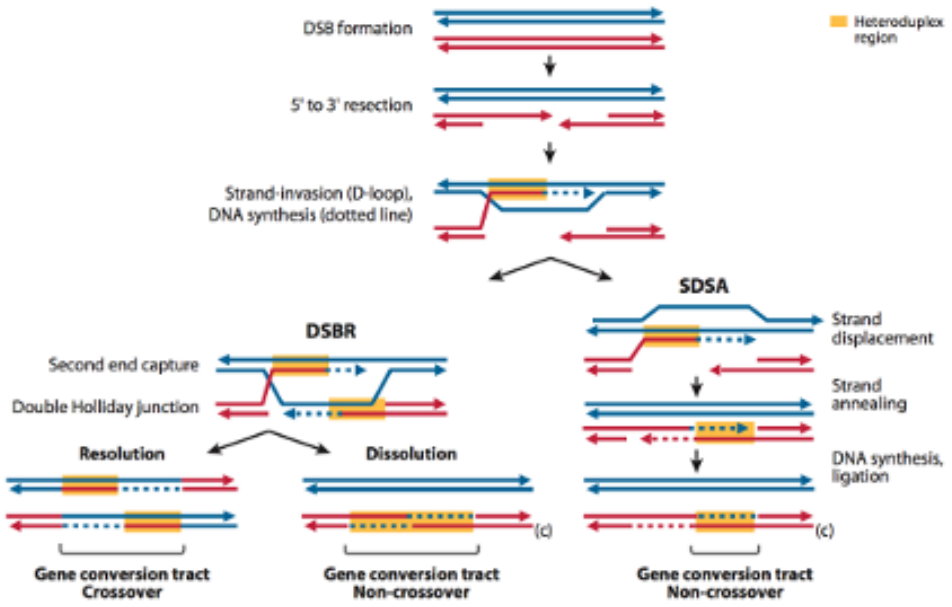

Figure 1.3: Mechanisms of recombination and gene conversion.

and gene conversion (adapted from (Duret and Galtier, 2009)).

One would expect that the donor allele in each meiosis event to be chosen at random, and thus gene conversion is not expected to have an effect on evolution. But, findings show that a bias towards G/C alleles during gene conversion is present in mammalian genomes (Duret and Galtier, 2009). Whenever a G/C allele is matched to a A/T allele during gene conversion, the G/C allele has a tendency to be preserved on the chromosome. This G/C biased gene conversion (BGC) mimics natural selection, as heterozygotes produce more progeny with the over-represented allele, which in turn increases that allele's frequency and fixation probability. Moreover, such a fixation bias can lead to fixation of deleterious alleles, counter-acting effect of natural selection.

1.4 Motivation

A mutational bias may refer to several distinct but related events. Mutation rate bias refers to a systematic difference in per base per generation (or more generally per locus per unit time) rate. This bias may be depend on genomic context (neighbouring simple sequence motifs, or functional regions like recombination hotspots) or be independent of context.

A substitution bias refers to differential rate of substitution of alleles within certain genomic contexts and/or of alleles due to certain types of mutations. This may be result of a mutation rate bias, e.g. among neutral regions, those with higher mutation rates are expected to accumulate more substitution with respect to those with lower mutation rates. Alternatively, a substitution bias may be caused by a fixation bias, differential probability for fixation of mutations, once they arise. From a population genetics view, this may represent recurrent positive selection, but it may also be due to an intrinsic effect like preferential repair of allelic mismatches, and thus can be considered as a special case of mutational biases.

In the previous section a variety of mutational bias examples were discussed:

- A context-independent mutation rate bias among mutation types without a directional preference (transition bias)

- A context-dependent mutation rate bias among certain mutation types with a preferential direction (CpG->TpG)

- A context independent fixation bias among mutation types with a directional preference (gBGC)

However, to our knowledge, there is no report of a context-dependent genome-wide fixation bias as yet. Here the human genome is investigated for the possible existence of a **fixation bias** caused by neighbouring base **context**. In other words, we seek for sequence patterns subject to fixation biases across the genome, which could arise due to selection or gene conversion-related processes. Our aims are:

- Testing for a sequence context-dependent fixation bias for all possible mutation / context combinations,

- Studying common sequence and functional characteristics of any mutation / context combinations that may show fixation biases.

CHAPTER 2

METHODS

This chapter begins with explaining structure of the study by defining the mutation pattern, which is basic unit of the study, and demonstrating how they are analysed to reveal a possible, sequence context-dependent fixation bias. Then, it continues by describing the data used and how the information on fixation bias is extracted. The later sections describe analyses for characterisation of the results.

2.1 Mutation Patterns

A mutation pattern, in this study, is defined as a nucleotide sequence pattern of a given length, associated with a point mutation in the middle position of the sequence, $abXcd \rightarrow abYcd$. 5 bp long sequences are used throughout this study, and each such sequence pattern is called a pentamer, $abXcd$. A mutation pattern, for example, will be denoted as:

$$AGCGT \rightarrow AGGGT$$

where AGCGT is the inferred ancestral sequence and AGGGT is the inferred derived sequence. $C \rightarrow G$ is the associated point mutation in this case.

For a mutation pattern $abXcd \rightarrow abYcd$, another pattern with same context bases (first two and last two bases in the pentamer) and reverse point mutation is called its reverse pattern: $abYcd \rightarrow abXcd$. Reverse patterns are compared against each other

throughout the analysis.

$AGCGT \rightarrow AGGGT$	focal (forward) mutation pattern
$AGGGT \rightarrow AGCGT$	its reverse pattern

For a mutation pattern $abXcd \rightarrow abYcd$, the mutation pattern on the complementary strand, is called its complementary pattern: $d'c'X'b'a' \rightarrow d'c'Y'b'a'$, where x' is the complementary base of x for any $x \in \{A, G, C, T\}$. Figure 2.1 illustrates an example mutation pattern and its related patterns on a hypothetical genomic background. Dashes on the alternative allele row denote absence of alternative allele (no polymorphism). The first four highlighted pentamers are fixed and latter two are polymorphic. When polymorphism is present, derived allelic state may be reference allele as well as alternative allele. Both the focal pattern and the complementary pattern are written in their $5' \rightarrow 3'$ direction. These two patterns are coupled in the genome, hence are coupled in our analysis. So for the previous example they are in the following form:

$AGCGT \rightarrow AGGGT$	focal mutation pattern
$ACGCT \rightarrow ACCCT$	its complementary pattern

2.2 Structure of Analysis

The main question of this study is whether certain mutation patterns are more likely to be fixed in human genome evolution in comparison to their reverse mutation patterns. A McDonald-Kreitman test framework is employed to answer this question (Nielsen and Slatkin, 2013).

Difference between the test used in this study and the original McDonald-Kreitman test is that; in this study selection on alleles across many loci is tested by comparing alleles' fixation odds with some other alleles'; whereas in the original test, selection on a locus is tested by comparing fixation odds of non-synonymous and synonymous

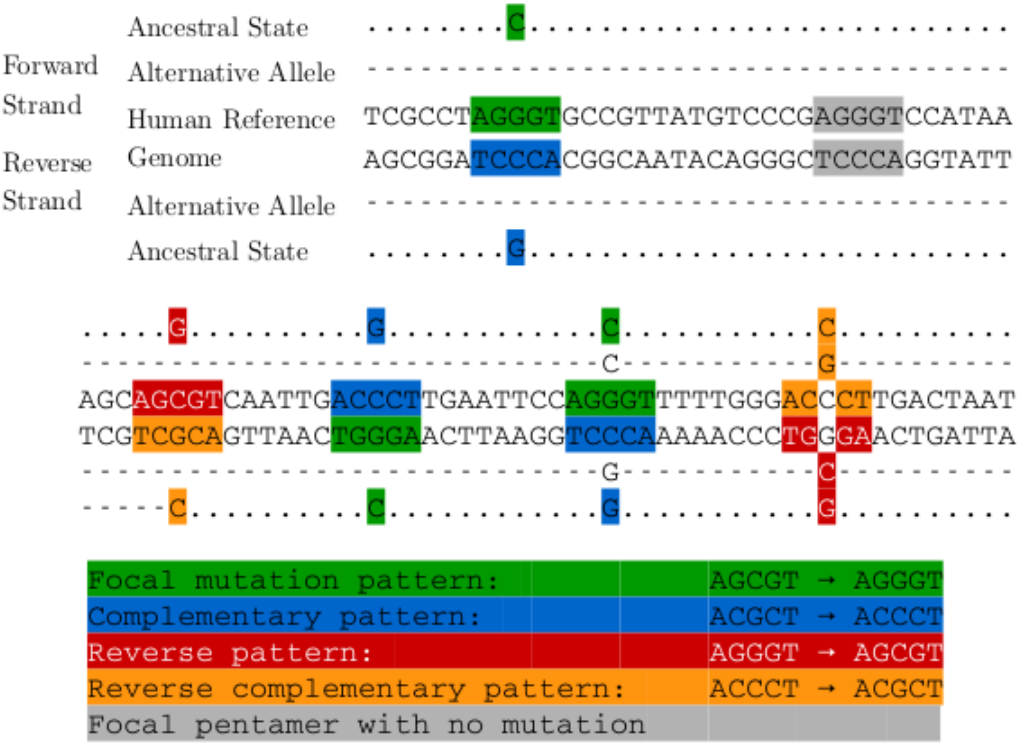

Figure 2.1: An example mutation pattern in a hypothetical genomic background

mutations on the same locus. A locus in this study is a single bp long position, whereas in the original usage of McDonald-Kreitman test it is the whole coding sequence of a gene, or a family of related genes combined.

Also the term selection is used here, in a broader sense, as a differential tendency (bias) of an allele to spread within a population. Such a bias is due to natural selection in most cases, but may be also due to intrinsic properties of the living system (see Introduction). The latter case is actually the only explanation for alleles on non-functional loci.

For each mutation pattern and its complementary pattern, we first calculate their aggregate fixed state frequency (F_f) and their aggregate polymorphic state frequency (P_f) across the whole genome. The ratio of these frequencies, ($\frac{F_f}{P_f}$), is then compared to the same ratio for the reverse pattern and reverse complementary pattern ($\frac{F_r}{P_r}$). These ratios are also odds, as each pattern is either fixed or polymorphic (Table 2.1). The ratio of these odds ($\frac{F_f}{P_f}/\frac{F_r}{P_r}$) can be used to infer a fixation bias, based on the theory described below.

$$abXcd \rightarrow abYcd + d'c'X'b'a' \rightarrow d'c'Y'b'a' \quad \text{focal pattern pair's F/P}$$
$$(1a) \qquad\qquad (1b)$$
$$abYcd \rightarrow abXcd + d'c'Y'b'a' \rightarrow d'c'X'b'a' \quad \text{reverse pattern pair's F/P}$$
$$(2a) \qquad\qquad (2b)$$

As in the original McDonald-Kreitman test, it is assumed that polymorphic events represent relatively neutral events (or the background mutation rate) more than fixed events. This assumption is based on the understanding that natural selection has had less time to act on polymorphisms. For instance, a slightly deleterious mutation can be represented as polymorphism, but is not expected to spread to the whole population and reach fixation. In an unbiased situation an odds ratio (OR) around 1 is expected. $OR > 1$ indicates a fixation bias in favour of the forward pattern relative to the

Table 2.1: Structure of the comparison table

Mutation Pattern	Fixed frequency	Polymorphic frequency
AGCGT → AGGGT ACGCT → ACCCT	F_f	P_f
AGGGT → AGCGT ACCCT → ACGCT	F_r	P_r

reverse pattern, whereas $OR < 1$ indicates a bias towards reverse pattern. Statistical significance of the computed odds ratios can be determined by the Fisher's exact test.

There are $4^5 = 1024$ distinct pentamers. For each pentamer, there are 3 possible point mutations, which produce together 3072 distinct mutation patterns. When the reverse pattern and the complementary pattern of a mutation pattern coincides ($1b = 2a$ and $1a = 2b$), a self to self comparison occurs. For example:

$$\underset{(1a)}{AGCCT \rightarrow AGGCT} + \underset{(1b)}{AGGCT \rightarrow AGCCT} \qquad \text{focal pattern pair}$$

$$\underset{(2a)}{AGGCT \rightarrow AGCCT} + \underset{(2b)}{AGCCT \rightarrow AGGCT} \qquad \text{reverse pattern pair}$$

This case occurs whenever $a = d'$ and $b = c'$ and $X = Y'$. Each equation has 4 solutions (a, b and $X \in \{A, G, C, T\}$, and they account for $4^3 = 64$ focal patterns causing self to self comparison. After elimination of these cases, 3008 distinct mutation patterns remain.

Computing OR for each of the mutation patterns as focal pattern would, however, lead to 4-fold repeated comparisons. Two-fold redundancy occurs due to complementary patterns, with each pattern coming up once as focal pattern and once as complementary pattern. Because complementary patterns are essentially the same in this study (we do not consider strand bias), I merge the data for focal and complementary patterns and use only one as representative.

Another 2-fold redundancy is due to handling each reverse pattern as focal pattern as well. Again, using either of these is sufficient, because the computed odds ratios indicate a bias (if it exists) in favour of either the focal pattern or the reverse pattern. If we repeat comparisons with switched focal-reverse pattern roles, for each odds ratio we also calculate its inverse ratio (e.g. if the odds ratio for $AGCCT \rightarrow AGGCT$ is 2, the odds ratio for $AGGCT \rightarrow AGCCT$ will be 0.5). This artificial symmetry in the results may mask any less obvious pattern in the data, especially in graphical representations. Eliminating repeated comparisons leaves us with 752 distinct comparisons.

2.3 Processing Data

The human genome hg19/GRCh37 is used as reference genome. For determining mutations and finding ancestral alleles, 46-way multiple alignment dataset of 45 vertebrate species' genomes to hg19/GRCh37 from the UCSC Genome Browser is used (Blanchette et al., 2004), (Karolchik et al., 2004). The 1000 Genomes project phase 3 release data (Auton et al., 2015) is used to determine mutations that are polymorphic within modern human population. The 1000 Genomes project also uses hg19/GRCh37 as reference genome.

For retrieving, transforming and filtering the data, I used the Python language (Python Software Foundation, https://www.python.org/), specific Unix software such as bedtools (Quinlan and Hall, 2010), as well as awk and bash commands. I used the R language (R Development Core Team, 2008) to conduct statistical analyses.

Genome annotations for exon and promoter regions and repeat positions were retrieved from the UCSC Table Browser or Ensembl Biomart (Yates et al., 2016), Ensembl Release 75.

2.3.1 Identifying Mutations

Chimpanzee and orang-utan were chosen as outgroup species to determine ancestral alleles. Chimpanzee is a classical choice, being the closest species to human. Choice of orang-utan, but not a closer species (gorilla) is because the gorilla genome is sequenced at lower depth (has lower quality) than orang-utan genome (ponAbe2, 2007). Meanwhile, choosing a more distant species would increase the probability of misidentifying the ancestral allele due to homoplasies. Also a distant species would provide fewer number of homologous positions, which would in turn decrease sample size (number of genomic positions surveyed).

Using a parsimony approach to infer ancestral state of a genomic position, a nucleotide in present-day human genomes is considered as a mutation, if at the homologous locus chimpanzee and orang-utan share the same allele, and either some or all humans carry a different allele. In this case, chimpanzee and orang-utan alleles are considered as the ancestral state. The cases where chimpanzee and orang-utan alleles do not match each other are excluded as unknown ancestral state (Figure 2.2(a)), along with some other cases to be explained later in this section.

Each mutation is also classified as either polymorphic or fixed, according to its presence/absence in the 1000 Genomes project phase 3 release dataset. Mutations on genomic positions not captured in the 1000 Genomes project are classified as fixed, whereas positions that are among 1000 Genomes SNP's are classified as polymorphic.

2.3.2 Domain of Analysis

In order to prevent any ambiguity in identification of ancestral alleles and sequence context bases, a restricted group of mutations are included in the analysis. Mutations considered in the study are those that are single nucleotide mutations, having two alleles (biallelic, including the ancestral allele) and at least 2 bp apart from any mutation, either polymorphic or fixed (Figure 2.2(b)). Figure 2.2 lists examples for each of mentioned cases. Bases in black are unchanged among homologous sites, green

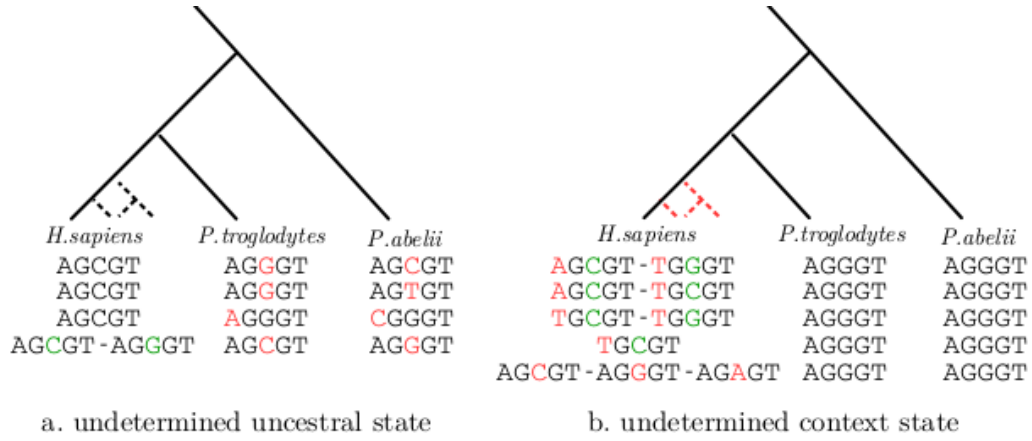

Figure 2.2: Allele genealogy

coloured ones indicate an unproblematic change and red coloured ones highlight reason for exclusion from analysis. Panel (a) shows examples for cases where ancestral state of the pentamer sequence can not be determined by parsimony approach, panel (b) shows examples for cases, where two polymorphic mutations, a fixed and a polymorphic mutation, a polymorphic and a fixed mutation, two fixed mutations, and a polymorphic position with 3 alleles are present within a pentamer, respectively.

Point mutations constitute majority of variation in genome, are easier to detect with higher confidence relative to structural variants and are homogeneous in their structure. In cases with more than two alleles or with mutated context bases, concerns arise about the state of the pentamer pattern at the time of the mutation. In the multiple derived allele case, it is unknown whether both of them are derived from the ancestral allele or some of them are derived from another derived allele. In case where mutations are in each others' context sequence, it is unknown which of them emerged and possibly fixed prior to the other one, hence what the actual context was during the emergence and spread of the focal mutation.

Table 2.2: Mutation counts before and after filters

Polymorphic Mutations		Fixed Mutations	
Count	%	Count	%
81,271,745	100	7,349,796	100
64,293,247	79	3,733,476	51
31,399,150	39		
31,331,142	39		
22,542,585	28		

2.3.3 Filtering Data

During assembly of population genomic data, paralogous sites (with mutation(s) between their sequences) can be mistakenly identified as heterozygous alleles in homologous chromosomes in a single site , leading to sites with false excess heterozygosity. Detecting positions with excess heterozygosity and excluding them increases reliability of the data. Thus, polymorphic sites that deviate significantly from Hardy-Weinberg equilibrium are filtered-out from the data.

Table 2.2 lists number of mutations after each of the applied filters. Column for polymorphic mutations (each row is a subset of the raw above) lists numbers of 1000 Genomes Phase 3 variation count, biallelic SNPs that are at least 3 bp apart, pentamers that do not overlap with repeat regions, alleles that are in Hardy Weinberg equilibrium and mutations with all context bases in ancestral state, respectively. Column for fixed mutations lists mutations that are not found in the 1000 Genomes project data, the human reference allele is derived and the ancestral state is inferable, with all context bases in ancestral state, in first row and pentamers that do not overlap with repeat regions in second row.

Genotype and allele frequencies for testing fit to Hardy-Weinberg equilibrium are retrieved from variant call format (.vcf) files, released with phase 3 of 1000 Genomes project. For each of the 5 “super populations” (AFR:African, AMR:Ad Mixed American, EAS:East Asian, EUR:European, SAS:South Asian) the expected (derived from observed allele frequencies p, q using $p^2 : 2pq : q^2$ genotype ratio of Hardy-Weinberg

equilibrium) and observed genotype frequencies are compared with a Chi-square test using the scipy module of Python (Jones et al., 01). Mutations that deviate from HWE in at least two super populations are excluded from the analysis. Highly repetitive genomic regions create similar concerns about genotype reliability. Different regions with similar sequences may be identified as alternative alleles to each other. Mutations with any overlap in their pentamer sequence with repeat positions are excluded from the analysis as well. Repeat positions are retrieved from the UCSC Table Browser (Karolchik et al., 2004) (hg19 Repeats group, repeatmasker track, rmsk table).

2.4 Genome Subsets for Analysis

Results may be changed when occurrences of mutations from only certain regions of genome are considered, or rare alleles are excluded from analysis. Also results may be distributed disparately among different point mutation types. The dataset was partitioned accordingly to investigate the effect of genomic region context, minor allele frequency and mutated bases on fixation bias, as described below.

2.4.1 Genomic Context

Comparisons were done using occurrences of mutation patterns (a) in the whole genome (excluding sex chromosomes and mitochondrial DNA), (b) using occurrences in exons only and, (c) occurrences in promoters only. Results from these restricted regions may point to a possibly functional link of an observed bias. Annotations of exonic and promoter regions were retrieved from Ensembl Genes 75 and Ensembl Regulation 75 databases, respectively (from the Ensembl Archive version Feb 2014 (GRCh37.p13) using Biomart).

Table 2.3: Counts and percentages of mutations in data partitions

			Count	%
Fixed	whole genome		3,733,476	100.00
	exons		369,057	9.89
	promoters		186,680	5.00
Polymorphic	whole genome	no maf filter	22,542,585	100.00
		0.001 maf	7,006,216	31.08
		0.01 maf	3,151,100	13.98
		0.05 maf	1,768,819	7.85
	exons	no maf filter	3,135,752	13.91
		0.001 maf	771,372	3.42
		0.01 maf	308,368	1.37
		0.05 maf	163,397	0.72
	promoters	no maf filter	1,215,129	5.39
		0.001 maf	365,604	1.62
		0.01 maf	156,246	0.69
		0.05 maf	84,731	0.38

2.4.2 Mutation Type

Some mutation patterns are subject to known mutation or fixation biases. In order to avoid the effect of these biases on results, mutation patterns involving certain point mutations are examined exclusively.

Transitions occur more frequently than transversions. Using transversions only decreases the risk of ancestral sequence misidentification, which can be due to repeated mutations at the same position. This poses a significant problem for the analysis as it will inflate both fixation counts and derived allele frequencies. Considering mutation patterns involving $A/G \leftrightarrow C/T$ only, thus provides a more reliable subset of the results.

Furthermore gBGC (see Introduction Section 1.3) is expected to introduce a fixation bias when $A \leftrightarrow C$ and $G \leftrightarrow T$ mutations are compared. Hence, further narrowing the focus and only considering mutation patterns with $A \leftrightarrow T$ and $G \leftrightarrow C$ mutations frees results of a possible confounding effect of gBGC.

2.4.3 Minor Allele Frequency

When counting occurrences of a polymorphic mutation, three different thresholds were used as condition for inclusion in the analysis. For each of the three thresholds (0.001, 0.01 and 0.05), only the polymorphic mutations with a higher minor allele frequency (MAF) than the respective thresholds were included. This way, possibility of bias through excess rare alleles is eliminated. Excess of rare alleles may result from false positives of SNP calling (e.g. due to sequencing error) from population whole genome sequencing data.

2.5 Derived Allele Frequency

Comparing derived allele frequencies of polymorphic mutations of a mutation pattern with its reverse mutation pattern is another approach to test whether certain mutation patterns are more likely to become fixed in the population relative to their reverse patterns.

Derived allele frequencies were computed from the genotype data available in the same vcf files of the 1000 Genomes project (see Section 2.3 above). Here I first determined which of the alleles in the vcf file is the derived one. Determination of the derived allele was done in same fashion as determination of the ancestral allele described earlier in this chapter (Section 2.3.1). Allele frequencies were obtained from 2504 individuals from 27 global populations surveyed in the project (Auton et al., 2015).

2.6 Context Metrics

A preliminary result that became apparent when we studied the MK test odds ratio distributions was that, mutation patterns creating more repetitions of a base in derived pentamer sequence with respect to its ancestral sequence are more likely to be fixed.

I then sought an explicit way to measure and demonstrate this observation. We name

the phenomenon of creation of repeated bases through mutation homopolymerization. I thus defined a score to assess strength of homopolymerization.

Score 1: Change in the number of adjacent bases identical to the middle base.

This score ranges between [-3,3]. It measures the extension of a repeated base sequence due to the derived allele. Mutations that connect same bases on two sides of the context get higher values for this score. Mutations that create 4-mer and a 3-mer can not be differentiated if both increase ancestral homopolymer length by 2. Also a disconnected base of the same type does not affect this score.

Score 2: Change in the number of bases identical to the middle base.

Ranging between [-1,1], this score measures same type base density without requiring the bases to be adjacent. It tells increase/decrease of same type base density due to mutation, and provides no information about that density itself.

Score 3: Maximum number of adjacent bases identical to the middle base, in the ancestral or derived pentamer motif, with positive values if adjacent base length is greater in derived motif and negative values if ancestral motif has greater homopolymer length. Zero is assigned in case of a tie.

This score is meant to complement score type 1, differentiating mutation patterns according to the length of the extended/disrupted homopolymer. This score ranges between [-5,5].

Score 4: Same as score 3, except that, bases that are identical to the middle base do not have to be adjacent.

Using these scores, mutation patterns are grouped to compare distribution of test results with other groups, i.e. to test whether a fixation bias is associated with higher values of a score type, by comparing test result distributions of groups with a score level of 3 vs 2 vs 1 etc.

CHAPTER 3

RESULTS

3.1 Exploratory Analysis

As a preliminary analysis, frequencies of all 1024 pentamer motifs throughout human genome are counted. Then, using a G/C ratio calculated from the same data (hg19 fasta file), expected frequencies for the motifs are generated. Distribution of observed frequencies deviate significantly ($\chi^2 = 657630000$, $df = 1023$, p-value $< 2.2e-16$) from expected frequency distribution. 278 motifs out of 1024 (27%), are observed either at least twice of expected or at most half of expected frequency. This deviation suggests a strong non-randomness in motif composition of human genome. Throughout this study, I will examine, if this non-randomness may be attributed to a fixation bias towards certain pentamer motifs. Figure 3.1 shows observed and expected frequencies of motifs with highest and lowest observed/expected ratio.

Odds of fixation (Prob(fixation)/Prob(polymorphic)) is a central statistic in the McDonald - Kreitman framework (see Methods). **Odds of fixation** for mutation patterns (abXcd -> abYcd) range between 0 and infinity , with the expected value of 1. Logarithm of this statistic constitutes a symmetric distribution around the expected value of 0, ranging between - infinity and + infinity, and may be approximated by a normal distribution. For sake of simplicity, odds of fixation is referred as odds, hereafter.

The **odds ratio**, which is the ratio of the fixation odds of compared pairs of mutation patterns, is another central test statistic and shows similar properties as odds described above. Logarithms of these statistics are more useful both for visual and inferential

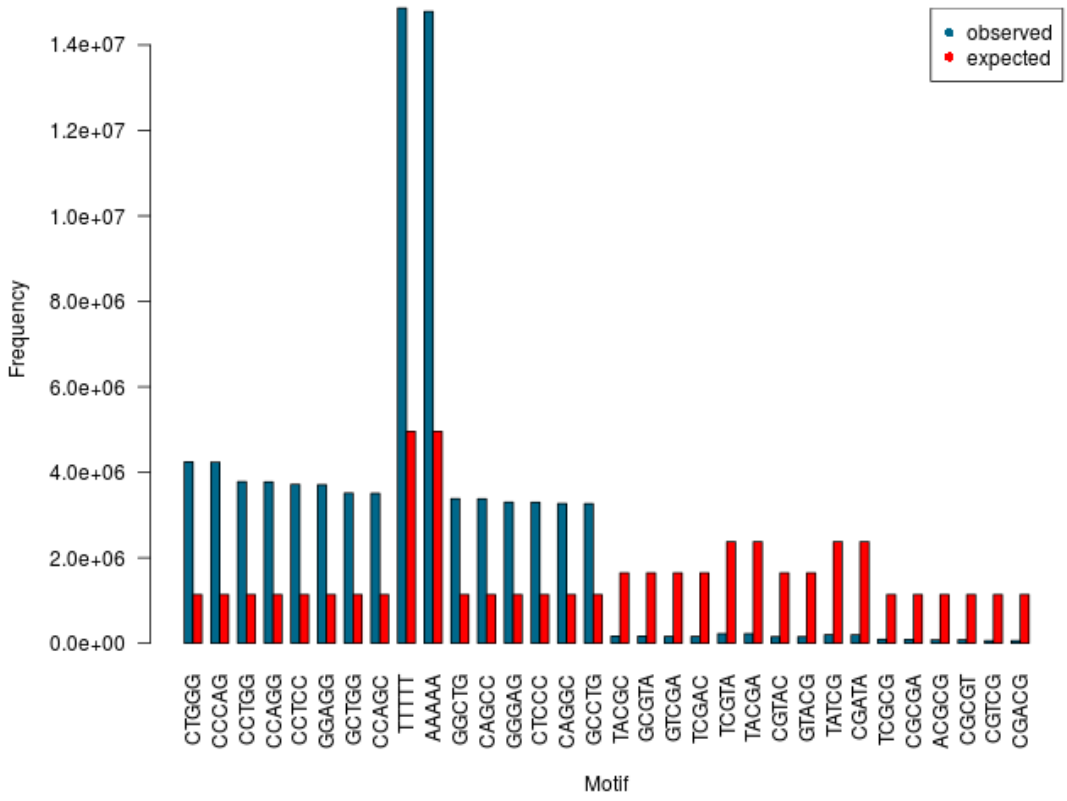

Figure 3.1: Expected and observed frequencies of pentamer motifs

purposes, as their distribution is symmetrical around mean and covers all real numbers. Thus, natural logarithms of these statistics are used throughout the study.

3.1.1 Fixation Odds Ratios

As a first step in the analysis, for each possible mutation pattern together with its complementary pattern, I tabulated frequencies of fixed and polymorphic occurrences, frequencies of those occurrences for their reverse pattern, the odds ratio, the Fisher's exact test p-value and homopolymerization scores (see Methods). Table 3.1 shows the first 15 comparisons, sorted by descending odds ratio values. The full table for 752 unique mutations pattern comparisons is provided in Appendix. I removed repeated comparisons due to two-fold appearance of each mutation pattern once as focal and once as complementary pattern. P-values are two sided Fisher's exact p-values with correction for multiple testing using Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995).

The Fisher's exact test is expected to produce astronomical significance levels, as there are very high values in contingency table for each comparison. With increasing sample size, statistical tests tend to report even small deviations from null hypothesis as statistically significant. The resulting statistically significant biases with very small effect sizes may not carry any biological significance, and instead reflect small methodological perturbations (see Discussion).

As an accompanying, more conservative method to assess statistical significance, an empirical distribution of odds ratios was generated from the data, by computing odds ratios from fixation and polymorphism frequencies of randomly selected two mutation patterns, a total of 10,000 times. The simulation showed that an odds ratio of 2.0315 (log odds ratio of 0.7088) or higher could be considered as a significant deviation ($p < 0.05$) from the null hypothesis of OR = 1 (log OR = 0). Out of 752 mutation pattern comparisons, 83 (11%) were found to be individually showing a significant fixation bias.

Overall, this analysis implies the presence of strong fixation biases for specific muta-

Table 3.1: McDonald-Kreitman test results

	fix1	poly1	fix2	poly2	OR	pval	s1	s2	s3	s4
TACGC->TAGGC + GCGTA->GCCTA vs TAGGC->TACGC + GCCTA->GCGTA	163	186	368	1750	4.1642937	1.554612e-29	1	0	2	0
TACGG->TAGGG + CCGTA->CCCTA vs TAGGG->TACGG + CCCTA->CCGTA	171	161	858	3047	3.7704557	3.172366e-28	1	1	3	3
TTCGT->TTGGT + ACGAA->ACCAA vs TTGGT->TTCGT + ACCAA->ACGAA	319	333	1197	4518	3.6148392	1.522898e-48	1	1	2	2
TGCGT->TGGGT + ACGCA->ACCCA vs TGGGT->TGCGT + ACCCA->ACGCA	312	327	1082	4041	3.5624574	4.604498e-46	2	1	3	3
TTCGC->TTGGC + GCGAA->GCCAA vs TTGGC->TTCGC + GCCAA->GCGAA	182	211	510	1993	3.3690115	3.224565e-25	1	0	2	0
CTCGT->CTGGT + ACGAG->ACCAG vs CTGGT->CTCGT + ACCAG->ACGAG	431	472	1045	3793	3.3138188	3.444028e-54	1	0	2	0
CGTCA->CGGCA + TGACG->TGCCG vs CGGCA->CGTCA + TGCCG->TGACG	55	39	95	221	3.2702543	1.843622e-06	1	1	2	2
TACGA->TAGGA + TCGTA->TCCTA vs TAGGA->TACGA + TCCTA->TCGTA	265	299	933	3369	3.1992231	4.780565e-34	1	1	2	2
TACGA->TAAGA + TCGTA->TCTTA vs TAAGA->TACGA + TCTTA->TCGTA	309	336	899	3053	3.1221067	1.708626e-36	1	1	2	3
GTCGG->GTGGG + CCGAC->CCCAC vs GTGGG->GTCGG + CCCAC->CCGAC	276	316	908	3218	3.0944604	2.375309e-33	1	1	3	4
GACGG->GAAGG + CCGTC->CCTTC vs GAAGG->GACGG + CCTTC->CCGTC	323	368	724	2514	3.0466662	1.608307e-35	1	1	2	2
TACGT->TAGGT + ACGTA->ACCTA vs TAGGT->TACGT + ACCTA->ACGTA	297	371	938	3568	3.0441964	1.745565e-35	1	1	2	2
AACGG->AAAGG + CCGTT->CCTTT vs AAAGG->AACGG + CCTTT->CCGTT	408	474	883	3062	2.9840288	9.268518e-43	1	1	3	3
AACGT->AAGGT + ACGTT->ACCTT vs AAGGT->AACGT + ACCTT->ACGTT	379	451	1248	4422	2.9769674	7.351755e-43	1	1	2	2
GACGA->GAAGA + TCGTC->TCTTC vs GAAGA->GACGA + TCTTC->TCGTC	276	347	864	3233	2.9753659	7.063028e-32	1	1	2	3

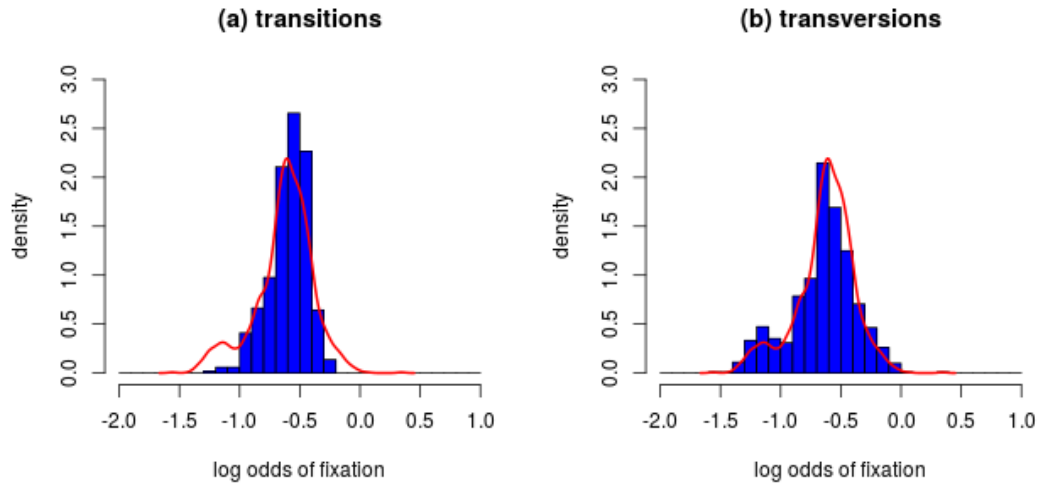

Figure 3.2: Distribution of fixation odds

tion patterns. Below, I first ask if these could arise due to mutation rate differences among patterns, and I also test the presence of a known fixation bias.

3.1.2 Distributions of Odds

The McDonald-Kreitman method is developed to detect fixation biases, and is not supposed to be affected by mutation rate biases (see Methods). Odds of mutations that differ in their rates are compared against each other, as negative control, and thus mutation rate heterogeneity is not expected to become a confounding factor in the used method. However, it is also plausible that mutation rate differences influence fixation and polymorphism detection rates differentially, e.g. due to differences in sequencing technology used to determine fixation and polymorphism.

To address this here I compared fixation odds for transition, known to have high mutation rates, and transversions, with generally lower mutation rates (see Introduction). Figure 3.2 shows the histogram of odds (fixation/polymorphism) of all mutation patterns involving a transition (3.2(a)) or a transversion (3.2(b)) mutation together with distribution density estimation for odds of all mutation patterns (the red curve).

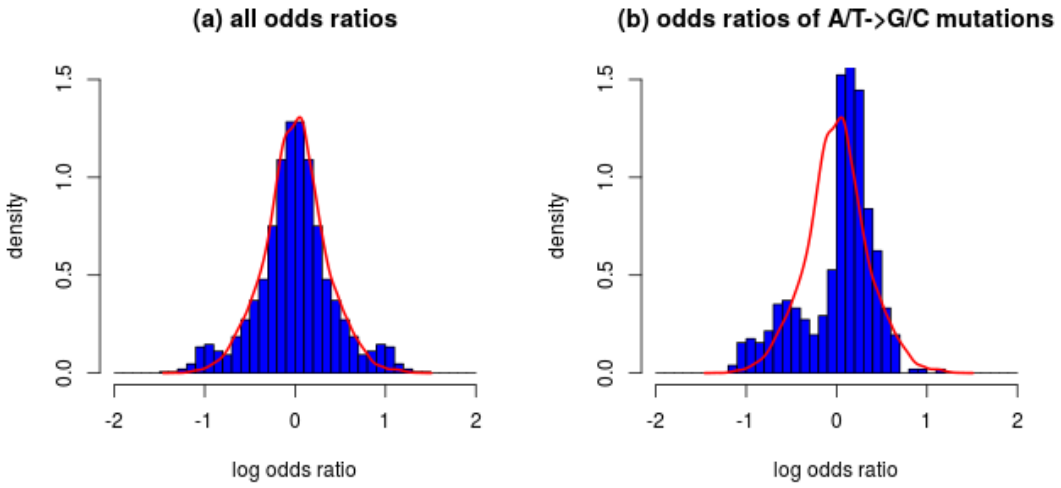

Figure 3.3: Distribution of odds ratios

Mutation patterns that have transitions and those having transversions have comparable odds distributions, with similar mean and mode values. Still, they also have somehow different dispersion around same mean. On the density estimation for all mutation patterns and histogram of transversions, a second smaller mode appears on the negative tail. This slightly bimodal distribution hints at the existence of a category of mutation patterns that are systematically less fixed and/or highly polymorphic.

3.1.3 Distributions of Odds Ratios

In a similar manner to the exploratory analysis above, I investigated the distribution of odds ratios among all 3072 mutation patterns. In order to demonstrate the sensitivity of the used method to fixation biases, I further studied the odds ratio distribution of mutation patterns expected to be affected by G/C biased gene conversion (see Introduction).

G/C biased gene conversion is a phenomenon independent of neighbouring sequence context, but it is expected to manifest itself also within pentamer sequence context, such that for each given set of context bases, a mutation pattern with $abA/Tcd- >$

abG/Ccd is expected to have higher odds relative to $abG/Ccd- > abA/Tcd$, where a-d stand for any nucleotide.

Figure 3.3, shows the data on a histogram, together with distribution of all odds ratios in the data and null distribution of odds ratios that is produced with permutations of all mutation patterns as compared patterns, representing an empirical null distribution where patterns are unrelated. Odds ratios of (a) all 3072 mutations patterns, and (b) 512 abA/Tcd -> abG/Ccd type mutations patterns are drawn in a histogram. Density estimation of an empirical null distribution is also shown on top of the histogram as a red line.

Distribution of all odds ratios (3.3(a)) fits fairly well to the empirical null distribution, suggesting that the compared mutation patterns have random-like odds ratios overall. On the tails of the observed OR distribution, however, there are small peaks; which is in fact one peak, because the distribution is symmetric. This peak points out to a small subset of mutation patterns with a distinct fixation behaviour from the rest of distribution. It is notable that excess of mutation patterns with higher ORs are not in the form of a fatter tail. That would be explained simply with a higher than expected variance of distribution. Instead, the observed distribution appears like the composition of samples from two populations, one major population complying to the null distribution and another minor population with a very distinct mean. Analysis on later parts of this chapter concentrates on characterization of that minor population.

Distribution of ORs for patterns affected by gBGC is concentrated on the positive side of the log odds ratio range (Figure 3.3(b)). As such, this distribution complies with the expectation of higher fixation odds for mutations creating a G/C allele with respect to those creating an A/T allele. The size of the effect may look small, and a somewhat conflicting fatter tail on the negative side is observed. It should be noted that gene conversion is dependent on recombination events and the fixation bias is strong enough to overcome genetic drift only on recombination hotspots (Duret and Galtier, 2009). Still it is notable that a fixation bias operating only on a small fraction of genome is capable of leaving a genome-wide signature.

3.2 Relation Between Fixation and Homopolymer Extension

A first inspection of Table 3.1 suggests that, homopolymer extending mutation patterns densely populate comparisons with high odds ratios. In order to reveal the relationship between odds ratio and homopolymerization score, I compared the odds ratio distribution for each score level using violin plots (Hintze and Nelson, 1998). A violin plot is a combination of a boxplot with a frequency density estimate (smoothed histogram). Quartiles from distribution are plotted as in box plot, together with actual data points and vertical curves, that are density estimations from actual data points.

The results shown in Figure 3.4(a) summarise McDonald - Kreitman comparisons that include all fixed or polymorphic mutations found anywhere in genome (except for repeat regions), and polymorphisms with a minor allele frequency threshold of 0.001. In following sections, I added McDonald - Kreitman comparisons that include mutations that are of specific type, on specific subsets of genome, and with varied minor allele frequency thresholds for polymorphisms, in order to investigate possible effects of these factors on relation of fixation probability and homopolymerization score of mutation patterns. In addition, in Figure 3.4, I used score type 1 (see Methods) as homopolymerization score, as it is more informative than other score types defined in the Methods section. The analyses are also repeated using different score types and the results compared in later sections.

In Figure 3.4(a), distribution of odds ratios across different score levels are compared, as described before. Violin plots in Figure 3.4(b-d) show the same distributions after mutation patterns are grouped according to the involved mutation type. This has an exploratory motivation, to check the effect of mutation type on the relationship between fixation and polymorphism. But we also know that mutation types have different properties. For instance, when using transversion type mutations, which occur at lower frequency than transitions, the inferred direction of mutation (correctly determining the derived state) is more reliable (see Introduction). Meanwhile, G/C biased gene conversion, as demonstrated in the previous section, influences fixation probability of mutations that involve A/T and G/C alleles. With this rationale, McDonald -

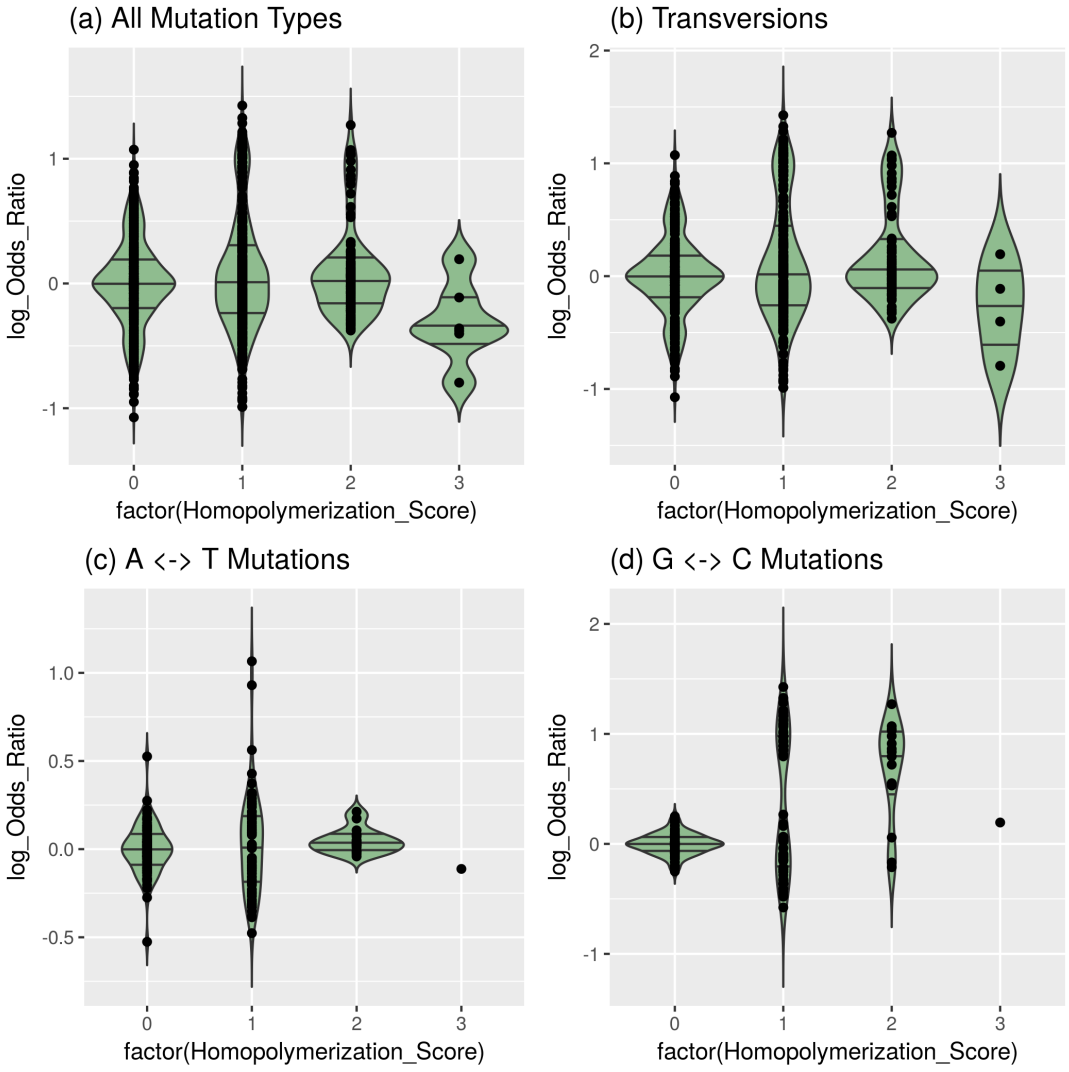

Figure 3.4: Homopolymerization - Odds ratio relationship

Kreitman comparisons between $A <-> T$ mutations and between $G <-> C$ mutations is a more reliable way of studying fixation-homopolymerization relationship, and additionally, these comparisons will not be affected by the known G/C biased gene conversion effect.

In this analysis I used a non-redundant set of data. For this, from all 3072 possible mutation pattern pair comparisons described before, I excluded self-to-self comparisons (due to complementary pattern and reverse pattern being identical), I resolved the two-fold appearance of mutation patterns due to joined complementary patterns by using only C or T as the focal mutated base, and I resolved the two-fold appearance of forward and reverse patterns by not including comparisons with negative homopolymerization scores. Removing duplicate comparisons is not required for score values of 0, as they are included as a control distribution, where the effect of homopolymerization does not exist. This yielded a total of 752 distinct comparisons. 348 of these have score level 1, 90 have score level 2, and 6 have score level 3. The number of comparisons involving transversion mutations for score levels 1, 2, and 3 are 232, 60 and 4 respectively, and for mutations between $G <-> C$ or $A <-> T$, these are 58, 15, and 1, respectively, for each mutation type.

Figure 3.4(d) shows a strongly positive relationship between odds ratio and homopolymerization score values for mutations between G and C. Such a relationship is very weak or absent for other mutation types (Figure 3.4(c)), and the trend observed in $G <-> C$ mutations is masked when all mutation types are involved (Figure 3.4(a)), due to random OR distributions of other mutation types. That mutations that create and/or extend G/C homopolymers may fix faster than expected given polymorphism is a new observation.

3.2.1 Comparison of Different Homopolymerization Definitions

Here I addressed whether the observed relationship between fixation and homopolymerisation depends on how exactly I defined homopolymerization. The homopolymerization effect of a mutation can be formalised in 4 alternative but related ways,

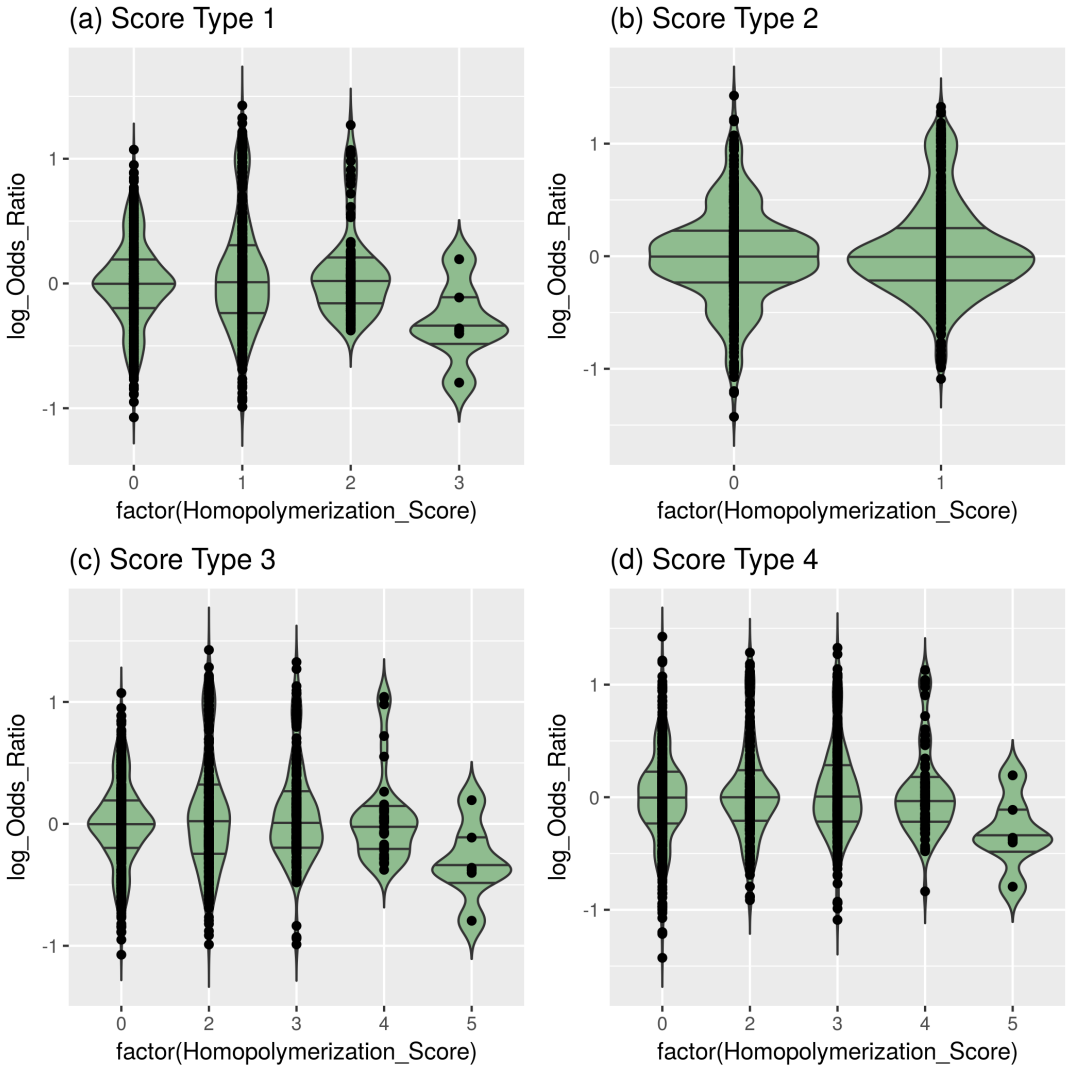

Figure 3.5: Effect of alternative homopolymerization definition

as described in Methods. Two alternative ways of defining a homopolymer are: either require the same base repeats to be strictly connected (as in score types 1 and 3), or consider more loosely the number of same bases within the pentamer sequence (as in score types 2 and 4). Another consideration is whether to use the change in homopolymer length (as in 1 and 2) or the homopolymer length itself (as in 3 and 4). Combination of these two factors with two alternative possibilities each, results in these 4 score types.

Figure 3.5 shows the effect of alternative homopolymerization score definition on the odds ratio vs. homopolymerization relationship. Score types do not have one-to-one correspondence between their levels. An inspection of their definitions, however, reveals that type 1's level 3 is equivalent to level 5 of type 3 and 4. Mutation patterns that belong to type 1's level 2, are split between type 3's levels 3 and 4, and also between type 4's levels 3, 4 with a few belonging to level 0. In a similar way, type 1's level 1 patterns reside in type 3's level 2 and 3, and type 4's levels 0, 3 and 4. Type 2 presents a broad classification of patterns as increasing/decreasing middle base's density within the pentamer. With these overlaps and distinctions, score types 1, 3 and 4 provide similar trends of OR-homopolymerization relationship. Distributions of corresponding or approximately corresponding score levels have very close summary statistics like quartiles, modes and skewness.

3.2.2 Effect of Polymorphic Mutations with Low Frequencies

In this section, I examined the sensitivity of my analysis to polymorphic mutations with low minor allele frequencies. It is possible that low frequency polymorphisms are more frequent in some mutation patterns than others in a systematic way, which could create a false positive fixation signal. For example, low frequency polymorphisms are in general more likely to be a sequencing error miscalled as a SNP. Sequencing error rates are known to vary among sequence patterns, so that possible sequencing errors would not only create random noise, but also a bias towards some mutation patterns having excess polymorphic occurrences. If this is the case, exclusion of low frequency polymorphisms should change the observed odds ratio distri-

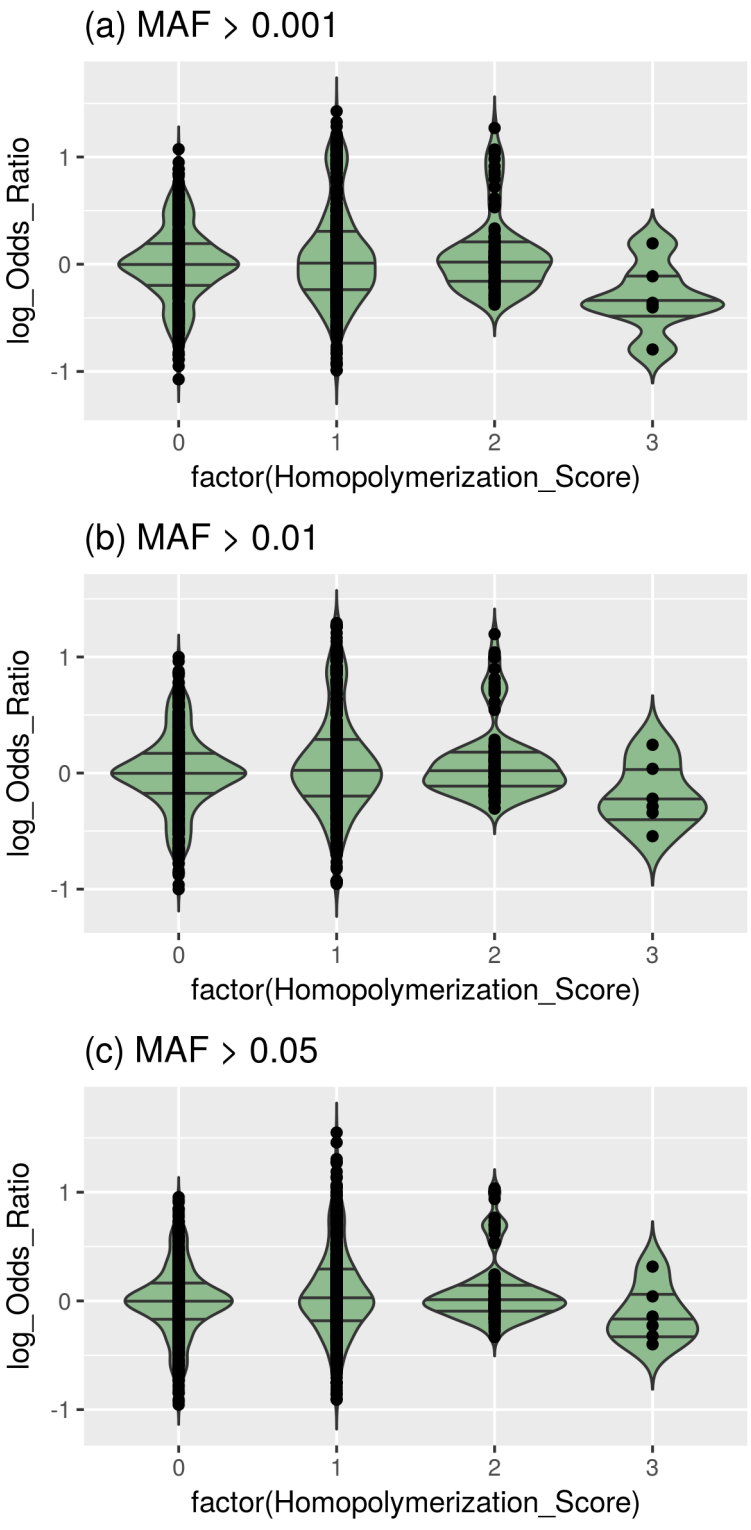

Figure 3.6: Effect of minor allele frequency threshold

butions compared to the original result in Figure 3.4(a).

I filtered polymorphic mutations with minor allele frequency thresholds of 0.001, 0.01 and 0.05, and repeated the McDonald - Kreitman analyses with the corresponding subsets of polymorphic mutations in the data. Figure 3.6 shows the resulting distributions, which indicates no qualitative change with respect to the use of filtering, in terms of number and amplitude of distribution density peaks, ranges and quartiles.

Analysis with minor allele frequency threshold of 0.001 is used throughout the study for convenience reasons, as it increases sample size.

3.2.3 Fixation on Exons and Promoters

The results discussed so far indicate that a fixation bias related to homopolymer extension may exist between G<->C mutations. This bias is not sensitive to how the homopolymer extension is defined, the inclusion/exclusion of low frequency polymorphisms, and as exploratory analysis suggests, it is likely not affected by mutation rate heterogeneity.

A natural question is then, whether this fixation bias is related to functional regions of the genome. Using mutations from only exonic regions and from only promoter regions, I repeated McDonald - Kreitman analyses for G<->C mutations, and the results are summarised in Figure 3.7(a-c). This showed that the trend of positive relation between fixation-homopolymerization in genome-wide data is not observed on mutations in exon and promoter regions. I discuss possible reasons for this effect in the Discussion section.

In order to directly demonstrate relationship between homopolymerization and odds ratio of fixation, I conducted simple linear regression analysis. Figure 3.7(d-f) shows regression lines and data points for patterns involving G<->C mutations. On genome wide data and among patterns involving G<->C mutation, homopolymerization score significantly explains variation in odds ratio of fixation (Adjusted R-squared: 0.3246 F-statistic: 58.2 on 1 and 118 DF, p-value: 6.695e-12).

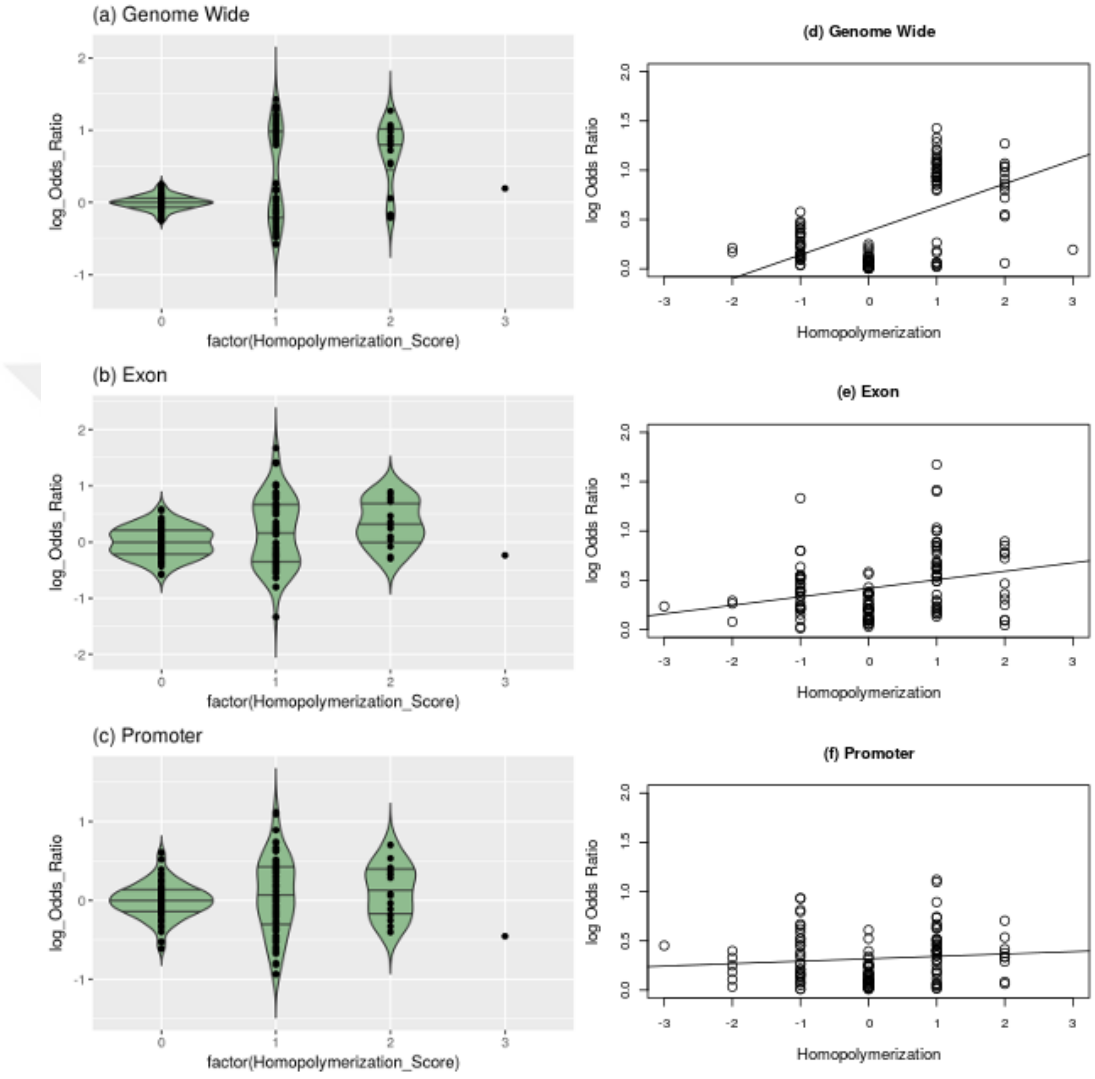

Figure 3.7: Effect of homopolymerization in exon and promoter regions

On exons and promoters, effect of homopolymerization is insignificant. These results are consistent with previous comparisons of odds ratio distributions among different homopolymerization levels. On right panel of Figure 3.7, score values range between [-3,3] and log odds ratio values range between [0,infinity]. These are different then the ranges used by violin plots, because in this case, the problem of redundant mutation pattern comparisons is solved by imposing constraint of log odds ratio > 1, in regression analysis. This is done so, to provide a higher range of values for the explanatory variable. Using previous approach of imposing constraint on comparisons with respect to homopolymerization scores (>= 0), also produced consistent results (not presented here).

3.3 Derived Allele Frequency Distribution

Distribution of allele frequencies provide information about fixation events of alleles of certain type or on certain genomic context. This method is independent of fixed allelic differences between species, hence independent of McDonald - Kreitman method. Here, as a way of confirming previously demonstrated fixation bias, I checked DAF distribution of 4 mutation patterns that have highest odds ratio values (all of them have positive homopolymerization scores and involve a G<->C mutation).

In figure 3.8, panels (a) and (c) compare these patterns' distributions with overall DAF distribution, whereas panels (b) and (d) compare these focal patterns' distributions with reverse patterns' distributions. As all of the distributions are rich in very low frequency alleles, removing positions with low minor allele frequencies increases resolution between distributions. Mutation patterns with high OR values have relatively lower frequency of low frequency derived alleles and slightly higher frequency of derived alleles with intermediate and high frequency. This observation is consistent with previous findings.

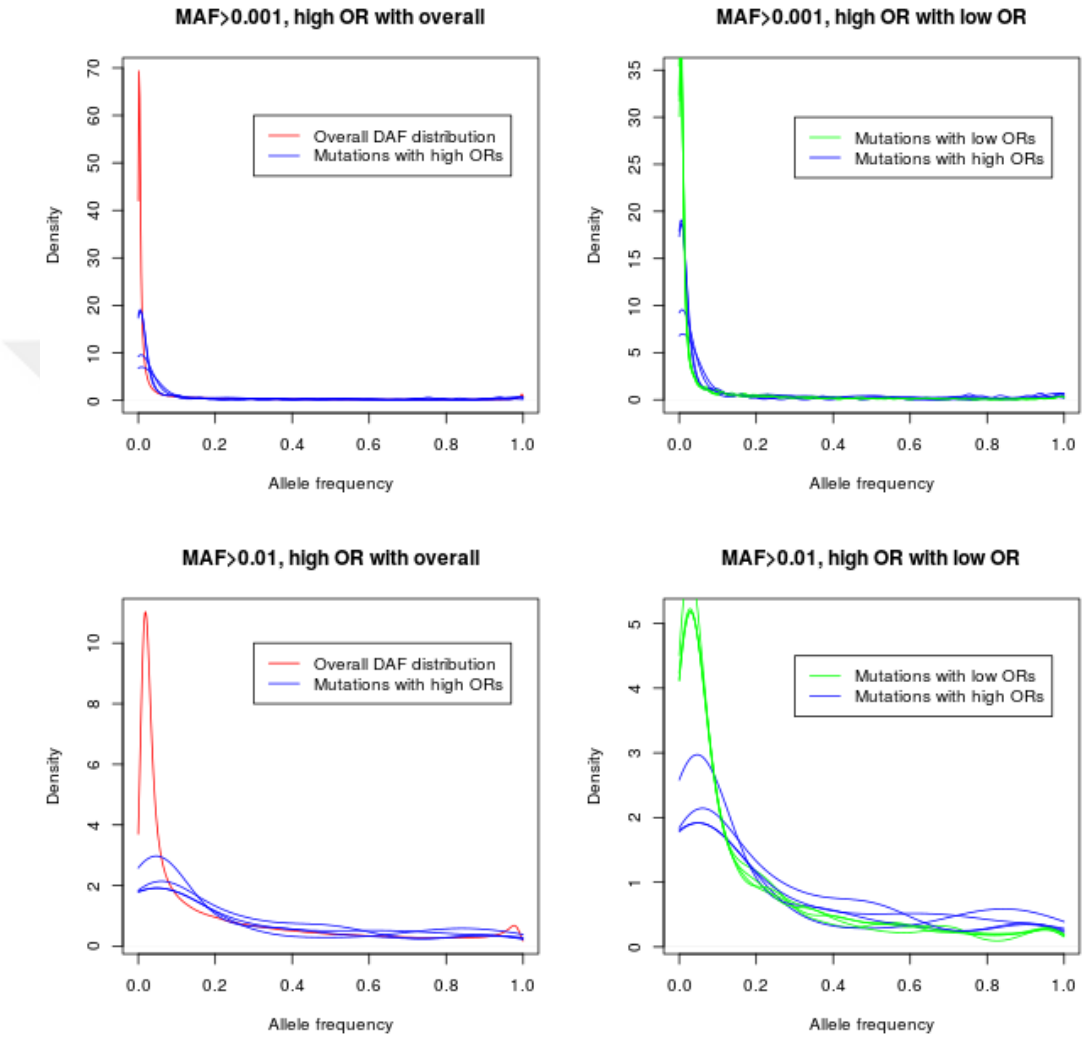

Figure 3.8: Comparison of derived allele frequency distributions

CHAPTER 4

DISCUSSION AND CONCLUSION

In this study, single nucleotide mutations in human genome are grouped regarding the pentamer sequence context they reside in. Fixation propensities of mutation types are tested by comparing ratio of fixation and polymorphism counts of the mutations belonging to that type, with the fixation/polymorphism ratio of the reverse mutation type. A fixation bias in favor of homopolymer-extending mutations is observed among mutation types involving a G<->C mutation. However, the observed fixation bias is absent in exons and promoters. A genome wide bias with no apparent connection to a biological function, at first suggests the presence of a technical artifact. Here, I discuss potential sources of error and measures I take to prevent their intervention. Then I proceed with a discussion of analysis methodology and possible explanations of the observed bias.

4.1 Potential Sources of Error

A systematic underestimation or overestimation of polymorphisms of a certain mutation pattern would easily create a signal of fixation bias against mutation pattern with overestimated polymorphism (and towards the pattern with underestimated polymorphism). One way to address this possibility is to analyse fixation and polymorphism counts of all comparisons, to see whether such an underestimation or overestimation pattern exists. Polymorphism counts are expected to reflect the background mutation frequency of the corresponding pattern. Assuming the mutation rate is equal, the mutation frequency of a pattern should be proportional to the frequency of the

sequence motif that could give rise to that mutation. The frequency of AGCGT->AGGGT mutations is proportional to frequency of AGCGT motifs in the genome, for instance. Also, fixation counts should be proportional to the mutation rate, under the assumption of selective neutrality: The more mutations of a pattern occur, the more chance there is for that pattern to drift to fixation. In supplementary analysis not presented here, I found correlation between fixation counts and polymorphism counts across all mutation patterns, as expected. I also found correlation between polymorphism counts with the frequency of the ancestral motif of the corresponding pattern in the whole genome. Then, I normalised fixation and polymorphism counts by dividing each by its ancestral motif frequency in the genome. This way, mutation frequencies are converted to a mutation rate per motif, and have closer values to each other. Principal component analysis of this fixation and polymorphism mutation rate dataset produced no distinct clusters, meaning that there are no groups of mutation patterns with distinctively high or low fixation and polymorphism frequencies (data not shown). Also, visually inspecting the data (e.g. Table 3.1), McDonald-Kreitman comparisons with high odds ratio values appear to include both cases with high reverse mutation pattern polymorphism counts, and low forward mutation pattern polymorphism counts. This analysis suggested that there is no anomaly in the data and a systematic excess of polymorphisms is not the reason for high odds ratio values.

4.1.1 Sequencing Errors

Many next generation sequencing (NGS) platforms are, alongside many benefits they possess, highly error-prone: For example Illumina HiSeq typically has an error rate an order of magnitude higher than one in thousand bases, the rate of Sanger sequencing (Manley et al., 2016). Error rates are also heterogeneous among sequence motifs and toward some bases (Nakamura et al., 2011). The comparative genomic dataset I used are of high quality, but the polymorphism dataset is NGS-based and of relatively moderate quality. Therefore, sequencing errors may introduce false positives in the polymorphism dataset, and these false positives may be biased, e.g. enriched in certain patterns.

On the other hand, sequencing flaws of NGS platforms are well studied and base reads are generated with quality scores reflecting confidence in the corresponding base. The 1000 Genomes Project Consortium includes in the output data polymorphisms that have sufficient quality scores from multiple independent sequencing processes. As a precaution, I also removed SNPs that are within repeat regions of the genome, as error rates on these regions are considerably higher. Moreover, a false positive variant caused by sequencing error is expected to have a low frequency within population. I therefore repeated McDonald-Kreitman tests using different thresholds of minor allele frequencies, where SNPs with low frequencies were excluded from analysis. This exclusion did not change the results, providing another layer of confidence regarding the polymorphism data.

Another way of inferring false polymorphisms is through alignment of reads derived from paralogous loci to a single locus, and treating mismatches between the paralogous sequences as alleles of the same locus. Genotypes inferred this way will be entirely heterozygotes, and the genotype distribution in such loci would deviate substantially from the distribution predicted by the Hardy-Weinberg principle. I applied a filter that excludes SNPs with such deviation. A filter of this kind is widely used to overcome the described source of error in polymorphism inference.

4.1.2 Ancestral State Inference Errors

The ancestral state of human alleles is determined by comparison with corresponding chimpanzee and orangutan alleles. Orthologous alleles for the 3 species are retrieved from 46-way multiz alignment. When chimpanzee and orangutan alleles match on a locus, that allele is assumed to be the common ancestral allele of human, chimpanzee and orangutan. In case of mismatch between chimpanzee and orangutan, corresponding positions are excluded from the study, as having an undetermined ancestral state. Using two outgroups is helpful to avoid the effect of homoplasy (Figure 4.1).

In order to exclude other ambiguous cases, whenever there were multiple mutations (fixed differences or human polymorphisms) residing within same pentamer frame,

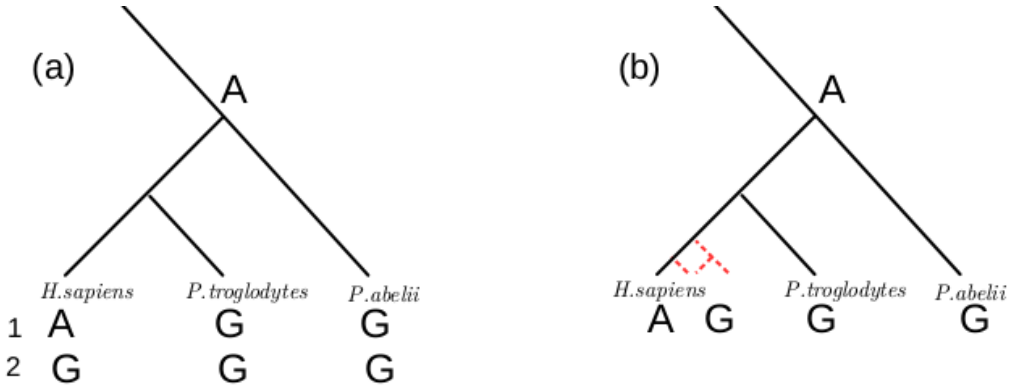

Figure 4.1: A hypothetical homoplasy event

we excluded that frame from the analysis. Likewise, polymorphisms with more than 2 alleles, and polymorphisms with neither of the alleles matching the inferred ancestral allele are excluded. These mutation filterings are expected to provide us a relatively reliable subset of data with respect to ancestral context motifs.

Figure 4.1 presents a hypothetical homoplasy event. Actual ancestral allele is an "A", recurrent substitutions to a "G" on chimpanzee and orangutan lineages would lead my method to infer the ancestral state as "G". This would then result in calling a G->A mutation instead of the actual A->G mutation. Increasing mutation rates increase probability of homoplasy, and different rates of mutation types would produce different amount of false positive mutation calls.

In the case of the McDonald-Kreitman test used here, if false calls affect polymorphism calls and fixation calls in equal proportions, they should cause no net effect in odds ratio values. In panel 4.1(a), a false G->A fixation is reported (1), and also an actual A->G fixation is missed (2). In panel 4.1(b), a G->A polymorphism is reported instead of the actual A->G polymorphism. As for each homoplasy background, occurrences of fixations and polymorphisms are expected to be proportional to their genome-wide occurrences, and therefore fixation/polymorphism ratio A->G and G->A mutations stays the same.

As such, errors due to homoplasy are not expected to create a bias towards fixation

of specific patterns, but will introduce noise in the data. Therefore, mutation patterns involving transversion mutations, which occur at lower rate and are less prone to homoplasy, are expected to have less noise caused by random ancestral allele misidentifications.

4.2 Alternative Methodology

Instead of comparing focal-reverse mutation pairs (e.g. abXcd -> abYcd vs. abYcd -> abXcd) as I did in this analysis, one could also use alternative approaches to study the effect of sequence context on mutation fixation rates. For example, all possible context 4mers for each one of the 12 nucleotide substitution types could be compared with respect to their distribution of fixation odds (e.g. abXcd -> abYcd vs. [a'bXcd -> a'bYcd, a'b'Xcd -> a'b'Ycd, ..., a'b'Xc'd' -> a'b'Yc'd']). This would provide a more direct measurement of the effect of context sequence on fixation probability. Then, contexts could be grouped as homopolymer extending and shortening ones and odds ratio for each of 12 mutation types computed. Consistent results are expected from both approaches, but there is one key reason for comparing forward and backward mutations with fixed context sequences. Pentamers in the mutation pattern pairs are the agents that are actually competing with each other in the evolutionary process. This way we can directly measure among allelic competitions, which pentamer motifs becomes fixed more often.

Another widely employed method for checking fixation bias is comparing derived allele frequency distribution of focal and reverse mutations (Nielsen and Slatkin, 2013). This method could also be used as the primary method for making all pairwise comparisons, using a summary statistic that describes derived allele frequencies distribution. We would expect higher derived allele frequencies for the mutation patterns with higher fixation odds ratios in the McDonald-Kreitman test. Thus, this approach can also be used as a tool to confirm significant results from McDonald-Kreitman test.

4.3 Source of Signal

If the observed bias is a real biological effect, rather than a technical one, its evolutionary and mechanistic cause or causes remain to be determined. Actually, unless a plausible biological association is identified, the possibility of a technical bias will persist. Biological effects can be of two types:

- Positive selection due to functional properties of G/C homopolymers,

- Gene conversion effects, arising from yet unknown mutation repair mechanism biases (such as biased gBGC).

4.3.1 Selection Related Mechanisms

To investigate a possible selective cause for the observed fixation bias, I searched the signal in functional regions, by checking exons and promoters. Perhaps surprisingly, a significant association between homopolymerization and fixation bias is observed throughout genome, but not in exons or promoters. In fact, it was previously reported that G and C homopolymers are deficient in coding regions in many organisms' genomes, possibly due to selection for stability. Repeats of more than 4 mononucleotides are prone to insertion/deletion mutations (Ackermann and Chao, 2006).

Hence the association should be enriched somewhere else in the genome. Enhancers and other cis and trans-acting regulatory sequences (e.g. insulators, non-coding RNA elements) outside the proximity of transcription start sites remain to be checked.

One possibility is a link with G-quadruplexes. This is worth investigating, as the observed bias is strongly, if not exclusively, associated with mutation patterns creating G or C homopolymers. G/C homopolymers are the defining motifs of G-quadruplexes, which are suggested to play a role in transcription factor binding or origin of replication by opening the chromatin (Huppert and Balasubramanian, 2007), (Das et al., 2016), (Chambers et al., 2015). However, these structures are mostly in promoters,

and absence of fixation bias in promoters reduces probability of the bias being related to G-quadruplexes.

Another biological explanation may be that the identified G/C homopolymers are linked to structural properties of DNA, such as interaction with histones. Checking for clustering of newly fixed G/C homopolymers around other G/C homopolymer motifs, would give a clue about a link to other possible structural roles, as trimer motifs are not expected to have structural effects by themselves and would require clustering of G/C homopolymer motifs.

4.3.2 Non-selective Mechanisms

A mechanistic explanation of a fixation bias as described here, without a functional link that would imply positive selection, should involve some kind of biased gene conversion effect. An enrichment of fixed G/C homopolymers on recombination hotspots would support this scenario. During homologous recombination, whenever paternal and maternal loci have different alleles, mismatches occur at gene conversion tracts and if these are repaired in a biased fashion, the result is a fixation bias. Such events happen at recombination hotspots at much higher rates relative to the rest of genome.If a context sequence-biased repair machinery exists and is responsible for the observed effect, this would be noticeable in hotspot regions.

Assuming such a biased gene conversion is the cause of homopolymer fixation, regions with higher recombination rates should possess higher number of biased fixations relative to low recombination regions, and also relative to regions under the effect of strong negative selection. Effective population size among loci within genome may vary considerably. Loci that are subject to background selection have lower effective population size (fewer variants contribute to gene pool) and are more likely to have an allele fixed randomly. In contrast, at loci with higher recombination rates, independently segregating alleles have higher chance of escaping elimination by genetic drift and higher chance of increasing in frequency by gene conversion. However, my preliminary analysis suggests no enrichment for homopolymer fixation in recom-

bination hotspots (data not shown).

Another assumption of McDonald-Kreitman test not explicitly stated is that mutation rates are stable since the last common ancestor of species involved in the analysis. Fixations may have occurred anytime since the last common ancestor, but polymorphisms are expected to be of a relatively recent origin. Hence, polymorphisms, which are used as a normalising factor against difference in mutation rates, actually reflect recent mutation rates. If some mutation patterns have undergone an increase in their mutation rates, McDonald-Kreitman test would report a fixation bias against that mutation pattern. This would not be an actual fixation bias, of course, but would still point out a biologically significant event.

4.4 Conclusion

A genome-wide scan revealed a fixation bias towards G/C homopolymers through G<->C mutations. Positive selection for G/C homopolymers with a functional link to G-quadruplex structures, which could be a potential explanation, turned out to be not very likely. A gene conversion effect in conjunction with recombination does not seem likely, either. A mechanistic explanation for this fixation bias remains to be explored.

REFERENCES

Ackermann, M. and Chao, L. (2006). DNA Sequences Shaped by Selection for Stability. *PLoS Genetics*, 2(2):e22.

Andolfatto, P. (2005). Adaptive evolution of non-coding DNA in Drosophila. *Nature*, 437(7062):1149–1152.

Auton, A., Abecasis, G. R., Altshuler, D. M., and Durbin, Richard M., . J. A. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1):289–300.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F. A., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., Haussler, D., and Miller, W. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome research*, 14(4):708–15.

Casillas, S. and Barbadilla, A. (2017). Molecular Population Genetics. *Genetics*, 205(3).

Chambers, V. S., Marsico, G., Boutell, J. M., Di Antonio, M., Smith, G. P., and Balasubramanian, S. (2015). High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nature Biotechnology*, 33(8):877–881.

Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C., and Patrinos, G. P. (2007). Gene conversion: mechanisms, evolution and human disease. *Nature Reviews Genetics*, 8(10):762–775.

Das, K., Srivastava, M., Raghavan, S. C., Hsieh, C., and Lieber, M. (2016). GNG Motifs Can Replace a GGG Stretch during G-Quadruplex Formation in a Context Dependent Manner. *PLOS ONE*, 11(7):e0158794.

Duret, L. and Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual review of genomics and human genetics*, 10:285–311.

Ehrlich, M. and Wang, R. (1981). 5-Methylcytosine in eukaryotic DNA. *Science*, 212(4501):1350–1357.

Fay, J. C. and Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, 155(3):1405–13.

Fitch, W. M. (1967). Evidence suggesting a non-random character to nucleotide replacements in naturally occurring mutations. *Journal of Molecular Biology*, 26(3):499–507.

Hamilton, M. B. (2009). *Population genetics*. Wiley-Blackwell.

Hintze, J. L. and Nelson, R. D. (1998). Violin Plots: A Box Plot-Density Trace Synergism Statistical Computing and Graphics Violin Plots: A Box Plot-Density Trace Synergism. *Source: The American Statistician*, 52(2):181–184.

Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B. M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., Ni, P., Wang, B., Ou, X., Huasang, Luosang, J., Cuo, Z. X. P., Li, K., Gao, G., Yin, Y., Wang, W., Zhang, X., Xu, X., Yang, H., Li, Y., Wang, J., Wang, J., and Nielsen, R. (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, 512(7513):194–197.

Huppert, J. L. and Balasubramanian, S. (2007). G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Research*, 35(2):406–413.

Jones, E., Oliphant, T., Peterson, P., et al. (2001–). SciPy: Open source scientific tools for Python. [Online; accessed <today>].

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., and Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic acids research*, 32(Database issue):D493–6.

Kimura, M. (1968). Evolutionary Rate at the Molecular Level. *Nature*, 217(5129):624–626.

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press.

Kimura, M. (1991). The neutral theory of molecular evolution: a review of recent evidence.

Krawczak, M., Ball, E. V., and Cooper, D. N. (1998). Neighboring-Nucleotide Effects on the Rates of Germ-Line Single-Base-Pair Substitution in Human Genes. *The American Journal of Human Genetics*, 63(2):474–488.

Lawrie, D. S., Messer, P. W., Hershberg, R., and Petrov, D. A. (2013). Strong Purifying Selection at Synonymous Sites in D. melanogaster. *PLoS Genetics*, 9(5):e1003527.

Lynch, M. (2007). *The Origins of Genome Architecture 2007*, volume 302. Sinauer Associates.

Manley, L. J., Ma, D., and Levine, S. S. (2016). Monitoring Error Rates In Illumina Sequencing. *Journal of biomolecular techniques : JBT*, 27(4):125–128.

McCandlish, D. M. and Stoltzfus, A. (2014). Modeling Evolution Using the Probability of Fixation: History and Implications. *The Quarterly Review of Biology*, 89(3):225–252.

McDonald, J. H. and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, 351(6328):652–654.

McVean, G. A., Altshuler (Co-Chair), D. M., Durbin (Co-Chair), R. M., Abecasis, Marth, G. T., and McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M. C., Hirai, A., Takahashi, H., Altaf-Ul-Amin, M., Ogasawara, N., and Kanaya, S. (2011). Sequence-specific error profile of Illumina sequencers. *Nucleic acids research*, 39(13):e90.

Nei, M. (1987). *Molecular evolutionary genetics*. Columbia University Press.

Nielsen, R. and Slatkin, M. (2013). *An introduction to population genetics : theory and applications*. Sinauer Associates.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6):841–2.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Sawyer, S. A. and Hartl, D. L. (1992). Population genetics of polymorphism and divergence. *Genetics*, 132(4).

Schuster, S. C. (2008). Next-generation sequencing transforms today ' s biology. *Nature methods*, 5(1):16–18.

Service, S. K., Teslovich, T. M., Fuchsberger, C., Ramensky, V., Yajnik, P., Koboldt, D. C., Larson, D. E., Zhang, Q., Lin, L., Welch, R., Ding, L., McLellan, M. D., O'Laughlin, M., Fronick, C., Fulton, L. L., Magrini, V., Swift, A., Elliott, P., Jarvelin, M.-R., Kaakinen, M., McCarthy, M. I., Peltonen, L., Pouta, A., Bonnycastle, L. L., Collins, F. S., Narisu, N., Stringham, H. M., Tuomilehto, J., Ripatti, S., Fulton, R. S., Sabatti, C., Wilson, R. K., Boehnke, M., and Freimer, N. B. (2014). Re-sequencing Expands Our Understanding of the Phenotypic Impact of Variants at GWAS Loci. *PLoS Genetics*, 10(1):e1004147.

Sethupathy, P. and Hannenhalli, S. (2008). A tutorial of the poisson random field model in population genetics. *Advances in bioinformatics*, 2008:257864.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595.

Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics*, pages 1–9.

Vogel, F. (1972). Non-Randomness of Base Replacement in Point Mutation. *J. molec. Evolution*, 1:334–367.

Wall, J. D. and Slatkin, M. (2012). Paleopopulation Genetics. *Annual Review of Genetics*, 46(1):635–649.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591.

Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Keenan, S., Lavidas, I., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Nuhn, M., Parker, A., Patricio, M., Pignatelli, M., Rahtz, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Birney, E., Harrow, J., Muffato, M., Perry, E., Ruffier, M., Spudich, G., Trevanion, S. J., Cunningham, F., Aken, B. L., Zerbino, D. R., and Flicek, P. (2016). Ensembl 2016. *Nucleic Acids Research*, 44(D1):D710–D716.

APPENDIX

Table A.1: McDonald-Kreitman test results

Mutation pattern	fix1	poly1	fix2	poly2	OR	pval	s1
TACGC->TAGGC + GCGTA->GCCTA vs TAGGC->TACGC + GCCTA->GCGTA	163	186	368	1,750	4.16	1.55E-29	1
TACGG->TAGGG + CCGTA->CCCTA vs TAGGG->TACGG + CCCTA->CCGTA	171	161	858	3,047	3.77	3.17E-28	1
TTCGT->TTGGT + ACGAA->ACCAA vs TTGGT->TTCGT + ACCAA->ACGAA	319	333	1,197	4,518	3.61	1.52E-48	1
TGCGT->TGGGT + ACGCA->ACCCA vs TGGGT->TGCGT + ACCCA->ACGCA	312	327	1,082	4,041	3.56	4.60E-46	2
TTCGC->TTGGC + GCGAA->GCCAA vs TTGGC->TTCGC + GCCAA->GCGAA	182	211	510	1,993	3.37	3.22E-25	1
CTCGT->CTGGT + ACGAG->ACCAG vs CTGGT->CTCGT + ACCAG->ACGAG	431	472	1,045	3,793	3.31	3.44E-54	1
CGTCA->CGGCA + TGACG->TGCCG vs CGGCA->CGTCA + TGCCG->TGACG	55	39	95	221	3.27	1.84E-06	1
TACGA->TAGGA + TCGTA->TCCTA vs TAGGA->TACGA + TCCTA->TCGTA	265	299	933	3,369	3.20	4.78E-34	1
TACGA->TAAGA + TCGTA->TCTTA vs TAAGA->TACGA + TCTTA->TCGTA	309	336	899	3,053	3.12	1.71E-36	1
GTCGG->GTGGG + CCGAC->CCCAC vs GTGGG->GTCGG + CCCAC->CCGAC	276	316	908	3,218	3.09	2.38E-33	1
GACGG->GAAGG + CCGTC->CCTTC vs GAAGG->GACGG + CCTTC->CCGTC	323	368	724	2,514	3.05	1.61E-35	1
TACGT->TAGGT + ACGTA->ACCTA vs TAGGT->TACGT + ACCTA->ACGTA	297	371	938	3,568	3.04	1.75E-35	1
AACGG->AAAGG + CCGTT->CCTTT vs AAAGG->AACGG + CCTTT->CCGTT	408	474	883	3,062	2.98	9.27E-43	1
AACGT->AAGGT + ACGTT->ACCTT vs AAGGT->AACGT + ACCTT->ACGTT	379	451	1,248	4,422	2.98	7.35E-43	1
GACGA->GAAGA + TCGTC->TCTTC vs GAAGA->GACGA + TCTTC->TCGTC	276	347	864	3,233	2.98	7.06E-32	1
CTCGC->CTGGC + GCGAG->GCCAG vs CTGGC->CTCGC + GCCAG->GCGAG	414	498	524	1,875	2.97	3.95E-38	1
TGCGT->TGAGT + ACGCA->ACTCA vs TGAGT->TGCGT + ACTCA->ACGCA	506	604	671	2,343	2.92	1.71E-45	0
GTCGC->GTGGC + GCGAC->GCCAC vs GTGGC->GTCGC + GCCAC->GCGAC	212	249	245	842	2.92	5.69E-19	1
AACGG->AAGGG + CCGTT->CCCTT vs AAGGG->AACGG + CCCTT->CCGTT	256	283	1,506	4,857	2.92	1.11E-29	1
AGCGT->AGGGT + ACGCT->ACCCT vs AGGGT->AGCGT + ACCCT->ACGCT	274	357	1,319	5,012	2.92	1.32E-32	2
CGTAG->CGAAG + CTACG->CTTCG vs CGAAG->CGTAG + CTTCG->CTACG	36	40	16	52	2.90	5.49E-03	1
GACGC->GAGGC + GCGTC->GCCTC vs GAGGC->GACGC + GCCTC->GCGTC	222	290	711	2,686	2.89	4.76E-25	1
TGCGG->TGGGG + CCGCA->CCCCA vs TGGGG->TGCGG + CCCCA->CCGCA	246	307	1,788	6,325	2.83	1.33E-28	2
TACGT->TAAGT + ACGTA->ACTTA vs TAAGT->TACGT + ACTTA->ACGTA	429	452	1,033	3,082	2.83	2.55E-40	1
GACGT->GAAGT + ACGTC->ACTTC vs GAAGT->GACGT + ACTTC->ACGTC	323	382	873	2,898	2.81	5.49E-32	1
AGCGG->AGGGG + CCGCT->CCCCT vs AGGGG->AGCGG + CCCCT->CCGCT	278	387	1,564	6,106	2.80	7.16E-32	2
TACGC->TAAGC + GCGTA->GCTTA vs TAAGC->TACGC + GCTTA->GCGTA	179	222	382	1,324	2.79	1.05E-17	1
GTCGT->GTGGT + ACGAC->ACCAC vs GTGGT->GTCGT + ACCAC->ACGAC	187	257	527	2,018	2.79	8.10E-20	1
GACGG->GAGGG + CCGTC->CCCTC vs GAGGG->GACGG + CCCTC->CCGTC	217	271	1,499	5,156	2.75	4.83E-24	1
ATCGA->ATGGA + TCGAT->TCCAT vs ATGGA->ATCGA + TCCAT->TCGAT	431	515	358	1,173	2.74	2.46E-29	1
AACGC->AAAGC + GCGTT->GCTTT vs AAAGC->AACGC + GCTTT->GCGTT	345	406	663	2,136	2.74	1.92E-30	1
GACGC->GAAGC + GCGTC->GCTTC vs GAAGC->GACGC + GCTTC->GCGTC	283	345	526	1,750	2.73	4.68E-25	1
ACCGG->ACAGG + CCGGT->CCTGT vs ACAGG->ACCGG + CCTGT->CCGGT	316	403	636	2,182	2.69	5.65E-28	-1
CCCGT->CCAGT + ACGGG->ACTGG vs CCAGT->CCCGT + ACTGG->ACGGG	370	428	893	2,776	2.69	1.32E-32	-1
GGCGT->GGGGT + ACGCC->ACCCC vs GGGGT->GGCGT + ACCCC->ACGCC	165	217	668	2,343	2.67	8.34E-17	2
TTCGA->TTGGA + TCGAA->TCCAA vs TTGGA->TTCGA + TCCAA->TCGAA	316	387	745	2,420	2.65	8.75E-28	1
CACGG->CAAGG + CCGTG->CCTTG vs CAAGG->CACGG + CCTTG->CCGTG	651	777	644	2,032	2.64	2.99E-43	1
AACGC->AAGGC + GCGTT->GCCTT vs AAGGC->AACGC + GCCTT->GCGTT	217	285	769	2,664	2.64	4.56E-21	1
CACGG->CAGGG + CCGTG->CCCTG vs CAGGG->CACGG + CCCTG->CCGTG	464	582	1,560	5,155	2.63	2.48E-42	1
TACGC->TATGC + GCGTA->GCATA vs TATGC->TACGC + GCATA->GCGTA	3,620	4,940	4,344	15,328	2.59	2.01E-251	0
CGTAT->CGGAT + ATACG->ATCCG vs CGGAT->CGTAT + ATCCG->ATACG	24	33	46	163	2.57	6.05E-03	1
CCCGA->CCAGA + TCGGG->TCTGG vs CCAGA->CCCGA + TCTGG->TCGGG	420	483	710	2,090	2.56	7.46E-31	-1
CACGC->CAGGC + GCGTG->GCCTG vs CAGGC->CACGC + GCCTG->GCGTG	402	573	789	2,878	2.56	1.11E-32	1
GTCGA->GTGGA + TCGAC->TCCAC vs GTGGA->GTCGA + TCCAC->TCGAC	272	317	295	880	2.56	4.62E-18	1
GACGA->GAGGA + TCGTC->TCCTC vs GAGGA->GACGA + TCCTC->TCGTC	270	346	1,155	3,781	2.55	1.40E-24	1
ATCGT->ATGGT + ACGAT->ACCAT vs ATGGT->ATCGT + ACCAT->ACGAT	270	385	659	2,393	2.55	3.51E-23	1
GACGT->GAGGT + ACGTC->ACCTC vs GAGGT->GACGT + ACCTC->ACGTC	193	257	911	3,083	2.54	4.00E-18	1
CGTAA->CGAAA + TTACG->TTTCG vs CGAAA->CGTAA + TTTCG->TTACG	35	54	15	59	2.53	1.62E-02	1
TACGG->TAAGG + CCGTA->CCTTA vs TAAGG->TACGG + CCTTA->CCGTA	220	287	613	2,026	2.53	7.36E-19	1
CCCGG->CCAGG + CCGGG->CCTGG vs CCAGG->CCCGG + CCTGG->CCGGG	564	712	856	2,733	2.53	4.80E-40	-1
AACGA->AAGGA + TCGTT->TCCTT vs AAGGA->AACGA + TCCTT->TCGTT	451	492	1,603	4,398	2.51	3.71E-36	1
GCCGT->GCAGT + ACGGC->ACTGC vs GCAGT->GCCGT + ACTGC->ACGGC	290	354	883	2,690	2.50	1.39E-23	-1
TGCGC->TGGGC + GCGCA->GCCCA vs TGGGC->TGCGC + GCCCA->GCGCA	181	262	703	2,534	2.49	2.18E-16	2
AACGA->AAAGA + TCGTT->TCTTT vs AAAGA->AACGA + TCTTT->TCGTT	470	607	1,335	4,265	2.47	4.23E-37	1
CTCGG->CTGGG + CCGAG->CCCAG vs CTGGG->CTCGG + CCCAG->CCGAG	568	789	1,137	3,856	2.44	3.66E-41	1
TTCGG->TTGGG + CCGAA->CCCAA vs TTGGG->TTCGG + CCCAA->CCGAA	218	295	1,007	3,327	2.44	6.38E-19	1
CACGA->CAGGA + TCGTG->TCCTG vs CAGGA->CACGA + TCCTG->TCGTG	509	681	1,036	3,376	2.44	1.20E-36	1
TGCGC->TGAGC + GCGCA->GCTCA vs TGAGC->TGCGC + GCTCA->GCGCA	442	614	340	1,149	2.43	1.24E-23	0
TACGT->TATGT + ACGTA->ACATA vs TATGT->TACGT + ACATA->ACGTA	6,603	8,711	9,867	31,553	2.42	0.00E+00	0
TCCGT->TCAGT + ACGGA->ACTGA vs TCAGT->TCCGT + ACTGA->ACGGA	247	302	1,365	4,020	2.41	2.32E-20	-1
AGCGA->AGGGA + TCGCT->TCCCT vs AGGGA->AGCGA + TCCCT->TCGCT	342	439	1,070	3,259	2.37	2.63E-25	2

Table A.1 (continued)

ATCGC->ATGGC + GCGAT->GCCAT vs ATGGC->ATCGC + GCCAT->GCGAT	217	283	284	873	2.36	1.94E-13	1
CACGT->CAGGT + ACGTG->ACCTG vs CAGGT->CACGT + ACCTG->ACGTG	513	745	1,407	4,787	2.34	8.98E-37	1
TACGG->TATGG + CCGTA->CCATA vs TATGG->TACGG + CCATA->CCGTA	4,647	6,524	6,143	20,109	2.33	9.57E-265	0
AGCGT->AGAGT + ACGCT->ACTCT vs AGAGT->AGCGT + ACTCT->ACGCT	441	630	808	2,673	2.32	2.14E-28	0
CCCGC->CCAGC + GCGGG->GCTGG vs CCAGC->CCCGC + GCTGG->GCGGG	556	763	628	1,989	2.31	4.44E-30	-1
TGCGA->TGGGA + TCGCA->TCCCA vs TGGGA->TGCGA + TCCCA->TCGCA	236	355	1,048	3,637	2.31	2.01E-18	2
ACCGT->ACAGT + ACGGT->ACTGT vs ACAGT->ACCGT + ACTGT->ACGGT	309	435	1,137	3,635	2.27	5.94E-22	-1
AGCGG->AGAGG + CCGCT->CCTCT vs AGAGG->AGCGG + CCTCT->CCGCT	566	901	647	2,339	2.27	7.26E-31	0
ATCGG->ATGGG + CCGAT->CCCAT vs ATGGG->ATCGG + CCCAT->CCGAT	226	343	466	1,593	2.25	8.28E-15	1
TGTCG->TGGCG + CGACA->CGCCA vs TGGCG->TGTCG + CGCCA->CGACA	29	32	78	193	2.24	9.96E-03	1
CTCGA->CTGGA + TCGAG->TCCAG vs CTGGA->CTCGA + TCCAG->TCGAG	593	792	579	1,714	2.22	2.99E-27	1
AGCGC->AGGGC + GCGCT->GCCCT vs AGGGC->AGCGC + GCCCT->GCGCT	286	415	692	2,224	2.21	3.93E-18	2
CCCCC->CCACC + GGGGG->GGTGG vs CCACC->CCCCC + GGTGG->GGGGG	1,285	2,004	862	2,976	2.21	3.93E-51	-3
GCCGG->GCAGG + CCGGC->CCTGC vs GCAGG->GCCGG + CCTGC->CCGGC	435	614	761	2,374	2.21	1.07E-24	-1
CACGA->CAAGA + TCGTG->TCTTG vs CAAGA->CACGA + TCTTG->TCGTG	475	627	739	2,110	2.16	3.96E-24	1
TGCGA->TGAGA + TCGCA->TCTCA vs TGAGA->TGCGA + TCTCA->TCGCA	507	725	753	2,323	2.16	8.59E-26	0
CACGC->CAAGC + GCGTG->GCTTG vs CAAGC->CACGC + GCTTG->GCGTG	484	719	371	1,187	2.15	1.38E-19	1
ACCGA->ACAGA + TCGGT->TCTGT vs ACAGA->ACCGA + TCTGT->TCGGT	344	491	964	2,959	2.15	1.28E-20	-1
TGCGG->TGAGG + CCGCA->CCTCA vs TGAGG->TGCGG + CCTCA->CCGCA	698	1,021	570	1,773	2.13	2.02E-27	0
GTCGT->GTAGT + ACGAC->ACTAC vs GTAGT->GTCGT + ACTAC->ACGAC	185	228	378	970	2.08	1.12E-09	0
GGCGA->GGGGA + TCGCC->TCCCC vs GGGGA->GGCGA + TCCCC->TCGCC	195	300	818	2,587	2.06	6.86E-12	2
TACGA->TATGA + TCGTA->TCATA vs TATGA->TACGA + TCATA->TCGTA	5,385	7,618	7,243	20,858	2.04	2.12E-216	0
AACGT->AAAGT + ACGTT->ACTTT vs AAAGT->AACGT + ACTTT->ACGTT	626	819	1,626	4,295	2.02	1.25E-29	1
GGCGT->GGAGT + ACGCC->ACTCC vs GGAGT->GGCGT + ACTCC->ACGCC	355	516	573	1,681	2.02	5.81E-16	0
TGTAT->TGGAT + ATACA->ATCCA vs TGGAT->TGTAT + ATCCA->ATACA	864	1,184	1,808	4,979	2.01	1.99E-38	1
CGTCT->CGGCT + AGACG->AGCCG vs CGGCT->CGTCT + AGCCG->AGACG	45	60	117	313	2.00	4.86E-03	1
CGTCT->CGCCT + AGACG->AGGCG vs CGCCT->CGTCT + AGGCG->AGACG	270	337	376	937	2.00	4.96E-11	1
TCCGG->TCAGG + CCGGA->CCTGA vs TCAGG->TCCGG + CCTGA->CCGGA	204	311	803	2,441	1.99	2.27E-11	-1
CGTTA->CGGTA + TAACG->TACCG vs CGGTA->CGTTA + TACCG->TAACG	40	40	48	96	1.99	2.40E-02	0
ACCGC->ACAGC + GCGGT->GCTGT vs ACAGC->ACCGC + GCTGT->GCGGT	286	440	498	1,526	1.99	3.99E-13	-1
CGTAC->CGGAC + GTACG->GTCCG vs CGGAC->CGTAC + GTCCG->GTACG	14	19	45	120	1.96	1.29E-01	1
GGTAT->GGGAT + ATACC->ATCCC vs GGGAT->GGTAT + ATCCC->ATACC	310	493	1,345	4,173	1.95	5.16E-16	1
AGTAT->AGGAT + ATACT->ATCCT vs AGGAT->AGTAT + ATCCT->ATACT	995	1,412	1,958	5,413	1.95	4.04E-40	1
AGCGT->AGTGT + ACGCT->ACACT vs AGTGT->AGCGT + ACACT->ACGCT	4,586	6,279	4,539	12,053	1.94	6.66E-141	0
AGCGA->AGAGA + TCGCT->TCTCT vs AGAGA->AGCGA + TCTCT->TCGCT	538	827	1,015	2,998	1.92	8.00E-22	0
CTCGT->CTAGT + ACGAG->ACTAG vs CTAGT->CTCGT + ACTAG->ACGAG	227	301	572	1,454	1.92	5.65E-10	0
GGCGA->GGAGA + TCGCC->TCTCC vs GGAGA->GGCGA + TCTCC->TCGCC	455	702	698	2,062	1.91	2.11E-17	0
AGCGC->AGAGC + GCGCT->GCTCT vs AGAGC->AGCGC + GCTCT->GCGCT	472	780	499	1,578	1.91	2.95E-16	0
TCCGT->TCTGT + ACGGA->ACAGA vs TCTGT->TCCGT + ACAGA->ACGGA	5,723	7,854	4,495	11,792	1.91	1.63E-151	-1
GGCGG->GGAGG + CCGCC->CCTCC vs GGAGG->GGCGG + CCTCC->CCGCC	738	1,237	727	2,308	1.89	2.50E-23	0
CGCTT->CGATT + AAGCG->AATCG vs CGATT->CGCTT + AATCG->AAGCG	80	107	26	66	1.89	3.78E-02	0
TATAT->TAGAT + ATATA->ATCTA vs TAGAT->TATAT + ATCTA->ATATA	1,510	1,870	1,727	4,038	1.89	3.24E-44	0
AACGG->AATGG + CCGTT->CCATT vs AATGG->AACGG + CCATT->CCGTT	6,676	9,831	7,410	20,557	1.88	4.96E-200	0
ACCGT->ACTGT + ACGGT->ACAGT vs ACTGT->ACCGT + ACAGT->ACGGT	4,364	6,097	4,546	11,887	1.87	5.09E-123	-1
AACGT->AATGT + ACGTT->ACATT vs AATGT->AACGT + ACATT->ACGTT	8,656	12,064	11,122	28,996	1.87	2.05E-262	0
GTTAC->GTGAC + GTAAC->GTCAC vs GTGAC->GTTAC + GTCAC->GTAAC	571	623	608	1,235	1.86	1.43E-15	-1
CACGT->CAAGT + ACGTG->ACTTG vs CAAGT->CACGT + ACTTG->ACGTG	684	1,020	986	2,729	1.86	1.24E-22	1
TCCGA->TCAGA + TCGGA->TCTGA vs TCAGA->TCCGA + TCTGA->TCGGA	160	247	1,124	3,219	1.85	4.17E-08	-1
CGTGA->CGGGA + TCACG->TCCCG vs CGGGA->CGTGA + TCCCG->TCACG	84	120	59	156	1.85	6.34E-03	2
GACGG->GATGG + CCGTC->CCATC vs GATGG->GACGG + CCATC->CCGTC	6,859	10,324	5,319	14,712	1.84	3.53E-163	0
TTCGT->TTTGT + ACGAA->ACAAA vs TTTGT->TTCGT + ACAAA->ACGAA	5,337	7,382	5,441	13,752	1.83	6.49E-137	1
TTCGT->TTAGT + ACGAA->ACTAA vs TTAGT->TTCGT + ACTAA->ACGAA	288	383	1,122	2,720	1.82	1.62E-11	0
TGCGT->TGTGT + ACGCA->ACACA vs TGTGT->TGCGT + ACACA->ACGCA	4,896	6,875	4,576	11,707	1.82	1.44E-120	0
TGTTA->TGCTA + TAACA->TAGCA vs TGCTA->TGTTA + TAGCA->TAACA	4,886	8,150	4,658	14,131	1.82	1.44E-127	-1
ATTAC->ATGAC + GTAAT->GTCAT vs ATGAC->ATTAC + GTCAT->GTAAT	1,716	2,013	1,251	2,652	1.81	4.68E-35	-1
TTCGG->TTAGG + CCGAA->CCTAA vs TTAGG->TTCGG + CCTAA->CCGAA	231	328	608	1,560	1.81	7.43E-09	0
GCCGA->GCAGA + TCGGC->TCTGC vs GCAGA->GCCGA + TCTGC->TCGGC	283	422	936	2,509	1.80	5.88E-11	-1
TATAA->TAGAA + TTATA->TTCTA vs TAGAA->TATAA + TTCTA->TTATA	2,530	3,015	2,563	5,482	1.79	4.23E-58	0
GCCGC->GCAGC + GCGGC->GCTGC vs GCAGC->GCCGC + GCTGC->GCGGC	442	690	615	1,718	1.79	2.18E-13	-1
GTCCG->GTGCG + CGGAC->CGCAC vs GTGCG->GTCCG + CGCAC->CGGAC	78	119	29	79	1.78	4.76E-02	-1
GACGT->GATGT + ACGTC->ACATC vs GATGT->GACGT + ACATC->ACGTC	5,699	8,523	4,598	12,121	1.76	3.03E-119	0
AGTAC->AGGAC + GTACT->GTCCT vs AGGAC->AGTAC + GTCCT->GTACT	646	849	1,207	2,792	1.76	1.19E-18	1
CGTAT->CGAAT + ATACG->ATTCG vs CGAAT->CGTAT + ATTCG->ATACG	29	52	19	60	1.75	1.59E-01	1
AACGC->AATGC + GCGTT->GCATT vs AATGC->AACGC + GCATT->GCGTT	6,004	8,849	5,579	14,414	1.75	1.75E-130	0
GTTAT->GTGAT + ATAAC->ATCAC vs GTGAT->GTTAT + ATCAC->ATAAC	833	987	990	2,056	1.75	1.53E-19	-1
TCCGG->TCTGG + CCGGA->CCAGA vs TCTGG->TCCGG + CCAGA->CCGGA	5,973	9,027	4,288	11,354	1.75	9.13E-116	-1
TTCGC->TTTGC + GCGAA->GCAAA vs TTTGC->TTCGC + GCAAA->GCGAA	3,574	5,114	2,911	7,287	1.75	3.06E-72	1
CGCGT->CGAGT + ACGCG->ACTCG vs CGAGT->CGCGT + ACTCG->ACGCG	63	104	41	118	1.74	3.58E-02	0
GGCGC->GGGGC + GCGCC->GCCCC vs GGGGC->GGCGC + GCCCC->GCGCC	226	373	634	1,817	1.74	4.03E-08	2
CTCGG->CTAGG + CCGAG->CCTAG vs CTAGG->CTCGG + CCTAG->CCGAG	357	527	383	978	1.73	5.56E-09	0
CTCGC->CTAGC + GCGAG->GCTAG vs CTAGC->CTCGC + GCTAG->GCGAG	274	377	242	576	1.73	1.57E-06	0
CGTTA->CGCTA + TAACG->TAGCG vs CGCTA->CGTTA + TAGCG->TAACG	111	157	127	309	1.72	1.77E-03	-1
TGCGC->TGTGC + GCGCA->GCACA vs TGTGC->TGCGC + GCACA->GCGCA	3,738	5,468	2,571	6,457	1.72	2.78E-65	0

Table A.1 (continued)

TGTAG->TGGAG + CTACA->CTCCA vs TGGAG->TGTAG + CTCCA->CTACA	831	1,161	1,314	3,134	1.71	1.04E-20	1
AATAA->AAGAA + TTATT->TTCTT vs AAGAA->AATAA + TTCTT->TTATT	3,989	4,913	9,179	19,270	1.70	7.28E-100	0
CGCGT->CGGGT + ACGCG->ACCCG vs CGGGT->CGCGT + ACCCG->ACGCG	38	74	79	262	1.70	4.91E-02	2
ACCGG->ACTGG + CCGGT->CCAGT vs ACTGG->ACCGG + CCAGT->CCGGT	3,964	6,012	3,600	9,272	1.70	1.12E-76	-1
AATAT->AAGAT + ATATT->ATCTT vs AAGAT->AATAT + ATCTT->ATATT	1,523	1,879	4,369	9,146	1.70	4.48E-40	0
CACGT->CATGT + ACGTG->ACATG vs CATGT->CACGT + ACATG->ACGTG	9,349	14,352	9,765	25,378	1.69	3.37E-189	0
CGTCA->CGACA + TGACG->TGTCG vs CGACA->CGTCA + TGTCG->TGACG	47	82	24	71	1.69	1.12E-01	0
AATAC->AAGAC + GTATT->GTCTT vs AAGAC->AATAC + GTCTT->GTATT	652	839	1,657	3,607	1.69	1.49E-17	0
GGTAC->GGGAC + GTACC->GTCCC vs GGGAC->GGTAC + GTCCC->GTACC	245	355	1,137	2,786	1.69	2.45E-08	1
TCCGC->TCTGC + GCGGA->GCAGA vs TCTGC->TCCGC + GCAGA->GCGGA	4,945	7,336	3,328	8,306	1.68	5.26E-79	-1
TGTCA->TGCCA + TGACA->TGGCA vs TGCCA->TGTCA + TGGCA->TGACA	5,175	7,957	6,239	16,120	1.68	2.78E-108	1
CACGG->CATGG + CCGTG->CCATG vs CATGG->CACGG + CCATG->CCGTG	9,879	15,151	8,303	21,274	1.67	3.48E-172	0
GGTCT->GGCCT + AGACC->AGGCC vs GGCCT->GGTCT + AGGCC->AGACC	2,960	3,907	5,504	12,092	1.66	9.19E-66	1
AGTAA->AGGAA + TTACT->TTCCT vs AGGAA->AGTAA + TTCCT->TTACT	2,042	2,690	2,342	5,129	1.66	1.07E-38	1
CGCGA->CGAGA + TCGCG->TCTCG vs CGAGA->CGCGA + TCTCG->TCGCG	67	122	37	112	1.66	6.16E-02	0
CGTCC->CGCCC + GGACG->GGGCG vs CGCCC->CGTCC + GGGCG->GGACG	240	328	608	1,379	1.66	7.00E-07	1
GACGA->GATGA + TCGTC->TCATC vs GATGA->GACGA + TCATC->TCGTC	5,604	8,667	5,057	12,957	1.66	2.60E-98	0
CGTAG->CGGAG + CTACG->CTCCG vs CGGAG->CGTAG + CTCCG->CTACG	39	53	68	153	1.65	7.13E-02	1
TGTAA->TGGAA + TTACA->TTCCA vs TGGAA->TGTAA + TTCCA->TTACA	1,356	1,872	2,039	4,651	1.65	1.51E-28	1
ATCGG->ATAGG + CCGAT->CCTAT vs ATAGG->ATCGG + CCTAT->CCGAT	263	379	328	781	1.65	2.99E-06	0
ATTCG->ATGCG + CGAAT->CGCAT vs ATGCG->ATTCG + CGCAT->CGAAT	70	72	55	93	1.64	6.20E-02	-1
CTTAT->CTGAT + ATAAG->ATCAG vs CTGAT->CTTAT + ATCAG->ATAAG	1,199	1,572	1,233	2,653	1.64	3.89E-21	-1
GTTAA->GTGAA + TTAAC->TTCAC vs GTGAA->GTTAA + TTCAC->TTAAC	1,701	2,017	1,130	2,196	1.64	3.81E-23	-1
GGTAG->GGGAG + CTACC->CTCCC vs GGGAG->GGTAG + CTCCC->CTACC	663	1,039	1,562	4,011	1.64	1.25E-16	1
TTCGA->TTAGA + TCGAA->TCTAA vs TTAGA->TTCGA + TCTAA->TCGAA	283	395	1,017	2,319	1.63	5.15E-08	0
TTCGG->TTTGG + CCGAA->CCAAA vs TTTGG->TTCGG + CCAAA->CCGAA	5,117	7,762	4,325	10,712	1.63	1.62E-81	1
ATTAT->ATGAT + ATAAT->ATCAT vs ATGAT->ATTAT + ATCAT->ATAAT	3,500	4,335	3,190	6,444	1.63	1.05E-53	-1
TGTAC->TGGAC + GTACA->GTCCA vs TGGAC->TGTAC + GTCCA->GTACA	471	699	911	2,201	1.63	3.97E-11	1
ATTAA->ATGAA + TTAAT->TTCAT vs ATGAA->ATTAA + TTCAT->TTAAT	5,291	6,419	3,565	7,020	1.62	9.85E-68	-1
AACGA->AATGA + TCGTT->TCATT vs AATGA->AACGA + TCATT->TCGTT	8,483	12,652	9,848	23,803	1.62	1.03E-148	0
TCCGA->TCTGA + TCGGA->TCAGA vs TCTGA->TCCGA + TCAGA->TCGGA	5,496	8,010	4,855	11,454	1.62	5.76E-85	-1
GGCGC->GGAGC + GCGCC->GCTCC vs GGAGC->GGCGC + GCTCC->GCGCC	489	791	439	1,148	1.62	6.71E-09	0
CACCC->CAGCC + GGGTG->GGCTG vs CAGCC->CACCC + GGCTG->GGGTG	1,871	2,846	1,210	2,974	1.62	8.94E-26	-1
CTCGA->CTAGA + TCGAG->TCTAG vs CTAGA->CTCGA + TCTAG->TCGAG	252	363	429	998	1.61	5.09E-06	0
GGTTT->GGGTT + AAACC->AACCC vs GGGTT->GGTTT + AACCC->AAACC	1,179	2,339	2,261	7,239	1.61	2.47E-27	0
CGTTC->CGATC + GAACG->GATCG vs CGATC->CGTTC + GATCG->GAACG	36	68	20	61	1.61	1.93E-01	-1
GTTCG->GTGCG + CGAAC->CGCAC vs GTGCG->GTTCG + CGCAC->CGAAC	27	43	44	113	1.61	1.60E-01	-1
CATAT->CAGAT + ATATG->ATCTG vs CAGAT->CATAT + ATCTG->ATATG	806	1,159	2,005	4,612	1.60	6.66E-18	0
GTCGG->GTAGG + CCGAC->CCTAC vs GTAGG->GTCGG + CCTAC->CCGAC	221	345	261	650	1.59	7.67E-05	0
CGCCC->CGGCC + GGGCG->GGCCG vs CGGCC->CGCCC + GGCCG->GGGCG	133	239	77	220	1.59	1.16E-02	-1
GGTTG->GGGTG + CAACC->CACCC vs GGGTG->GGTTG + CACCC->CAACC	608	1,153	2,365	7,119	1.59	1.09E-15	1
TCCGC->TCAGC + GCGGA->GCTGA vs TCAGC->TCCGC + GCTGA->GCGGA	168	299	657	1,853	1.58	4.60E-05	-1
ATCGT->ATAGT + ACGAT->ACTAT vs ATAGT->ATCGT + ACTAT->ACGAT	328	449	728	1,578	1.58	2.13E-07	0
CATAC->CAGAC + GTATG->GTCTG vs CAGAC->CATAC + GTCTG->GTATG	490	644	832	1,730	1.58	1.29E-09	0
GTTCC->GTCCC + GGAAC->GGGAC vs GTCCC->GTTCC + GGGAC->GGAAC	1,582	2,334	3,461	8,060	1.58	2.82E-31	1
TGTTT->TGCTT + AAACA->AAGCA vs TGCTT->TGTTT + AAGCA->AAACA	10,155	16,612	9,273	23,843	1.57	8.39E-145	-1
TATAC->TAGAC + GTATA->GTCTA vs TAGAC->TATAC + GTCTA->GTATA	597	793	608	1,269	1.57	1.85E-09	0
AGTCG->AGCCG + CGACT->CGGCT vs AGCCG->AGTCG + CGGCT->CGACT	142	180	365	726	1.57	1.02E-03	1
CACGC->CATGC + GCGTG->GCATG vs CATGC->CACGC + GCATG->GCGTG	7,578	11,807	5,589	13,642	1.57	1.01E-94	0
GGTTC->GGGTC + GAACC->GACCC vs GGGTC->GGTTC + GACCC->GAACC	434	877	1,556	4,911	1.56	6.70E-11	1
TGCGG->TGTGG + CCGCA->CCACA vs TGTGG->TGCGG + CCACA->CCGCA	6,039	9,025	4,200	9,802	1.56	2.31E-71	0
ACCGC->ACTGC + GCGGT->GCAGT vs ACTGC->ACCGC + GCAGT->GCGGT	3,880	6,062	2,797	6,820	1.56	1.31E-47	-1
GACGC->GATGC + GCGTC->GCATC vs GATGC->GACGC + GCATC->GCGTC	5,144	8,268	3,352	8,388	1.56	1.70E-59	0
TTTAC->TTGAC + GTAAA->GTCAA vs TTGAC->TTTAC + GTCAA->GTAAA	1,438	1,827	1,065	2,106	1.56	3.03E-17	-1
CACCG->CAGCG + CGGTG->CGCTG vs CAGCG->CACCG + CGCTG->CGGTG	146	229	97	236	1.55	1.11E-02	-1
ACCGA->ACTGA + TCGGT->TCAGT vs ACTGA->ACCGA + TCAGT->TCGGT	4,614	6,902	5,037	11,649	1.55	1.62E-64	-1
TGTCT->TGCCT + AGACA->AGGCA vs TGCCT->TGTCT + AGGCA->AGACA	6,552	10,323	8,244	20,072	1.55	5.73E-98	1
CCCGT->CCTGT + ACGGG->ACAGG vs CCTGT->CCCGT + ACAGG->ACGGG	5,223	7,665	4,098	9,293	1.55	3.67E-62	-1
TTTAT->TTGAT + ATAAA->ATCAA vs TTGAT->TTTAT + ATCAA->ATAAA	3,149	4,149	2,706	5,501	1.54	6.87E-38	-1
CGTCA->CGCCA + TGACG->TGGCG vs CGCCA->CGTCA + TGGCG->TGACG	193	245	291	570	1.54	6.17E-04	1
CATAA->CAGAA + TTATG->TTCTG vs CAGAA->CATAA + TTCTG->TTATG	1,688	2,249	2,386	4,892	1.54	2.94E-25	0
AGCGG->AGTGG + CCGCT->CCACT vs AGTGG->AGCGG + CCACT->CCGCT	5,883	8,825	4,334	10,003	1.54	7.59E-67	0
GCCGT->GCTGT + ACGGC->ACAGC vs GCTGT->GCCGT + ACAGC->ACGGC	4,250	6,171	2,996	6,687	1.54	9.81E-47	-1
GGTTA->GGGTA + TAACC->TACCC vs GGGTA->GGTTA + TACCC->TAACC	437	800	1,432	4,025	1.54	8.21E-10	1
GGTAT->GGAAT + ATACC->ATTCC vs GGAAT->GGTAT + ATTCC->ATACC	439	733	716	1,835	1.53	3.09E-08	1
GTTAG->GTGAG + CTAAC->CTCAC vs GTGAG->GTTAG + CTCAC->CTAAC	781	1,029	793	1,602	1.53	9.20E-11	-1
GTTCA->GTCCA + TGAAC->TGGAC vs GTCCA->GTTCA + TGGAC->TGAAC	1,832	2,799	2,605	6,096	1.53	4.31E-28	0
CGTTG->CGGTG + CAACG->CACCG vs CGGTG->CGTTG + CACCG->CAACG	34	61	101	276	1.52	1.31E-01	0
GATAA->GAGAA + TTATC->TTCTC vs GAGAA->GATAA + TTCTC->TTATC	1,538	2,007	4,056	8,032	1.52	1.14E-25	0
GGTAA->GGGAA + TTACC->TTCCC vs GGGAA->GGTAA + TTCCC->TTACC	692	1,081	1,764	4,171	1.51	1.12E-12	1
CTTCG->CTGCG + CGAAG->CGCAG vs CTGCG->CTTCG + CGCAG->CGAAG	62	93	113	256	1.51	6.05E-02	-1
AACCC->AAGCC + GGGTT->GGCTT vs AAGCC->AACCC + GGCTT->GGGTT	1,428	2,268	914	2,190	1.51	6.62E-15	-1
GACCC->GAGCC + GGGTC->GGCTC vs GAGCC->GACCC + GGCTC->GGGTC	967	1,427	911	2,020	1.50	5.96E-12	-1

Table A.1 (continued)

GGTCT->GGGCT + AGACC->AGCCC vs GGGCT->GGTCT + AGCCC->AGACC	572	1,024	2,100	5,634	1.50	2.12E-11	1
TTTTT->TTGTT + AAAAA->AACAA vs TTGTT->TTTTT + AACAA->AAAAA	6,667	9,531	5,752	12,291	1.49	9.66E-70	-3
GATAT->GAGAT + ATATC->ATCTC vs GAGAT->GATAT + ATCTC->ATATC	712	1,009	1,932	4,086	1.49	4.81E-12	0
CGTTC->CGCTC + GAACG->GAGCG vs CGCTC->CGTTC + GAGCG->GAACG	199	270	408	826	1.49	6.46E-04	-1
CACGA->CATGA + TCGTG->TCATG vs CATGA->CACGA + TCATG->TCGTG	7,160	11,097	7,336	16,930	1.49	1.33E-81	0
GTTCT->GTCCT + AGAAC->AGGAC vs GTCCT->GTTCT + AGGAC->AGAAC	2,797	4,322	4,193	9,642	1.49	1.51E-37	0
AGTAG->AGGAG + CTACT->CTCCT vs AGGAG->AGTAG + CTCCT->CTACT	1,395	2,126	1,614	3,657	1.49	1.56E-17	1
TGTCC->TGCCC + GGACA->GGGCA vs TGCCC->TGTCC + GGGCA->GGACA	3,943	6,210	6,592	15,430	1.49	2.92E-54	1
CCCGG->CCTGG + CCGGG->CCAGG vs CCTGG->CCCGG + CCAGG->CCGGG	8,944	13,448	5,765	12,865	1.48	7.57E-79	-1
CGTAC->CGCAC + GTACG->GTGCG vs CGCAC->CGTAC + GTGCG->GTACG	115	145	258	483	1.48	1.17E-02	0
CGCCT->CGACT + AGGCG->AGTCG vs CGACT->CGCCT + AGTCG->AGGCG	129	261	28	84	1.48	1.41E-01	-1
GGTTA->GGCTA + TAACC->TAGCC vs GGCTA->GGTTA + TAGCC->TAACC	2,009	3,101	2,365	5,394	1.48	3.03E-24	-1
TTTAA->TTGAA + TTAAA->TTCAA vs TTGAA->TTTAA + TTCAA->TTAAA	4,987	6,747	2,521	5,028	1.47	4.80E-36	-1
ACTTC->ACATC + GAAGT->GATGT vs ACATC->ACTTC + GATGT->GAAGT	883	1,576	1,351	3,548	1.47	1.20E-12	-1
ATTCA->ATCCA + TGAAT->TGGAT vs ATCCA->ATTCA + TGGAT->TGAAT	3,347	4,922	4,093	8,840	1.47	4.06E-38	0
CTTCA->CTCCA + TGAAG->TGGAG vs CTCCA->CTTCA + TGGAG->TGAAG	2,900	4,374	5,599	12,363	1.46	2.17E-38	0
GGTTC->GGATC + GAACC->GATCC vs GGATC->GGTTC + GATCC->GAACC	337	947	555	2,282	1.46	4.45E-06	-1
CCCCC->CCTCC + GGGGG->GGAGG vs CCTCC->CCCCC + GGAGG->GGGGG	4,728	7,620	2,137	5,034	1.46	1.82E-32	-3
CTTCG->CTCCG + CGAAG->CGGAG vs CTCCG->CTTCG + CGGAG->CGAAG	125	189	325	717	1.46	8.21E-03	0
ATTCT->ATCCT + AGAAT->AGGAT vs ATCCT->ATTCT + AGGAT->AGAAT	4,656	7,078	5,255	11,649	1.46	1.26E-49	0
TTTTA->TTGTA + TAAAA->TACAA vs TTGTA->TTTTA + TACAA->TAAAA	6,153	8,677	3,619	7,438	1.46	5.15E-46	-2
GGTCA->GGCCA + TGACC->TGGCC vs GGCCA->GGTCA + TGGCC->TGACC	2,652	3,841	4,806	10,142	1.46	2.43E-33	1
ATCGC->ATAGC + GCGAT->GCTAT vs ATAGC->ATCGC + GCTAT->GCGAT	208	366	245	627	1.45	2.04E-03	0
AGTAT->AGAAT + ATACT->ATTCT vs AGAAT->AGTAT + ATTCT->ATACT	1,107	1,844	1,208	2,922	1.45	1.11E-12	1
AGCGC->AGTGC + GCGCT->GCACT vs AGTGC->AGCGC + GCACT->GCGCT	4,329	6,695	2,889	6,471	1.45	4.19E-35	0
ATCGT->ATTGT + ACGAT->ACAAT vs ATTGT->ATCGT + ACAAT->ACGAT	3,535	5,308	3,788	8,234	1.45	1.17E-35	1
CGTCC->CGGCC + GGACG->GGCCG vs CGGCC->CGTCC + GGCCG->GGACG	50	70	155	313	1.44	1.16E-01	1
AACCT->AAGCT + AGGTT->AGCTT vs AAGCT->AACCT + AGCTT->AGGTT	2,343	3,628	1,953	4,345	1.44	5.89E-21	-1
CACCA->CAGCA + TGGTG->TGCTG vs CAGCA->CACCA + TGCTG->TGGTG	2,783	4,269	1,462	3,220	1.44	2.95E-19	-1
GACCG->GAGCG + CGGTC->CGCTC vs GAGCG->GACCG + CGCTC->CGGTC	75	120	71	163	1.43	1.12E-01	-1
TTTTT->TTCTT + AAAAA->AAGAA vs TTCTT->TTTTT + AAGAA->AAAAA	12,500	20,679	8,677	20,573	1.43	1.17E-97	-3
GTTTA->GTCTA + TAAAC->TAGAC vs GTCTA->GTTTA + TAGAC->TAAAC	1,610	2,560	1,751	3,985	1.43	2.23E-16	-2
GGTTT->GGCTT + AAACC->AAGCC vs GGCTT->GGTTT + AAGCC->AAACC	4,499	6,792	5,390	11,626	1.43	6.62E-44	-1
TGTCA->TGGCA + TGACA->TGCCA vs TGGCA->TGTCA + TGCCA->TGACA	947	1,530	2,272	5,243	1.43	9.54E-13	1
GGTCC->GGCCC + GGACC->GGGCC vs GGCCC->GGTCC + GGGCC->GGACC	2,093	2,850	5,206	10,123	1.43	1.82E-25	1
AGTTT->AGGTT + AAACT->AACCT vs AGGTT->AGTTT + AACCT->AAACT	2,093	3,510	4,561	10,886	1.42	6.27E-26	-1
TGCGA->TGTGA + TCGCA->TCACA vs TGTGA->TGCGA + TCACA->TCGCA	4,371	6,641	3,696	7,992	1.42	5.70E-36	0
ACTTA->ACATA + TAAGT->TATGT vs ACATA->ACTTA + TATGT->TAAGT	900	1,473	1,917	4,458	1.42	1.29E-11	-1
AGTCT->AGCCT + AGACT->AGGCT vs AGCCT->AGTCT + AGGCT->AGACT	4,829	6,834	6,407	12,834	1.42	2.60E-45	1
AGTCA->AGCCA + TGACT->TGGCT vs AGCCA->AGTCA + TGGCT->TGACT	4,054	5,686	5,841	11,581	1.41	5.00E-39	1
GGTTC->GGCTC + GAACC->GAGCC vs GGCTC->GGTTC + GAGCC->GAACC	2,354	3,448	4,612	9,541	1.41	8.03E-26	-1
CTTCC->CTCCC + GGAAG->GGGAG vs CTCCC->CTTCC + GGGAG->GGAAG	3,540	5,409	7,022	15,135	1.41	1.57E-38	1
ATTCC->ATCCC + GGAAT->GGGAT vs ATCCC->ATTCC + GGGAT->GGAAT	2,531	3,979	4,060	8,942	1.40	1.92E-25	1
GTTTC->GTCTC + GAAAC->GAGAC vs GTCTC->GTTTC + GAGAC->GAAAC	2,098	3,247	3,014	6,531	1.40	2.06E-20	-2
TACCC->TAGCC + GGGTA->GGCTA vs TAGCC->TACCC + GGCTA->GGGTA	833	1,413	635	1,507	1.40	4.11E-07	-1
TGTGA->TGGGA + TCACA->TCCCA vs TGGGA->TGTGA + TCCCA->TCACA	2,106	3,091	1,406	2,881	1.40	2.93E-14	2
ATTAG->ATGAG + CTAAT->CTCAT vs ATGAG->ATTAG + CTCAT->CTAAT	2,281	3,074	1,473	2,763	1.39	2.17E-14	-1
AGTCA->AGGCA + TGACT->TGCCT vs AGGCA->AGTCA + TGCCT->TGACT	1,111	1,768	2,805	6,212	1.39	5.81E-13	1
TCTTG->TCATG + CAAGA->CATGA vs TCATG->TCTTG + CATGA->CAAGA	776	1,123	1,961	3,941	1.39	4.93E-09	-1
ATTTT->ATGTT + AAAAT->AACAT vs ATGTT->ATTTT + AACAT->AAAAT	7,115	10,332	1,940	3,906	1.39	1.60E-24	-2
TTCGA->TTTGA + TCGAA->TCAAA vs TTTGA->TTCGA + TCAAA->TCGAA	5,271	7,814	4,890	10,036	1.38	6.31E-38	1
TTCGC->TTAGC + GCGAA->GCTAA vs TTAGC->TTCGC + GCTAA->GCGAA	158	247	503	1,088	1.38	8.77E-03	0
CTTAA->CTGAA + TTAAG->TTCAG vs CTGAA->CTTAA + TTCAG->TTAAG	2,607	3,669	1,384	2,692	1.38	3.04E-14	-1
AGTTG->AGGTG + CAACT->CACCT vs AGGTG->AGTTG + CACCT->CAACT	1,502	2,586	2,533	6,016	1.38	6.10E-15	0
TTTTG->TTGTG + CAAAA->CACAA vs TTGTG->TTTTG + CACAA->CAAAA	2,852	4,363	2,962	6,249	1.38	6.10E-22	-2
CACCT->CAGCT + AGGTG->AGCTG vs CAGCT->CACCT + AGCTG->AGGTG	2,010	3,281	2,491	5,606	1.38	2.74E-17	-1
TTTCA->TTCCA + TGAAA->TGGAA vs TTCCA->TTTCA + TGGAA->TGAAA	6,922	11,394	6,990	15,848	1.38	1.51E-51	-1
TCTTT->TCATT + AAAGA->AATGA vs TCATT->TCTTT + AATGA->AAAGA	1,639	2,381	3,052	6,106	1.38	1.15E-15	-1
TGTAT->TGAAT + ATACA->ATTCA vs TGAAT->TGTAT + ATTCA->ATACA	1,027	1,781	1,086	2,592	1.38	6.17E-09	1
GTTCC->GTGCC + GGAAC->GGCAC vs GTGCC->GTTCC + GGCAC->GGAAC	519	752	1,068	2,124	1.37	8.43E-06	-1
TATAG->TAGAG + CTATA->CTCTA vs TAGAG->TATAG + CTCTA->CTATA	1,582	2,381	779	1,606	1.37	1.70E-08	0
CCCGA->CCTGA + TCGGG->TCAGG vs CCTGA->CCCGA + TCAGG->TCGGG	5,900	9,164	4,759	10,121	1.37	1.45E-37	-1
GTTTT->GTCTT + AAAAC->AAGAC vs GTCTT->GTTTT + AAGAC->AAAAC	3,438	5,604	3,485	7,775	1.37	2.79E-25	-2
TTTAG->TTGAG + CTAAA->CTCAA vs TTGAG->TTTAG + CTCAA->CTAAA	1,980	2,739	1,249	2,363	1.37	2.13E-11	-1
AGTCC->AGCCC + GGACT->GGGCT vs AGCCC->AGTCC + GGGCT->GGACT	2,745	4,050	5,771	11,628	1.37	4.29E-25	1
GGTTT->GGATT + AAACC->AATCC vs GGATT->GGTTT + AATCC->AAACC	838	2,176	915	3,243	1.36	5.14E-08	-1
GGTCG->GGCCG + CGACC->CGGCC vs GGCCG->GGTCG + CGGCC->CGACC	135	180	508	924	1.36	2.54E-02	1
TCTTC->TCATC + GAAGA->GATGA vs TCATC->TCTTC + GATGA->GAAGA	878	1,406	1,869	4,082	1.36	4.82E-09	-1
CGTGT->CGGGT + ACACG->ACCCG vs CGGGT->CGTGT + ACCCG->ACACG	72	149	85	240	1.36	1.60E-01	2
CTCGT->CTTGT + ACGAG->ACAAG vs CTTGT->CTCGT + ACAAG->ACGAG	4,053	6,116	3,249	6,681	1.36	3.72E-25	1
GTCGT->GTTGT + ACGAC->ACAAC vs GTTGT->GTCGT + ACAAC->ACGAC	2,160	3,252	1,737	3,562	1.36	6.47E-14	1
TGTAC->TGAAC + GTACA->GTTCA vs TGAAC->TGTAC + GTTCA->GTACA	574	983	478	1,114	1.36	1.02E-04	1
GGCGT->GGTGT + ACGCC->ACACC vs GGTGT->GGCGT + ACACC->ACGCC	3,498	5,140	2,258	4,499	1.36	7.94E-19	0

Table A.1 (continued)

CGTTT->CGGTT + AAACG->AACCG vs CGGTT->CGTTT + AACCG->AAACG	75	129	93	217	1.36	1.61E-01	-1
CATCT->CAGCT + AGATG->AGCTG vs CAGCT->CATCT + AGCTG->AGATG	1,173	1,855	1,803	3,855	1.35	4.68E-10	0
ATTTA->ATGTA + TAAAT->TACAT vs ATGTA->ATTTA + TACAT->TAAAT	7,183	9,879	1,825	3,386	1.35	2.42E-19	-2
TGTCC->TGGCC + GGACA->GGCCA vs TGGCC->TGTCC + GGCCA->GGACA	661	1,062	1,657	3,590	1.35	6.44E-07	1
CTTCC->CTGCC + GGAAG->GGCAG vs CTGCC->CTTCC + GGCAG->GGAAG	1,232	2,079	2,711	6,162	1.35	1.31E-11	-1
ACTTT->ACATT + AAAGT->AATGT vs ACATT->ACTTT + AATGT->AAAGT	2,008	3,366	2,873	6,472	1.34	1.13E-15	-1
GGTAT->GGCAT + ATACC->ATGCC vs GGCAT->GGTAT + ATGCC->ATACC	3,052	4,635	3,995	8,130	1.34	2.41E-21	0
GATAC->GAGAC + GTATC->GTCTC vs GAGAC->GATAC + GTCTC->GTATC	280	415	1,000	1,985	1.34	1.59E-03	0
CATCA->CAGCA + TGATG->TGCTG vs CAGCA->CATCA + TGCTG->TGATG	1,107	1,645	1,671	3,322	1.34	9.44E-09	0
GTTGG->GTGGG + CCAAC->CCCAC vs GTGGG->GTTGG + CCCAC->CCAAC	899	1,382	1,624	3,339	1.34	7.77E-08	1
CGCGG->CGAGG + CCGCG->CCTCG vs CGAGG->CGCGG + CCTCG->CCGCG	166	301	66	160	1.34	1.37E-01	0
TTTTA->TTCTA + TAAAA->TAGAA vs TTCTA->TTTTA + TAGAA->TAAAA	7,678	13,148	4,990	11,416	1.34	4.66E-38	-2
AGTTA->AGGTA + TAACT->TACCT vs AGGTA->AGTTA + TACCT->TAACT	1,222	2,105	2,475	5,687	1.33	1.08E-10	0
GCTTC->GCATC + GAAGC->GATGC vs GCATC->GCTTC + GATGC->GAAGC	541	884	859	1,872	1.33	5.63E-05	-1
ACTAC->ACAAC + GTAGT->GTTGT vs ACAAC->ACTAC + GTTGT->GTAGT	548	893	549	1,193	1.33	2.46E-04	1
GTTAT->GTCAT + ATAAC->ATGAC vs GTCAT->GTTAT + ATGAC->ATAAC	2,529	4,016	3,248	6,872	1.33	2.23E-17	-1
CATCG->CACCG + CGATG->CGGTG vs CACCG->CATCG + CGGTG->CGATG	224	338	379	760	1.33	1.30E-02	1
TTTTC->TTCTC + GAAAA->GAGAA vs TTCTC->TTTTC + GAGAA->GAAAA	7,964	12,840	7,703	16,497	1.33	2.60E-45	-2
ATTGG->ATGGG + CCAAT->CCCAT vs ATGGG->ATTGG + CCCAT->CCAAT	1,303	2,010	1,038	2,127	1.33	1.11E-07	1
CTTCT->CTCCT + AGAAG->AGGAG vs CTCCT->CTTCT + AGGAG->AGAAG	5,162	7,802	9,797	19,659	1.33	1.97E-37	0
AACCA->AAGCA + TGGTT->TGCTT vs AAGCA->AACCA + TGCTT->TGGTT	2,804	4,507	1,774	3,785	1.33	1.28E-13	-1
AGTCC->AGGCC + GGACT->GGCCT vs AGGCC->AGTCC + GGCCT->GGACT	639	1,013	2,336	4,908	1.33	1.67E-06	1
CTTCA->CTGCA + TGAAG->TGCAG vs CTGCA->CTTCA + TGCAG->TGAAG	1,487	2,393	2,632	5,605	1.32	1.96E-11	-1
ATCGG->ATTGG + CCGAT->CCAAT vs ATTGG->ATCGG + CCAAT->CCGAT	2,936	5,043	2,265	5,144	1.32	1.25E-15	1
ATTGA->ATGGA + TCAAT->TCCAT vs ATGGA->ATTGA + TCCAT->TCAAT	2,622	3,654	1,127	2,075	1.32	1.50E-09	0
CTTAC->CTGAC + GTAAG->GTCAG vs CTGAC->CTTAC + GTCAG->GTAAG	757	1,072	768	1,436	1.32	4.46E-05	-1
TTTCT->TTCCT + AGAAA->AGGAA vs TTCCT->TTTCT + AGGAA->AGAAA	12,120	20,523	11,375	25,378	1.32	1.06E-64	-1
CGTGT->CGAGT + ACACG->ACTCG vs CGAGT->CGTGT + ACTCG->ACACG	84	131	37	76	1.32	3.40E-01	0
TGTTC->TGCTC + GAACA->GAGCA vs TGCTC->TGTTC + GAGCA->GAACA	5,322	8,635	7,162	15,276	1.31	8.27E-33	-1
CTTGT->CTGGT + ACAAG->ACCAG vs CTGGT->CTTGT + ACCAG->ACAAG	1,482	2,273	1,964	3,955	1.31	9.42E-10	0
GTTTT->GTGTT + AAAAC->AACAC vs GTGTT->GTTTT + AACAC->AAAAC	2,398	3,653	1,404	2,805	1.31	2.46E-10	-2
TGTTG->TGCTG + CAACA->CAGCA vs TGCTG->TGTTG + CAGCA->CAACA	8,240	14,168	6,494	14,641	1.31	1.68E-39	-1
CGTTT->CGCTT + AAACG->AAGCG vs CGCTT->CGTTT + AAGCG->AAACG	242	368	320	638	1.31	1.97E-02	-1
GTTCG->GTCCG + CGAAC->CGGAC vs GTCCG->GTTCG + CGGAC->CGAAC	43	77	124	291	1.31	2.73E-01	0
GGTAC->GGAAC + GTACC->GTTCC vs GGAAC->GGTAC + GTTCC->GTACC	310	584	381	940	1.31	6.24E-03	1
TATCT->TAGCT + AGATA->AGCTA vs TAGCT->TATCT + AGCTA->AGATA	1,653	2,612	1,182	2,441	1.31	3.84E-08	0
GTTCT->GTGCT + AGAAC->AGCAC vs GTGCT->GTTCT + AGCAC->AGAAC	818	1,259	1,461	2,938	1.31	3.02E-06	-1
GTTAC->GTCAC + GTAAC->GTGAC vs GTCAC->GTTAC + GTGAC->GTAAC	1,778	2,770	2,426	4,935	1.31	3.32E-11	-1
CGCAA->CGGAA + TTGCG->TTCCG vs CGGAA->CGCAA + TTCCG->TTGCG	39	72	85	205	1.31	3.38E-01	1
ACTTG->ACATG + CAAGT->CATGT vs ACATG->ACTTG + CATGT->CAAGT	856	1,442	2,015	4,431	1.31	4.47E-07	-1
GCTTA->GCCTA + TAAGC->TAGGC vs GCCTA->GCTTA + TAGGC->TAAGC	1,712	3,053	2,296	5,337	1.30	4.79E-11	0
GGTGT->GGGGT + ACACC->ACCCC vs GGGGT->GGTGT + ACCCC->ACACC	840	1,557	1,534	3,704	1.30	1.21E-06	2
CGTGG->CGCGG + CCACG->CCGCG vs CGCGG->CGTGG + CCGCG->CCACG	406	610	813	1,588	1.30	1.37E-03	0
AGTCT->AGGCT + AGACT->AGCCT vs AGGCT->AGTCT + AGCCT->AGACT	1,186	1,981	3,824	8,300	1.30	1.08E-09	1
GCCGG->GCTGG + CCGGC->CCAGC vs GCTGG->GCCGG + CCAGC->CCGGC	6,227	9,899	3,452	7,122	1.30	1.25E-22	-1
AGTAC->AGAAC + GTACT->GTTCT vs AGAAC->AGTAC + GTTCT->GTACT	577	1,068	609	1,463	1.30	4.23E-04	1
GTTGA->GTGGA + TCAAC->TCCAC vs GTGGA->GTTGA + TCCAC->TCAAC	1,324	1,864	616	1,124	1.30	5.08E-05	0
TCTAT->TCAAT + ATAGA->ATTGA vs TCAAT->TCTAT + ATTGA->ATAGA	835	1,208	985	1,845	1.29	3.57E-05	1
GCCGA->GCTGA + TCGGC->TCAGC vs GCTGA->GCCGA + TCAGC->TCGGC	4,421	7,095	3,509	7,281	1.29	2.31E-19	-1
AGCGA->AGTGA + TCGCT->TCACT vs AGTGA->AGCGA + TCACT->TCGCT	5,607	8,669	4,804	9,601	1.29	9.61E-25	0
TGTGC->TGGGC + GCACA->GCCCA vs TGGGC->TGTGC + GCCCA->GCACA	1,484	2,487	989	2,142	1.29	8.93E-07	2
AACCG->AAGCG + CGGTT->CGCTT vs AAGCG->AACCG + CGCTT->CGGTT	90	157	51	115	1.29	3.04E-01	-1
GCTAA->GCCAA + TTAGC->TTGGC vs GCCAA->GCTAA + TTGGC->TTAGC	2,587	4,013	2,712	5,434	1.29	3.80E-13	1
GGTAG->GGAAG + CTACC->CTTCC vs GGAAG->GGTAG + CTTCC->CTACC	484	855	705	1,607	1.29	8.88E-04	1
ATTTT->ATCTT + AAAAT->AAGAT vs ATCTT->ATTTT + AAGAT->AAAAT	7,042	11,504	5,231	11,018	1.29	1.33E-28	-2
GTCGC->GTAGC + GCGAC->GCTAC vs GTAGC->GTCGC + GCTAC->GCGAC	150	267	180	413	1.29	9.05E-02	0
GCTCA->GCCCA + TGAGC->TGGGC vs GCCCA->GCTCA + TGGGC->TGAGC	2,155	3,301	4,341	8,568	1.29	1.28E-13	2
TGTCG->TGCCG + CGACA->CGGCA vs TGCCG->TGTCG + CGGCA->CGACA	165	244	358	682	1.29	5.57E-02	1
CTCCG->CTGCG + CGGAG->CGCAG vs CTGCG->CTCCG + CGCAG->CGGAG	201	355	95	216	1.29	1.34E-01	-1
CCTTA->CCATA + TAAGG->TATGG vs CCATA->CCTTA + TATGG->TAAGG	621	901	973	1,815	1.29	2.71E-04	-1
TGCCG->TGGCG + CGGCA->CGCCA vs TGGCG->TGCCG + CGCCA->CGGCA	60	112	60	144	1.28	3.29E-01	0
TCTTA->TCCTA + TAAGA->TAGGA vs TCCTA->TCTTA + TAGGA->TAAGA	3,945	7,016	5,141	11,743	1.28	3.50E-21	0
GCTCG->GCCCG + CGAGC->CGGGC vs GCCCG->GCTCG + CGGGC->CGAGC	121	195	406	836	1.28	8.65E-02	2
TGTAT->TGCAT + ATACA->ATGCA vs TGCAT->TGTAT + ATGCA->ATACA	7,302	11,816	6,307	13,032	1.28	1.61E-29	0
CTTAG->CTGAG + CTAAG->CTCAG vs CTGAG->CTTAG + CTCAG->CTAAG	1,124	1,642	1,038	1,936	1.28	1.61E-05	-1
CCTTT->CCATT + AAAGG->AATGG vs CCATT->CCTTT + AATGG->AAAGG	1,247	1,890	1,716	3,320	1.28	5.58E-07	-1
TGTTA->TGGTA + TAACA->TACCA vs TGGTA->TGTTA + TACCA->TAACA	1,097	2,015	1,868	4,374	1.27	4.78E-07	0
CCCGC->CCTGC + GCGGG->GCAGG vs CCTGC->CCCGC + GCAGG->GCGGG	6,899	11,181	3,862	7,975	1.27	7.05E-22	-1
ATCGA->ATAGA + TCGAT->TCTAT vs ATAGA->ATCGA + TCTAT->TCGAT	347	567	527	1,097	1.27	8.68E-03	0
AATAG->AAGAG + CTATT->CTCTT vs AAGAG->AATAG + CTCTT->CTATT	2,182	3,341	1,777	3,459	1.27	5.65E-09	0
ACTAT->ACAAT + ATAGT->ATTGT vs ACAAT->ACTAT + ATTGT->ATAGT	1,068	1,593	1,251	2,371	1.27	1.26E-05	1
GTCGA->GTAGA + TCGAC->TCTAC vs GTAGA->GTCGA + TCTAC->TCGAC	162	254	384	762	1.27	6.73E-02	0
CGTTG->CGCTG + CAACG->CAGCG vs CGCTG->CGTTG + CAGCG->CAACG	490	871	362	813	1.26	9.88E-03	-1

Table A.1 (continued)

GTTAG->GTCAG + CTAAC->CTGAC vs GTCAG->GTTAG + CTGAC->CTAAC	2,279	3,607	2,866	5,730	1.26	9.54E-11	-1
TACCT->TAGCT + AGGTA->AGCTA vs TAGCT->TACCT + AGCTA->AGGTA	1,379	2,344	1,685	3,617	1.26	5.23E-07	-1
CTTAG->CTCAG + CTAAG->CTGAG vs CTCAG->CTTAG + CTGAG->CTAAG	4,564	6,833	6,750	12,759	1.26	5.31E-21	-1
AGTGA->AGGGA + TCACT->TCCCT vs AGGGA->AGTGA + TCCCT->TCACT	2,317	3,603	1,585	3,109	1.26	3.01E-08	2
TTTCC->TTCCC + GGAAA->GGGAA vs TTCCC->TTTCC + GGGAA->GGAAA	7,968	13,704	8,429	18,285	1.26	1.70E-32	0
ATTTA->ATCTA + TAAAT->TAGAT vs ATCTA->ATTTA + TAGAT->TAAAT	3,897	6,397	3,132	6,482	1.26	2.28E-14	-2
ATCCC->ATGCC + GGGAT->GGCAT vs ATGCC->ATCCC + GGCAT->GGGAT	1,449	2,551	449	996	1.26	8.00E-04	-1
TGTTC->TGGTC + GAACA->GACCA vs TGGTC->TGTTC + GACCA->GAACA	865	1,696	1,566	3,868	1.26	1.72E-05	0
CATCC->CAGCC + GGATG->GGCTG vs CAGCC->CATCC + GGCTG->GGATG	835	1,256	1,573	2,980	1.26	5.16E-05	0
AGTTC->AGGTC + GAACT->GACCT vs AGGTC->AGTTC + GACCT->GAACT	896	1,681	2,293	5,413	1.26	4.90E-06	0
GGTTG->GGCTG + CAACC->CAGCC vs GGCTG->GGTTG + CAGCC->CAACC	4,823	8,225	4,650	9,975	1.26	6.66E-19	-1
CGTTT->CGATT + AAACG->AATCG vs CGATT->CGTTT + AATCG->AAACG	42	102	36	110	1.26	4.99E-01	-1
CGTTC->CGGTC + GAACG->GACCG vs CGGTC->CGTTC + GACCG->GAACG	32	59	66	153	1.26	4.93E-01	0
TTTCA->TTGCA + TGAAA->TGCAA vs TTGCA->TTTCA + TGCAA->TGAAA	2,895	5,120	3,020	6,684	1.25	7.30E-12	-1
CATTG->CAGTG + CAATG->CACTG vs CAGTG->CATTG + CACTG->CAATG	878	1,349	1,817	3,492	1.25	3.97E-05	-1
GCTAG->GCCAG + CTAGC->CTGGC vs GCCAG->GCTAG + CTGGC->CTAGC	2,336	3,575	4,087	7,822	1.25	3.57E-11	1
CGTGC->CGCGC + GCACG->GCGCG vs CGCGC->CGTGC + GCGCG->GCACG	218	350	544	1,091	1.25	4.60E-02	0
GATTG->GAATG + CAATC->CATTC vs GAATG->GATTG + CATTC->CAATC	609	1,187	717	1,745	1.25	1.59E-03	0
CTTTA->CTCTA + TAAAG->TAGAG vs CTCTA->CTTTA + TAGAG->TAAAG	2,869	4,647	3,835	7,744	1.25	3.30E-12	-2
TCCGC->TCGGC + GCGGA->GCCGA vs TCGGC->TCCGC + GCCGA->GCGGA	141	278	263	646	1.25	1.13E-01	0
CTTTG->CTCTG + CAAAG->CAGAG vs CTCTG->CTTTG + CAGAG->CAAAG	6,728	11,237	5,468	11,365	1.24	1.35E-21	-2
ATCGC->ATTGC + GCGAT->GCAAT vs ATTGC->ATCGC + GCAAT->GCGAT	2,668	4,465	2,085	4,338	1.24	4.47E-09	1
TTTGA->TTGGA + TCAAA->TCCAA vs TTGGA->TTTGA + TCCAA->TCAAA	3,506	5,196	1,237	2,279	1.24	3.48E-07	-1
GCTAT->GCCAT + ATAGC->ATGGC vs GCCAT->GCTAT + ATGGC->ATAGC	2,310	3,761	3,472	7,021	1.24	3.53E-10	1
TTTTG->TTCTG + CAAAA->CAGAA vs TTCTG->TTTTG + CAGAA->CAAAA	11,489	19,817	6,214	13,312	1.24	1.50E-28	-2
TGTCT->TGGCT + AGACA->AGCCA vs TGGCT->TGTCT + AGCCA->AGACA	936	1,728	2,637	6,044	1.24	9.71E-06	1
CATAG->CAGAG + CTATG->CTCTG vs CAGAG->CATAG + CTCTG->CTATG	1,257	1,980	1,156	2,259	1.24	5.08E-05	0
GTCCC->GTGCC + GGGAC->GGCAC vs GTGCC->GTCCC + GGCAC->GGGAC	1,241	2,154	462	993	1.24	2.24E-03	-1
CATAG->CAAAG + CTATG->CTTTG vs CAAAG->CATAG + CTTTG->CTATG	1,058	1,722	759	1,528	1.24	6.00E-04	2
TCCCG->TCGCG + CGGGA->CGCGA vs TCGCG->TCCCG + CGCGA->CGGGA	113	204	56	125	1.24	3.89E-01	-2
CGTAC->CGAAC + GTACG->GTTCG vs CGAAC->CGTAC + GTTCG->GTACG	24	31	15	24	1.24	7.46E-01	1
AATTG->AAATG + CAATT->CATTT vs AAATG->AATTG + CATTT->CAATT	1,400	2,856	1,745	4,398	1.24	2.24E-06	1
GGTAC->GGCAC + GTACC->GTGCC vs GGCAC->GGTAC + GTGCC->GTACC	1,726	2,603	3,540	6,580	1.23	6.14E-08	0
TGTGT->TGGGT + ACACA->ACCCA vs TGGGT->TGTGT + ACCCA->ACACA	2,327	3,926	2,190	4,552	1.23	4.05E-08	2
AGTTC->AGCTC + GAACT->GAGCT vs AGCTC->AGTTC + GAGCT->GAACT	3,931	6,324	5,825	11,519	1.23	6.62E-15	-1
GCTCT->GCCCT + AGAGC->AGGGC vs GCCCT->GCTCT + AGGGC->AGAGC	3,065	4,808	5,592	10,762	1.23	2.34E-12	2
AGTTA->AGCTA + TAACT->TAGCT vs AGCTA->AGTTA + TAGCT->TAACT	3,468	5,480	3,449	6,683	1.23	4.44E-11	-1
GGTCG->GGACG + CGACC->CGTCC vs GGACG->GGTCG + CGTCC->CGACC	29	65	52	143	1.23	5.60E-01	0
AGTTT->AGCTT + AAACT->AAGCT vs AGCTT->AGTTT + AAGCT->AAACT	6,519	10,370	7,008	13,667	1.23	1.76E-20	-1
AGTTC->AGATC + GAACT->GATCT vs AGATC->AGTTC + GATCT->GAACT	642	1,963	923	3,458	1.23	1.02E-03	-1
CTTGA->CTGGA + TCAAG->TCCAG vs CTGGA->CTTGA + TCCAG->TCAAG	1,985	2,942	903	1,639	1.22	1.17E-04	0
GTTTG->GTCTG + CAAAC->CAGAC vs GTCTG->GTTTG + CAGAC->CAAAC	4,560	7,595	2,583	5,258	1.22	1.22E-10	-2
TTTTC->TTGTC + GAAAA->GACAA vs TTGTC->TTTTC + GACAA->GAAAA	2,977	4,838	3,133	6,218	1.22	1.14E-09	-2
AACAC->AAGAC + GTGTT->GTCTT vs AAGAC->AACAC + GTCTT->GTGTT	1,077	1,756	2,074	4,123	1.22	5.57E-05	0
CATCC->CACCC + GGATG->GGGTG vs CACCC->CATCC + GGGTG->GGATG	4,069	6,607	6,813	13,484	1.22	7.06E-15	1
GCTTT->GCATT + AAAGC->AATGC vs GCATT->GCTTT + AATGC->AAAGC	1,005	1,588	1,533	2,949	1.22	2.37E-04	-1
GGTAG->GGCAG + CTACC->CTGCC vs GGCAG->GGTAG + CTGCC->CTACC	3,463	5,510	5,629	10,896	1.22	2.10E-12	0
TTTCT->TTGCT + AGAAA->AGCAA vs TTGCT->TTTCT + AGCAA->AGAAA	3,205	5,687	4,061	8,762	1.22	6.01E-11	-1
GGCGG->GGGGG + CCGCC->CCCCC vs GGGGG->GGCGG + CCCCC->CCGCC	214	459	1,040	2,710	1.21	4.76E-02	3
GCTAC->GCCAC + GTAGC->GTGGC vs GCCAC->GCTAC + GTGGC->GTAGC	1,975	3,037	3,437	6,412	1.21	1.78E-07	1
GGTAA->GGAAA + TTACC->TTTCC vs GGAAA->GGTAA + TTTCC->TTACC	660	1,165	1,134	2,427	1.21	2.83E-03	1
TCTTA->TCATA + TAAGA->TATGA vs TCATA->TCTTA + TATGA->TAAGA	914	1,441	1,973	3,764	1.21	3.21E-04	-1
AATTA->AAGTA + TAATT->TACTT vs AAGTA->AATTA + TACTT->TAATT	2,788	4,161	2,154	3,884	1.21	4.51E-07	-1
AATCT->AAGCT + AGATT->AGCTT vs AAGCT->AATCT + AGCTT->AGATT	1,174	1,776	2,304	4,211	1.21	7.40E-05	0
TTTAT->TTCAT + ATAAA->ATGAA vs TTCAT->TTTAT + ATGAA->ATAAA	12,630	20,328	7,261	14,110	1.21	4.20E-24	-1
ATTAG->ATCAG + CTAAT->CTGAT vs ATCAG->ATTAG + CTGAT->CTAAT	4,928	8,020	3,858	7,569	1.21	9.22E-12	-1
CTCGG->CTTGG + CCGAG->CCAAG vs CTTGG->CTCGG + CCAAG->CCGAG	6,461	10,748	3,501	7,014	1.20	2.55E-12	1
CTTTT->CTCTT + AAAAG->AAGAG vs CTCTT->CTTTT + AAGAG->AAAAG	5,968	9,483	8,741	16,724	1.20	7.23E-18	-2
TATCG->TAGCG + CGATA->CGCTA vs TAGCG->TATCG + CGCTA->CGATA	27	51	33	75	1.20	7.11E-01	0
GGCGG->GGTGG + CCGCC->CCACC vs GGTGG->GGCGG + CCACC->CCGCC	7,066	11,035	3,253	6,105	1.20	1.19E-11	0
TCCGG->TCGGG + CCGGA->CCCGA vs TCGGG->TCCGG + CCCGA->CCGGA	164	336	437	1,076	1.20	1.35E-01	1
ATTAT->ATCAT + ATAAT->ATGAT vs ATCAT->ATTAT + ATGAT->ATAAT	6,434	10,449	4,865	9,491	1.20	3.41E-14	-1
CTTTC->CTCTC + GAAAG->GAGAG vs CTCTC->CTTTC + GAGAG->GAAAG	4,033	6,538	7,290	14,185	1.20	5.08E-13	-2
TGTAA->TGAAA + TTACA->TTTCA vs TGAAA->TGTAA + TTTCA->TTACA	1,151	1,926	1,376	2,763	1.20	4.92E-04	1
CTTAT->CTAAT + ATAAG->ATTAG vs CTAAT->CTTAT + ATTAG->ATAAG	848	1,223	792	1,370	1.20	6.69E-03	0
TTTGT->TTGGT + ACAAA->ACCAA vs TTGGT->TTTGT + ACCAA->ACAAA	3,442	5,583	2,902	5,642	1.20	2.19E-08	-1
AATTG->AAGTG + CAATT->CACTT vs AAGTG->AATTG + CACTT->CAATT	1,283	2,061	1,605	3,089	1.20	2.47E-04	-1
AGTAG->AGAAG + CTACT->CTTCT vs AGAAG->AGTAG + CTTCT->CTACT	879	1,531	964	2,006	1.19	3.71E-03	1
GCTTA->GCATA + TAAGC->TATGC vs GCATA->GCTTA + TATGC->TAAGC	462	759	1,032	2,024	1.19	1.97E-02	-1
CCTTG->CCATG + CAAGG->CATGG vs CCATG->CCTTG + CATGG->CAAGG	689	1,131	1,355	2,653	1.19	4.67E-03	-1
GTTCA->GTGCA + TGAAC->TGCAC vs GTGCA->GTTCA + TGCAC->TGAAC	753	1,131	1,443	2,584	1.19	4.08E-03	-1
GCTTC->GCCTC + GAAGC->GAGGC vs GCCTC->GCTTC + GAGGC->GAAGC	2,806	4,876	5,626	11,655	1.19	2.93E-09	0
CATAT->CAAAT + ATATG->ATTTG vs CAAAT->CATAT + ATTTG->ATATG	1,294	2,263	2,000	4,161	1.19	1.72E-04	2

Table A.1 (continued)

TATGT->TAGGT + ACATA->ACCTA vs TAGGT->TATGT + ACCTA->ACATA	2,585	4,154	1,088	2,079	1.19	2.34E-04	1
ACTTA->ACCTA + TAAGT->TAGGT vs ACCTA->ACTTA + TAGGT->TAAGT	2,734	4,782	3,717	7,730	1.19	7.48E-08	0
TGTCC->TGACC + GGACA->GGTCA vs TGACC->TGTCC + GGTCA->GGACA	753	2,103	543	1,803	1.19	1.19E-02	0
AGTAC->AGCAC + GTACT->GTGCT vs AGCAC->AGTAC + GTGCT->GTACT	3,198	4,940	4,288	7,872	1.19	1.47E-08	0
TGTAG->TGAAG + CTACA->CTTCA vs TGAAG->TGTAG + CTTCA->CTACA	582	1,041	678	1,441	1.19	2.21E-02	1
GGTAA->GGCAA + TTACC->TTGCC vs GGCAA->GGTAA + TTGCC->TTACC	3,070	4,961	3,678	7,062	1.19	4.81E-08	0
CTTAC->CTCAC + GTAAG->GTGAG vs CTCAC->CTTAC + GTGAG->GTAAG	2,967	4,627	5,554	10,291	1.19	6.24E-09	-1
TTTCC->TTGCC + GGAAA->GGCAA vs TTGCC->TTTCC + GGCAA->GGAAA	2,059	3,904	2,462	5,546	1.19	5.08E-06	-1
TATCC->TACCC + GGATA->GGGTA vs TACCC->TATCC + GGGTA->GGATA	3,743	6,483	4,097	8,419	1.19	2.76E-09	1
ATTGC->ATGGC + GCAAT->GCCAT vs ATGGC->ATTGC + GCCAT->GCAAT	1,337	2,221	667	1,314	1.19	6.39E-03	0
GTTAT->GTAAT + ATAAC->ATTAC vs GTAAT->GTTAT + ATTAC->ATAAC	740	1,113	732	1,305	1.19	1.67E-02	0
CATCT->CACCT + AGATG->AGGTG vs CACCT->CATCT + AGGTG->AGATG	7,083	12,061	8,340	16,817	1.18	1.59E-16	1
GTTTG->GTGTG + CAAAC->CACAC vs GTGTG->GTTTG + CACAC->CAAAC	1,489	2,342	1,491	2,777	1.18	4.87E-04	-2
CCCCG->CCGCG + CGGGG->CGCGG vs CCGCG->CCCCG + CGCGG->CGGGG	235	487	86	211	1.18	3.27E-01	-2
TATTA->TAGTA + TAATA->TACTA vs TAGTA->TATTA + TACTA->TAATA	2,132	3,577	1,135	2,251	1.18	4.58E-04	-1
ACTCT->ACACT + AGAGT->AGTGT vs ACACT->ACTCT + AGTGT->AGAGT	1,241	2,246	1,400	2,995	1.18	8.87E-04	0
GTTAA->GTCAA + TTAAC->TTGAC vs GTCAA->GTTAA + TTGAC->TTAAC	2,088	3,510	2,369	4,703	1.18	1.85E-05	-1
AACAA->AAGAA + TTGTT->TTCTT vs AAGAA->AACAA + TTCTT->TTGTT	4,730	7,845	4,712	9,229	1.18	2.65E-10	0
TACCG->TATCG + CGGTA->CGATA vs TATCG->TACCG + CGATA->CGGTA	166	293	142	296	1.18	3.21E-01	-1
GGTCA->GGGCA + TGACC->TGCCC vs GGGCA->GGTCA + TGCCC->TGACC	466	912	1,946	4,496	1.18	1.45E-02	1
TTTGC->TTGGC + GCAAA->GCCAA vs TTGGC->TTTGC + GCCAA->GCAAA	2,066	3,358	861	1,648	1.18	2.15E-03	-1
ACCGG->ACGGG + CCGGT->CCCGT vs ACGGG->ACCGG + CCCGT->CCGGT	189	348	330	715	1.18	1.99E-01	1
TGTAA->TGCAA + TTACA->TTGCA vs TGCAA->TGTAA + TTGCA->TTACA	5,447	9,213	5,575	11,090	1.18	2.33E-11	0
CCCTG->CCATG + CAGGG->CATGG vs CCATG->CCCTG + CATGG->CAGGG	1,453	2,142	1,347	2,335	1.18	1.43E-03	-1
TATTG->TAGTG + CAATA->CACTA vs TAGTG->TATTG + CACTA->CAATA	956	1,614	808	1,604	1.18	1.02E-02	-1
GATCC->GACCC + GGATC->GGGTC vs GACCC->GATCC + GGGTC->GGATC	1,717	2,843	3,933	7,648	1.17	2.21E-05	1
GATCG->GAGCG + CGATC->CGCTC vs GAGCG->GATCG + CGCTC->CGATC	20	33	80	155	1.17	7.11E-01	0
AATCC->AACCC + GGATT->GGGTT vs AACCC->AATCC + GGGTT->GGATT	3,294	5,042	5,359	9,625	1.17	3.45E-08	1
ATTCG->ATCCG + CGAAT->CGGAT vs ATCCG->ATTCG + CGGAT->CGAAT	83	134	171	324	1.17	4.16E-01	0
GCTCC->GCCCC + GGAGC->GGGGC vs GCCCC->GCTCC + GGGGC->GGAGC	2,041	3,342	4,858	9,329	1.17	3.84E-06	2
ATTTC->ATCTC + GAAAT->GAGAT vs ATCTC->ATTTC + GAGAT->GAAAT	3,629	6,309	4,310	8,785	1.17	3.23E-08	-2
CTTAT->CTCAT + ATAAG->ATGAG vs CTCAT->CTTAT + ATGAG->ATAAG	4,118	6,831	8,289	16,117	1.17	1.03E-10	-1
AATCC->AAGCC + GGATT->GGCTT vs AAGCC->AATCC + GGCTT->GGATT	688	1,091	1,490	2,768	1.17	1.09E-02	0
ATTTG->ATGTG + CAAAT->CACAT vs ATGTG->ATTTG + CACAT->CAAAT	4,073	6,145	1,589	2,807	1.17	4.85E-05	-2
CTTGC->CTGGC + GCAAG->GCCAG vs CTGGC->CTTGC + GCCAG->GCAAG	1,154	1,872	883	1,677	1.17	7.89E-03	0
AGTTG->AGATG + CAACT->CATCT vs AGATG->AGTTG + CATCT->CAACT	1,085	2,260	2,521	6,146	1.17	6.34E-04	-1
CCTTA->CCCTA + TAAGG->TAGGG vs CCCTA->CCTTA + TAGGG->TAAGG	1,827	3,393	3,392	7,368	1.17	2.51E-05	1
TCTAT->TCCAT + ATAGA->ATGGA vs TCCAT->TCTAT + ATGGA->ATAGA	5,410	8,649	6,448	12,056	1.17	4.52E-11	1
CTCCC->CTGCC + GGGAG->GGCAG vs CTGCC->CTCCC + GGCAG->GGGAG	2,835	5,018	1,417	2,933	1.17	1.77E-04	-1
CATCA->CACCA + TGATG->TGGTG vs CACCA->CATCA + TGGTG->TGATG	4,734	7,948	6,680	13,104	1.17	1.70E-10	1
GTTTA->GTGTA + TAAAC->TACAC vs GTGTA->GTTTA + TACAC->TAAAC	1,613	2,378	1,128	1,942	1.17	3.00E-03	-2
CCTAT->CCAAT + ATAGG->ATTGG vs CCAAT->CCTAT + ATTGG->ATAGG	538	866	649	1,219	1.17	5.23E-02	1
AGTAT->AGCAT + ATACT->ATGCT vs AGCAT->AGTAT + ATGCT->ATACT	5,936	9,609	5,673	10,709	1.17	1.16E-10	0
TATTC->TAGTC + GAATA->GACTA vs TAGTC->TATTC + GACTA->GAATA	2,189	3,563	744	1,409	1.16	7.13E-03	-1
GTCGG->GTTGG + CCGAC->CCAAC vs GTTGG->GTCGG + CCAAC->CCGAC	2,966	4,846	1,670	3,169	1.16	1.75E-04	1
GATCA->GACCA + TGATC->TGGTC vs GACCA->GATCA + TGGTC->TGATC	2,432	4,395	3,898	8,181	1.16	6.38E-06	1
TCTCA->TCACA + TGAGA->TGTGA vs TCACA->TCTCA + TGTGA->TGAGA	827	1,373	1,101	2,122	1.16	1.57E-02	0
CTCCT->CTGCT + AGGAG->AGCAG vs CTGCT->CTCCT + AGCAG->AGGAG	2,939	5,087	2,562	5,147	1.16	1.78E-05	-1
ATTAA->ATCAA + TTAAT->TTGAT vs ATCAA->ATTAA + TTGAT->TTAAT	4,702	7,835	3,791	7,328	1.16	1.12E-07	-1
TCTAC->TCAAC + GTAGA->GTTGA vs TCAAC->TCTAC + GTTGA->GTAGA	538	821	523	925	1.16	8.54E-02	1
AGTGT->AGGGT + ACACT->ACCCT vs AGGGT->AGTGT + ACCCT->ACACT	1,943	3,352	2,435	4,867	1.16	1.93E-04	2
GCCGC->GCTGC + GCGGC->GCAGC vs GCTGC->GCCGC + GCAGC->GCGGC	5,028	8,249	2,769	5,262	1.16	1.44E-06	-1
AATCA->AAGCA + TGATT->TGCTT vs AAGCA->AATCA + TGCTT->TGATT	1,311	2,089	2,032	3,750	1.16	1.82E-03	0
CCCGC->CCGGC + GCGGG->GCCGG vs CCGGC->CCCGC + GCCGG->GCGGG	470	1,077	309	820	1.16	1.24E-01	-1
ATTTG->ATCTG + CAAAT->CAGAT vs ATCTG->ATTTG + CAGAT->CAAAT	9,726	16,908	3,897	7,839	1.16	1.08E-09	-2
GGCGC->GGTGC + GCGCC->GCACC vs GGTGC->GGCGC + GCACC->GCGCC	4,119	6,601	1,728	3,204	1.16	9.05E-05	0
TGTTT->TGGTT + AAACA->AACCA vs TGGTT->TGTTT + AACCA->AAACA	2,250	4,406	3,508	7,944	1.16	2.19E-05	-1
GCTTG->GCCTG + CAAGC->CAGGC vs GCCTG->GCTTG + CAGGC->CAAGC	3,063	5,304	4,413	8,827	1.16	1.91E-06	0
TTTAG->TTCAG + CTAAA->CTGAA vs TTCAG->TTTAG + CTGAA->CTAAA	10,469	16,988	6,485	12,137	1.15	1.61E-12	-1
TACCG->TAGCG + CGGTA->CGCTA vs TAGCG->TACCG + CGCTA->CGGTA	36	61	41	80	1.15	7.42E-01	-1
CCTAG->CCCAG + CTAGG->CTGGG vs CCCAG->CCTAG + CTGGG->CTAGG	3,733	5,538	8,095	13,810	1.15	9.38E-08	1
GGTCC->GGGCC + GGACC->GGCCC vs GGGCC->GGTCC + GGCCC->GGACC	373	750	1,576	3,643	1.15	7.00E-02	1
GATTG->GAGTG + CAATC->CACTC vs GAGTG->GATTG + CACTC->CAATC	562	878	946	1,698	1.15	5.97E-02	-1
CGTTG->CGATG + CAACG->CATCG vs CGATG->CGTTG + CATCG->CAACG	34	72	51	124	1.15	7.60E-01	-1
GGTCG->GGGCG + CGACC->CGCCC vs GGGCG->GGTCG + CGCCC->CGACC	29	68	165	444	1.15	7.03E-01	1
GTTGT->GTGGT + ACAAC->ACCAC vs GTGGT->GTTGT + ACCAC->ACAAC	1,196	2,074	1,269	2,524	1.15	1.02E-02	0
TGTTG->TGGTG + CAACA->CACCA vs TGGTG->TGTTG + CACCA->CAACA	1,045	2,084	2,176	4,975	1.15	5.18E-03	0
TATCC->TAGCC + GGATA->GGCTA vs TAGCC->TATCC + GGCTA->GGATA	882	1,455	798	1,507	1.14	4.11E-02	0
TATCA->TAGCA + TGATA->TGCTA vs TAGCA->TATCA + TGCTA->TGATA	1,161	1,871	1,157	2,132	1.14	1.62E-02	0
CTTAA->CTCAA + TTAAG->TTGAG vs CTCAA->CTTAA + TTGAG->TTAAG	3,451	5,630	4,735	8,832	1.14	4.23E-06	-1
TACAC->TAGAC + GTGTA->GTCTA vs TAGAC->TACAC + GTCTA->GTGTA	623	1,092	1,392	2,789	1.14	4.03E-02	0
CTTTA->CTGTA + TAAAG->TACAG vs CTGTA->CTTTA + TACAG->TAAAG	2,586	4,158	1,480	2,720	1.14	1.97E-03	-2
ACTCA->ACACA + TGAGT->TGTGT vs ACACA->ACTCA + TGTGT->TGAGT	1,026	1,911	1,272	2,708	1.14	1.53E-02	0

Table A.1 (continued)

CTCGC->CTTGC + GCGAG->GCAAG vs CTTGC->CTCGC + GCAAG->GCGAG	3,983	6,687	2,124	4,075	1.14	1.28E-04	1
GCTCA->GCACA + TGAGC->TGTGC vs GCACA->GCTCA + TGTGC->TGAGC	523	883	732	1,412	1.14	9.21E-02	0
TACAC->TATAC + GTGTA->GTATA vs TATAC->TACAC + GTATA->GTGTA	7,297	11,645	9,288	16,933	1.14	4.73E-11	0
GTTGC->GTGGC + GCAAC->GCCAC vs GTGGC->GTTGC + GCCAC->GCAAC	769	1,234	618	1,132	1.14	7.44E-02	0
ACTAA->ACAAA + TTAGT->TTTGT vs ACAAA->ACTAA + TTTGT->TTAGT	1,332	2,094	1,518	2,723	1.14	9.41E-03	1
CCTAT->CCCAT + ATAGG->ATGGG vs CCCAT->CCTAT + ATGGG->ATAGG	2,779	4,320	5,409	9,594	1.14	1.93E-05	1
TACAG->TATAG + CTGTA->CTATA vs TATAG->TACAG + CTATA->CTGTA	9,648	15,755	16,751	31,207	1.14	1.09E-15	0
GATCT->GACCT + AGATC->AGGTC vs GACCT->GATCT + AGGTC->AGATC	3,056	5,327	4,652	9,247	1.14	1.28E-05	1
ATTAC->ATCAC + GTAAT->GTGAT vs ATCAC->ATTAC + GTGAT->GTAAT	3,254	5,503	3,237	6,233	1.14	5.74E-05	-1
TACAA->TAGAA + TTGTA->TTCTA vs TAGAA->TACAA + TTCTA->TTGTA	2,421	4,201	3,877	7,648	1.14	1.39E-04	0
GGTGC->GGGGC + GCACC->GCCCC vs GGGGC->GGTGC + GCCCC->GCACC	751	1,507	1,135	2,588	1.14	4.01E-02	2
GACCT->GAGCT + AGGTC->AGCTC vs GAGCT->GACCT + AGCTC->AGGTC	1,037	1,922	1,503	3,162	1.14	1.69E-02	-1
TATCT->TACCT + AGATA->AGGTA vs TACCT->TATCT + AGGTA->AGATA	8,492	15,252	6,388	13,001	1.13	2.46E-09	1
AACAG->AATAG + CTGTT->CTATT vs AATAG->AACAG + CTATT->CTGTT	10,031	16,103	18,544	33,694	1.13	1.08E-14	0
AACTG->AATTG + CAGTT->CAATT vs AATTG->AACTG + CAATT->CAGTT	6,942	11,469	20,413	38,149	1.13	6.99E-12	1
TCTCT->TCCCT + AGAGA->AGGGA vs TCCCT->TCTCT + AGGGA->AGAGA	6,513	10,454	10,099	18,315	1.13	3.26E-09	2
TTTAC->TTAAC + GTAAA->GTTAA vs TTAAC->TTTAC + GTTAA->GTAAA	1,268	1,834	888	1,451	1.13	4.56E-02	-1
GGTGT->GGAGT + ACACC->ACTCC vs GGAGT->GGTGT + ACTCC->ACACC	681	1,268	669	1,407	1.13	9.74E-02	0
GCTTT->GCCTT + AAAGC->AAGGC vs GCCTT->GCTTT + AAGGC->AAAGC	3,708	6,368	5,688	11,029	1.13	8.87E-06	-1
TTTCG->TTCCG + CGAAA->CGGAA vs TTCCG->TTTCG + CGGAA->CGAAA	146	304	234	550	1.13	4.36E-01	-1
CCTAG->CCAAG + CTAGG->CTTGG vs CCAAG->CCTAG + CTTGG->CTAGG	421	677	549	996	1.13	1.80E-01	1
TATCA->TACCA + TGATA->TGGTA vs TACCA->TATCA + TGGTA->TGATA	6,404	11,277	5,535	10,995	1.13	2.66E-07	1
AGTAA->AGCAA + TTACT->TTGCT vs AGCAA->AGTAA + TTGCT->TTACT	5,631	9,054	5,503	9,979	1.13	1.05E-06	0
ACTGC->ACAGC + GCAGT->GCTGT vs ACAGC->ACTGC + GCTGT->GCAGT	1,080	1,636	938	1,602	1.13	5.16E-02	0
GGCGA->GGTGA + TCGCC->TCACC vs GGTGA->GGCGA + TCACC->TCGCC	4,609	7,439	2,850	5,185	1.13	1.25E-04	0
AATCT->AACCT + AGATT->AGGTT vs AACCT->AATCT + AGGTT->AGATT	6,158	9,998	7,771	14,220	1.13	6.26E-08	1
AACTC->AAGTC + GAGTT->GACTT vs AAGTC->AACTC + GACTT->GAGTT	1,958	3,489	1,748	3,510	1.13	5.72E-03	0
CTCGA->CTTGA + TCGAG->TCAAG vs CTTGA->CTCGA + TCAAG->TCGAG	4,041	6,525	3,240	5,888	1.13	1.31E-04	1
AGTGC->AGGGC + GCACT->GCCCT vs AGGGC->AGTGC + GCCCT->GCACT	1,573	2,764	1,175	2,322	1.12	2.16E-02	2
TCTCC->TCCCC + GGAGA->GGGGA vs TCCCC->TCTCC + GGGGA->GGAGA	3,717	6,227	7,357	13,861	1.12	7.53E-06	2
GACTG->GATTG + CAGTC->CAATC vs GATTG->GACTG + CAATC->CAGTC	4,026	6,709	10,904	20,412	1.12	1.26E-06	1
AGCCC->AGGCC + GGGCT->GGCCT vs AGGCC->AGCCC + GGCCT->GGGCT	1,768	3,228	1,321	2,709	1.12	1.51E-02	-1
CATTA->CAGTA + TAATG->TACTG vs CAGTA->CATTA + TACTG->TAATG	1,346	2,211	1,160	2,140	1.12	3.15E-02	-1
ATTTC->ATGTC + GAAAT->GACAT vs ATGTC->ATTTC + GACAT->GAAAT	2,978	4,833	1,021	1,859	1.12	1.78E-02	-2
ATTCC->ATGCC + GGAAT->GGCAT vs ATGCC->ATTCC + GGCAT->GGAAT	1,267	2,106	1,222	2,278	1.12	3.57E-02	-1
TGTAC->TGCAC + GTACA->GTGCA vs TGCAC->TGTAC + GTGCA->GTACA	3,481	5,591	5,348	9,632	1.12	6.42E-05	0
AATTT->AAGTT + AAATT->AACTT vs AAGTT->AATTT + AACTT->AAATT	3,141	5,255	3,013	5,643	1.12	7.65E-04	-1
GATCG->GACCG + CGATC->CGGTC vs GACCG->GATCG + CGGTC->CGATC	87	165	195	414	1.12	5.49E-01	1
CGTAT->CGCAT + ATACG->ATGCG vs CGCAT->CGTAT + ATGCG->ATACG	133	230	187	362	1.12	5.08E-01	0
AACAT->AAGAT + ATGTT->ATCTT vs AAGAT->AACAT + ATCTT->ATGTT	1,556	2,811	2,967	5,999	1.12	6.45E-03	0
TTTTT->TTATT + AAAAA->AATAA vs TTATT->TTTTT + AATAA->AAAAA	2,990	5,513	3,774	7,785	1.12	3.72E-04	-3
CTTGG->CTGGG + CCAAG->CCCAG vs CTGGG->CTTGG + CCCAG->CCAAG	1,426	2,399	1,935	3,641	1.12	1.70E-02	1
TTTAA->TTCAA + TTAAA->TTGAA vs TTCAA->TTTAA + TTGAA->TTAAA	10,035	16,824	5,793	10,861	1.12	1.28E-07	-1
TGCCC->TGGCC + GGGCA->GGCCA vs TGGCC->TGCCC + GGCCA->GGGCA	1,065	2,071	911	1,981	1.12	6.01E-02	-1
GCTAT->GCAAT + ATAGC->ATTGC vs GCAAT->GCTAT + ATTGC->ATAGC	529	817	630	1,086	1.12	1.83E-01	1
CGTTA->CGATA + TAACG->TATCG vs CGATA->CGTTA + TATCG->TAACG	26	55	25	59	1.11	9.23E-01	-1
CTTAC->CTAAC + GTAAG->GTTAG vs CTAAC->CTTAC + GTTAG->GTAAG	550	880	364	649	1.11	2.71E-01	0
TCTCT->TCACT + AGAGA->AGTGA vs TCACT->TCTCT + AGTGA->AGAGA	1,099	1,892	1,451	2,783	1.11	4.62E-02	0
AATTT->AACTT + AAATT->AAGTT vs AACTT->AATTT + AAGTT->AAATT	11,970	19,604	9,775	17,827	1.11	9.09E-10	-1
TCTAA->TCAAA + TTAGA->TTTGA vs TCAAA->TCTAA + TTTGA->TTAGA	1,033	1,571	1,263	2,138	1.11	6.68E-02	1
AATAT->AAAAT + ATATT->ATTTT vs AAAAT->AATAT + ATTTT->ATATT	2,066	3,751	3,498	7,066	1.11	3.26E-03	2
GCTTG->GCATG + CAAGC->CATGC vs GCATG->GCTTG + CATGC->CAAGC	474	810	1,257	2,387	1.11	1.62E-01	-1
TCTTT->TCCTT + AAAGA->AAGGA vs TCCTT->TCTTT + AAGGA->AAAGA	9,072	15,348	11,127	20,882	1.11	1.17E-08	-1
TGTAG->TGCAG + CTACA->CTGCA vs TGCAG->TGTAG + CTGCA->CTACA	6,987	11,738	7,095	13,202	1.11	2.78E-06	0
CTTCT->CTGCT + AGAAG->AGCAG vs CTGCT->CTTCT + AGCAG->AGAAG	2,061	3,633	3,156	6,160	1.11	6.54E-03	-1
GACCA->GAGCA + TGGTC->TGCTC vs GAGCA->GACCA + TGCTC->TGGTC	1,398	2,347	1,331	2,473	1.11	5.07E-02	-1
GCTAA->GCAAA + TTAGC->TTTGC vs GCAAA->GCTAA + TTTGC->TTAGC	653	987	797	1,333	1.11	1.76E-01	1
AATCA->AACCA + TGATT->TGGTT vs AACCA->AATCA + TGGTT->TGATT	5,172	8,594	7,235	13,298	1.11	2.19E-05	1
CGTGA->CGCGA + TCACG->TCGCG vs CGCGA->CGTGA + TCGCG->TCACG	201	340	310	580	1.11	4.61E-01	0
TCTAC->TCCAC + GTAGA->GTGGA vs TCCAC->TCTAC + GTGGA->GTAGA	4,099	6,186	6,206	10,354	1.11	1.93E-04	1
TCTCA->TCCCA + TGAGA->TGGGA vs TCCCA->TCTCA + TGGGA->TGAGA	3,820	6,394	7,683	14,212	1.11	1.17E-04	2
ATTCT->ATGCT + AGAAT->AGCAT vs ATGCT->ATTCT + AGCAT->AGAAT	2,235	3,727	2,050	3,776	1.10	1.52E-02	-1
TATTG->TAATG + CAATA->CATTA vs TAATG->TATTG + CATTA->CAATA	848	1,700	973	2,154	1.10	1.16E-01	0
GTCGC->GTTGC + GCGAC->GCAAC vs GTTGC->GTCGC + GCAAC->GCGAC	1,829	3,131	1,023	1,933	1.10	6.01E-02	1
TATTT->TACTT + AAATA->AAGTA vs TACTT->TATTT + AAGTA->AAATA	16,817	30,340	9,009	17,939	1.10	2.19E-09	-1
GTTTC->GTGTC + GAAAC->GACAC vs GTGTC->GTTTC + GACAC->GAAAC	1,138	1,943	865	1,629	1.10	1.11E-01	-2
CACAT->CAGAT + ATGTG->ATCTG vs CAGAT->CACAT + ATCTG->ATGTG	1,101	2,066	3,055	6,323	1.10	3.65E-02	0
GATGC->GAGGC + GCATC->GCCTC vs GAGGC->GATGC + GCCTC->GCATC	1,249	1,959	860	1,486	1.10	1.18E-01	1
GGTTA->GGATA + TAACC->TATCC vs GGATA->GGTTA + TATCC->TAACC	430	1,000	1,167	2,988	1.10	1.96E-01	-1
AATTC->AAGTC + GAATT->GACTT vs AAGTC->AATTC + GACTT->GAATT	1,371	2,359	1,467	2,778	1.10	5.94E-02	-1
TCTAG->TCAAG + CTAGA->CTTGA vs TCAAG->TCTAG + CTTGA->CTAGA	575	831	654	1,040	1.10	2.46E-01	1
CGCTG->CGATG + CAGCG->CATCG vs CGATG->CGCTG + CATCG->CAGCG	108	207	37	78	1.10	8.02E-01	0
CATGT->CAGGT + ACATG->ACCTG vs CAGGT->CATGT + ACCTG->ACATG	2,197	3,775	1,446	2,732	1.10	3.65E-02	1

Table A.1 (continued)

TCTAA->TCCAA + TTAGA->TTGGA vs TCCAA->TCTAA + TTGGA->TTAGA	5,255	8,483	5,966	10,581	1.10	1.61E-04	1
CCTCA->CCCCA + TGAGG->TGGGG vs CCCCA->CCTCA + TGGGG->TGAGG	2,604	4,082	7,839	13,499	1.10	2.02E-03	2
ATTCA->ATGCA + TGAAT->TGCAT vs ATGCA->ATTCA + TGCAT->TGAAT	2,095	3,614	2,046	3,874	1.10	2.47E-02	-1
CGTAG->CGCAG + CTACG->CTGCG vs CGCAG->CGTAG + CTGCG->CTACG	227	400	363	702	1.10	4.66E-01	0
ATCGA->ATTGA + TCGAT->TCAAT vs ATTGA->ATCGA + TCAAT->TCGAT	3,495	5,860	3,548	6,526	1.10	3.38E-03	1
CCTAC->CCAAC + GTAGG->GTTGG vs CCAAC->CCTAC + GTTGG->GTAGG	407	712	357	685	1.10	3.86E-01	1
TTTAC->TTCAC + GTAAA->GTGAA vs TTCAC->TTTAC + GTGAA->GTAAA	6,122	10,293	4,924	9,078	1.10	2.31E-04	-1
TATCA->TAACA + TGATA->TGTTA vs TAACA->TATCA + TGTTA->TGATA	1,622	4,092	774	2,141	1.10	1.02E-01	1
TGTTC->TGATC + GAACA->GATCA vs TGATC->TGTTC + GATCA->GAACA	563	1,724	804	2,697	1.10	1.96E-01	-1
ACTAG->ACAAG + CTAGT->CTTGT vs ACAAG->ACTAG + CTTGT->CTAGT	634	958	741	1,226	1.09	2.50E-01	1
GTCCA->GTGCA + TGGAC->TGCAC vs GTGCA->GTCCA + TGCAC->TGGAC	1,276	2,216	514	977	1.09	2.09E-01	-1
TTCCC->TTGCC + GGGAA->GGCAA vs TTGCC->TTCCC + GGCAA->GGGAA	2,264	4,364	1,031	2,175	1.09	7.01E-02	-1
CACTG->CATTG + CAGTG->CAATG vs CATTG->CACTG + CAATG->CAGTG	8,288	13,551	22,393	40,062	1.09	7.50E-08	1
ACTAA->ACCAA + TTAGT->TTGGT vs ACCAA->ACTAA + TTGGT->TTAGT	3,873	6,381	4,325	7,790	1.09	2.45E-03	1
AATGT->AAGGT + ACATT->ACCTT vs AAGGT->AATGT + ACCTT->ACATT	2,649	4,557	2,386	4,484	1.09	1.91E-02	1
CCTTC->CCCTC + GAAGG->GAGGG vs CCCTC->CCTTC + GAGGG->GAAGG	3,280	5,736	8,327	15,899	1.09	1.18E-03	1
GATCT->GAGCT + AGATC->AGCTC vs GAGCT->GATCT + AGCTC->AGATC	926	1,500	1,251	2,208	1.09	1.54E-01	0
AACTA->AAGTA + TAGTT->TACTT vs AAGTA->AACTA + TACTT->TAGTT	2,992	5,389	2,760	5,411	1.09	1.48E-02	0
CCTAA->CCAAA + TTAGG->TTTGG vs CCAAA->CCTAA + TTTGG->TTAGG	786	1,293	883	1,581	1.09	2.19E-01	1
ACTCA->ACCCA + TGAGT->TGGGT vs ACCCA->ACTCA + TGGGT->TGAGT	2,891	4,673	6,084	10,691	1.09	5.94E-03	2
ACTAC->ACCAC + GTAGT->GTGGT vs ACCAC->ACTAC + GTGGT->GTAGT	2,541	4,180	4,246	7,591	1.09	1.40E-02	1
CGCCT->CGGCT + AGGCG->AGCCG vs CGGCT->CGCCT + AGCCG->AGGCG	75	147	108	230	1.09	7.85E-01	0
CTCTC->CTGTC + GAGAG->GACAG vs CTGTC->CTCTC + GACAG->GAGAG	6,964	11,307	3,278	5,782	1.09	3.23E-03	0
TACTC->TAGTC + GAGTA->GACTA vs TAGTC->TACTC + GACTA->GAGTA	1,342	2,516	1,184	2,411	1.09	1.28E-01	0
TCTGC->TCAGC + GCAGA->GCTGA vs TCAGC->TCTGC + GCTGA->GCAGA	889	1,387	897	1,520	1.09	2.21E-01	0
AGTTG->AGCTG + CAACT->CAGCT vs AGCTG->AGTTG + CAGCT->CAACT	8,885	15,802	5,485	10,589	1.09	2.22E-04	-1
TATAT->TAAAT + ATATA->ATTTA vs TAAAT->TATAT + ATTTA->ATATA	1,816	3,107	2,034	3,777	1.09	6.16E-02	2
CCCCA->CCACA + TGGGG->TGTGG vs CCACA->CCCCA + TGTGG->TGGGG	1,883	3,415	1,379	2,714	1.09	8.74E-02	-2
AGTAA->AGAAA + TTACT->TTTCT vs AGAAA->AGTAA + TTTCT->TTACT	1,411	2,611	1,824	3,659	1.08	9.05E-02	1
CACAC->CAGAC + GTGTG->GTCTG vs CAGAC->CACAC + GTCTG->GTGTG	1,004	1,840	1,828	3,626	1.08	1.41E-01	0
CACTC->CAGTC + GAGTG->GACTG vs CAGTC->CACTC + GACTG->GAGTG	2,073	3,742	1,373	2,682	1.08	9.33E-02	0
GATGT->GAGGT + ACATC->ACCTC vs GAGGT->GATGT + ACCTC->ACATC	1,440	2,448	1,075	1,975	1.08	1.61E-01	1
TACAT->TAGAT + ATGTA->ATCTA vs TAGAT->TACAT + ATCTA->ATGTA	1,081	1,965	2,799	5,495	1.08	1.15E-01	0
CCCCG->CCTCG + CGGGG->CGAGG vs CCTCG->CCCCG + CGAGG->CGGGG	690	1,242	160	311	1.08	5.60E-01	-2
TCTTC->TCCTC + GAAGA->GAGGA vs TCCTC->TCTTC + GAGGA->GAAGA	5,479	9,541	9,945	18,693	1.08	5.26E-04	0
TATTA->TACTA + TAATA->TAGTA vs TACTA->TATTA + TAGTA->TAATA	9,404	17,273	4,779	9,464	1.08	1.05E-03	-1
AGTAG->AGCAG + CTACT->CTGCT vs AGCAG->AGTAG + CTGCT->CTACT	7,117	12,047	6,638	12,108	1.08	8.56E-04	0
TCTGT->TCAGT + ACAGA->ACTGA vs TCAGT->TCTGT + ACTGA->ACAGA	1,294	2,086	1,579	2,742	1.08	1.53E-01	0
CACAA->CAGAA + TTGTG->TTCTG vs CAGAA->CACAA + TTCTG->TTGTG	2,642	4,759	3,436	6,660	1.08	3.38E-02	0
AGTTT->AGATT + AAACT->AATCT vs AGATT->AGTTT + AATCT->AAACT	1,566	4,567	1,864	5,849	1.08	9.01E-02	-1
CTTTT->CTGTT + AAAAG->AACAG vs CTGTT->CTTTT + AACAG->AAAAG	2,750	4,659	2,360	4,302	1.08	5.51E-02	-2
AATAT->AACAT + ATATT->ATGTT vs AACAT->AATAT + ATGTT->ATATT	14,653	24,020	13,131	23,158	1.08	2.99E-06	0
CCCTT->CCATT + AAGGG->AATGG vs CCATT->CCCTT + AATGG->AAGGG	1,675	2,598	1,473	2,458	1.08	1.46E-01	-1
TATGG->TAGGG + CCATA->CCCTA vs TAGGG->TATGG + CCCTA->CCATA	1,011	1,548	665	1,095	1.08	3.27E-01	1
ACTCT->ACCCT + AGAGT->AGGGT vs ACCCT->ACTCT + AGGGT->AGAGT	3,967	6,557	6,183	10,990	1.08	7.68E-03	2
TTTAT->TTAAT + ATAAA->ATTAA vs TTAAT->TTTAT + ATTAA->ATAAA	3,139	4,514	1,904	2,943	1.07	7.82E-02	-1
CATAC->CAAAC + GTATG->GTTTG vs CAAAC->CATAC + GTTTG->GTATG	882	1,635	665	1,325	1.07	3.15E-01	2
TATGA->TAGGA + TCATA->TCCTA vs TAGGA->TATGA + TCCTA->TCATA	2,366	3,589	902	1,469	1.07	1.98E-01	1
GATTC->GACTC + GAATC->GAGTC vs GACTC->GATTC + GAGTC->GAATC	3,145	5,603	4,034	7,705	1.07	2.80E-02	-1
AGTCC->AGACC + GGACT->GGTCT vs AGACC->AGTCC + GGTCT->GGACT	705	1,966	621	1,856	1.07	3.38E-01	0
AATTC->AACTC + GAATT->GAGTT vs AACTC->AATTC + GAGTT->GAATT	5,522	9,488	6,735	12,390	1.07	4.54E-03	-1
ACTCG->ACCCG + CGAGT->CGGGT vs ACCCG->ACTCG + CGGGT->CGAGT	128	200	377	630	1.07	7.24E-01	2
CCTTC->CCATC + GAAGG->GATGG vs CCATC->CCTTC + GATGG->GAAGG	682	1,226	1,112	2,136	1.07	3.36E-01	-1
TCTCG->TCCCG + CGAGA->CGGGA vs TCCCG->TCTCG + CGGGA->CGAGA	152	256	489	880	1.07	6.77E-01	2
TGCAT->TGGAT + ATGCA->ATCCA vs TGGAT->TGCAT + ATCCA->ATGCA	835	1,465	2,074	3,886	1.07	2.48E-01	1
CTTTG->CTGTG + CAAAG->CACAG vs CTGTG->CTTTG + CACAG->CAAAG	1,828	3,200	2,050	3,832	1.07	1.37E-01	-2
ACTCC->ACCCC + GGAGT->GGGGT vs ACCCC->ACTCC + GGGGT->GGAGT	2,057	3,418	5,042	8,940	1.07	7.09E-02	2
TATTT->TAGTT + AAATA->AACTA vs TAGTT->TATTT + AACTA->AAATA	3,942	6,772	1,687	3,090	1.07	1.08E-01	-1
TGCAA->TGGAA + TTGCA->TTCCA vs TGGAA->TGCAA + TTCCA->TTGCA	1,564	2,913	3,034	6,018	1.06	1.39E-01	1
CCCCG->CCACG + CGGGG->CGTGG vs CCACG->CCCCG + CGTGG->CGGGG	161	317	94	197	1.06	8.19E-01	-2
AATGA->AAGGA + TCATT->TCCTT vs AAGGA->AATGA + TCCTT->TCATT	3,055	4,823	2,413	4,054	1.06	9.88E-02	1
GACTC->GAATC + GAGTC->GATTC vs GAATC->GACTC + GATTC->GAGTC	814	1,299	1,342	2,278	1.06	3.43E-01	1
ACCGC->ACGGC + GCGGT->GCCGT vs ACGGC->ACCGC + GCCGT->GCGGT	180	386	196	447	1.06	7.36E-01	0
GATAT->GACAT + ATATC->ATGTC vs GACAT->GATAT + ATGTC->ATATC	5,524	9,734	6,633	12,421	1.06	1.20E-02	0
ATTGT->ATGGT + ACAAT->ACCAT vs ATGGT->ATTGT + ACCAT->ACAAT	2,172	3,919	1,733	3,320	1.06	1.73E-01	0
GATTT->GAGTT + AAATC->AACTC vs GAGTT->GATTT + AACTC->AAATC	1,647	2,980	1,219	2,341	1.06	2.58E-01	-1
CGCCG->CGACG + CGGCG->CGTCG vs CGACG->CGCCG + CGTCG->CGGCG	55	95	12	22	1.06	1.00E+00	-1
ACTAT->ACCAT + ATAGT->ATGGT vs ACCAT->ACTAT + ATGGT->ATAGT	4,045	6,835	4,950	8,874	1.06	4.01E-02	1
CATAT->CACAT + ATATG->ATGTG vs CACAT->CATAT + ATGTG->ATATG	9,614	16,325	10,608	19,101	1.06	1.62E-03	0
CGTGA->CGAGA + TCACG->TCTCG vs CGAGA->CGTGA + TCTCG->TCACG	54	110	25	54	1.06	9.37E-01	0
CGCGC->CGGGC + GCGCG->GCCCG vs CGGGC->CGCGC + GCCCG->GCGCG	73	156	91	206	1.06	8.40E-01	2
TATTT->TAATT + AAATA->AATTA vs TAATT->TATTT + AATTA->AAATA	2,554	6,065	1,773	4,453	1.06	1.65E-01	-1
AATCG->AACCG + CGATT->CGGTT vs AACCG->AATCG + CGGTT->CGATT	161	281	266	491	1.06	7.36E-01	1

Table A.1 (continued)

CACTA->CAGTA + TAGTG->TACTG vs CAGTA->CACTA + TACTG->TAGTG	2,224	4,080	1,784	3,460	1.06	1.99E-01	0
CGTGC->CGGGC + GCACG->GCCCG vs CGGGC->CGTGC + GCCCG->GCACG	80	148	91	178	1.06	8.40E-01	2
GCTCC->GCACC + GGAGC->GGTGC vs GCACC->GCTCC + GGTGC->GGAGC	494	910	470	914	1.06	5.73E-01	0
TACAA->TATAA + TTGTA->TTATA vs TATAA->TACAA + TTATA->TTGTA	9,024	15,376	14,172	25,484	1.06	2.50E-03	0
GATAG->GAGAG + CTATC->CTCTC vs GAGAG->GATAG + CTCTC->CTATC	1,426	2,470	1,145	2,090	1.05	3.60E-01	0
CGCAT->CGGAT + ATGCG->ATCCG vs CGGAT->CGCAT + ATCCG->ATGCG	21	41	70	144	1.05	9.34E-01	1
AATAG->AAAAG + CTATT->CTTTT vs AAAAG->AATAG + CTTTT->CTATT	1,368	2,400	1,167	2,156	1.05	3.72E-01	2
TACTG->TATTG + CAGTA->CAATA vs TATTG->TACTG + CAATA->CAGTA	5,955	10,470	17,205	31,814	1.05	1.18E-02	1
TGTTT->TGATT + AAACA->AATCA vs TGATT->TGTTT + AATCA->AAACA	1,449	4,164	1,579	4,767	1.05	3.04E-01	-1
GATGA->GAGGA + TCATC->TCCTC vs GAGGA->GATGA + TCCTC->TCATC	1,812	2,868	1,266	2,105	1.05	3.57E-01	1
GGTTG->GGATG + CAACC->CATCC vs GGATG->GGTTG + CATCC->CAACC	630	1,362	1,890	4,292	1.05	4.56E-01	-1
CATTC->CACTC + GAATG->GAGTG vs CACTC->CATTC + GAGTG->GAATG	5,402	9,386	6,509	11,879	1.05	4.82E-02	-1
CGTAA->CGCAA + TTACG->TTGCG vs CGCAA->CGTAA + TTGCG->TTACG	94	171	111	212	1.05	8.57E-01	0
CCTCT->CCCCT + AGAGG->AGGGG vs CCCCT->CCTCT + AGGGG->AGAGG	3,800	6,428	7,395	13,131	1.05	7.56E-02	2
TGTCT->TGACT + AGACA->AGTCA vs TGACT->TGTCT + AGTCA->AGACA	1,147	3,473	1,006	3,197	1.05	3.97E-01	0
TCTAG->TCCAG + CTAGA->CTGGA vs TCCAG->TCTAG + CTGGA->CTAGA	6,182	9,894	7,950	13,341	1.05	4.12E-02	1
TTTCG->TTGCG + CGAAA->CGCAA vs TTGCG->TTTCG + CGCAA->CGAAA	45	96	51	114	1.05	9.49E-01	-1
CATAA->CACAA + TTATG->TTGTG vs CACAA->CATAA + TTGTG->TTATG	7,467	12,990	6,657	12,122	1.05	4.54E-02	0
CATGA->CAGGA + TCATG->TCCTG vs CAGGA->CATGA + TCCTG->TCATG	2,317	3,596	1,235	2,005	1.05	3.87E-01	1
GTCTC->GTGTC + GAGAC->GACAC vs GTGTC->GTCTC + GACAC->GAGAC	1,937	3,529	551	1,050	1.05	5.31E-01	0
TATAG->TAAAG + CTATA->CTTTA vs TAAAG->TATAG + CTTTA->CTATA	848	1,498	678	1,252	1.05	5.75E-01	2
CTTTC->CTGTC + GAAAG->GACAG vs CTGTC->CTTTC + GACAG->GAAAG	1,958	3,510	1,408	2,638	1.05	3.82E-01	-2
GACAC->GATAC + GTGTC->GTATC vs GATAC->GACAC + GTATC->GTGTC	4,334	7,697	4,180	7,751	1.04	1.46E-01	0
ATCCT->ATGCT + AGGAT->AGCAT vs ATGCT->ATCCT + AGCAT->AGGAT	1,930	3,582	1,285	2,490	1.04	4.04E-01	-1
ACTAG->ACCAG + CTAGT->CTGGT vs ACCAG->ACTAG + CTGGT->CTAGT	3,823	6,186	5,113	8,626	1.04	1.62E-01	1
CTTTC->CTATC + GAAAG->GATAG vs CTATC->CTTTC + GATAG->GAAAG	587	1,083	583	1,121	1.04	6.69E-01	-2
GGTGA->GGAGA + TCACC->TCTCC vs GGAGA->GGTGA + TCTCC->TCACC	778	1,430	838	1,604	1.04	5.92E-01	0
AGTCG->AGACG + CGACT->CGTCT vs AGACG->AGTCG + CGTCT->CGACT	34	82	45	113	1.04	9.44E-01	0
GACAA->GAGAA + TTGTC->TTCTC vs GAGAA->GACAA + TTCTC->TTGTC	1,760	3,347	3,345	6,620	1.04	3.36E-01	0
GCTGG->GCAGG + CCAGC->CCTGC vs GCAGG->GCTGG + CCTGC->CCAGC	893	1,421	979	1,621	1.04	5.92E-01	0
ACTGG->ACAGG + CCAGT->CCTGT vs ACAGG->ACTGG + CCTGT->CCAGT	1,106	1,856	995	1,737	1.04	5.49E-01	0
ACTTC->ACCTC + GAAGT->GAGGT vs ACCTC->ACTTC + GAGGT->GAAGT	3,322	6,019	6,377	12,002	1.04	1.96E-01	0
GATTT->GACTT + AAATC->AAGTC vs GACTT->GATTT + AAGTC->AAATC	6,725	12,008	5,533	10,262	1.04	1.27E-01	-1
TATGC->TAGGC + GCATA->GCCTA vs TAGGC->TATGC + GCCTA->GCATA	1,606	2,467	482	768	1.04	6.77E-01	1
TACCA->TAGCA + TGGTA->TGCTA vs TAGCA->TACCA + TGCTA->TGGTA	1,559	2,931	1,492	2,907	1.04	5.06E-01	-1
TGCAG->TGGAG + CTGCA->CTCCA vs TGGAG->TGCAG + CTCCA->CTGCA	1,384	2,435	3,515	6,400	1.03	4.61E-01	1
TTTGG->TTGGG + CCAAA->CCCAA vs TTGGG->TTTGG + CCCAA->CCAAA	2,773	4,479	1,902	3,179	1.03	4.32E-01	0
TATAA->TAAAA + TTATA->TTTTA vs TAAAA->TATAA + TTTTA->TTATA	3,204	5,716	1,815	3,350	1.03	4.28E-01	2
GCTCG->GCGCG + CGAGC->CGCGC vs GCGCG->GCTCG + CGCGC->CGAGC	36	69	116	230	1.03	9.51E-01	0
CGTGT->CGCGT + ACACG->ACGCG vs CGCGT->CGTGT + ACGCG->ACACG	287	484	413	720	1.03	8.05E-01	0
AATAA->AACAA + TTATT->TTGTT vs AACAA->AATAA + TTGTT->TTATT	13,572	22,393	10,782	18,384	1.03	6.25E-02	0
GACAG->GATAG + CTGTC->CTATC vs GATAG->GACAG + CTATC->CTGTC	5,925	10,745	8,853	16,579	1.03	1.61E-01	0
GATAA->GACAA + TTATC->TTGTC vs GACAA->GATAA + TTGTC->TTATC	6,039	10,862	4,833	8,959	1.03	2.64E-01	0
GATTA->GAATA + TAATC->TATTC vs GAATA->GATTA + TATTC->TAATC	692	1,608	851	2,038	1.03	7.03E-01	0
AGCAA->AGGAA + TTGCT->TTCCT vs AGGAA->AGCAA + TTCCT->TTGCT	2,888	5,371	2,876	5,512	1.03	4.28E-01	1
CCTAA->CCCAA + TTAGG->TTGGG vs CCCAA->CCTAA + TTGGG->TTAGG	2,960	4,857	4,735	7,997	1.03	3.98E-01	1
CGTAA->CGGAA + TTACG->TTCCG vs CGGAA->CGTAA + TTCCG->TTACG	23	46	69	142	1.03	1.00E+00	1
CATCA->CAACA + TGATG->TGTTG vs CAACA->CATCA + TGTTG->TGATG	1,963	4,202	809	1,781	1.03	6.77E-01	1
TTTTC->TTATC + GAAAA->GATAA vs TTATC->TTTTC + GATAA->GAAAA	1,317	2,412	2,064	3,887	1.03	6.00E-01	-2
ACTTG->ACCTG + CAAGT->CAGGT vs ACCTG->ACTTG + CAGGT->CAAGT	4,771	8,530	5,332	9,792	1.03	3.45E-01	0
CGTCG->CGCCG + CGACG->CGGCG vs CGCCG->CGTCG + CGGCG->CGACG	37	70	176	342	1.03	9.55E-01	1
AATAC->AACAC + GTATT->GTGTT vs AACAC->AATAC + GTGTT->GTATT	8,143	13,300	7,326	12,287	1.03	2.45E-01	0
CGTGG->CGAGG + CCACG->CCTCG vs CGAGG->CGTGG + CCTCG->CCACG	89	156	60	108	1.03	9.58E-01	0
CCCCT->CCACT + AGGGG->AGTGG vs CCACT->CCCCT + AGTGG->AGGGG	1,888	3,483	1,522	2,882	1.03	6.28E-01	-2
TATCT->TAACT + AGATA->AGTTA vs TAACT->TATCT + AGTTA->AGATA	1,984	5,632	790	2,301	1.03	6.88E-01	1
CATAA->CAAAA + TTATG->TTTTG vs CAAAA->CATAA + TTTTG->TTATG	2,230	4,225	1,061	2,060	1.02	6.77E-01	2
GCTAC->GCAAC + GTAGC->GTTGC vs GCAAC->GCTAC + GTTGC->GTAGC	352	608	317	561	1.02	8.70E-01	1
ACTTT->ACCTT + AAAGT->AAGGT vs ACCTT->ACTTT + AAGGT->AAAGT	5,024	8,781	7,489	13,396	1.02	3.78E-01	-1
GCTAG->GCAAG + CTAGC->CTTGC vs GCAAG->GCTAG + CTTGC->CTAGC	328	547	452	771	1.02	8.77E-01	1
CTCAT->CTGAT + ATGAG->ATCAG vs CTGAT->CTCAT + ATCAG->ATGAG	1,455	2,535	2,332	4,155	1.02	6.79E-01	0
CGTGC->CGAGC + GCACG->GCTCG vs CGAGC->CGTGC + GCTCG->GCACG	52	120	25	59	1.02	1.00E+00	0
AACAG->AAGAG + CTGTT->CTCTT vs AAGAG->AACAG + CTCTT->CTGTT	4,895	8,471	5,564	9,837	1.02	4.58E-01	0
AATAC->AAAAC + GTATT->GTTTT vs AAAAC->AATAC + GTTTT->GTATT	1,131	2,162	1,101	2,149	1.02	7.68E-01	2
ATCAC->ATGAC + GTGAT->GTCAT vs ATGAC->ATCAC + GTCAT->GTGAT	1,101	2,053	838	1,593	1.02	8.20E-01	0
AGCAC->AGGAC + GTGCT->GTCCT vs AGGAC->AGCAC + GTCCT->GTGCT	980	1,746	1,305	2,370	1.02	8.03E-01	1
CATTT->CACTT + AAATG->AAGTG vs CACTT->CATTT + AAGTG->AAATG	12,869	22,161	8,897	15,617	1.02	3.33E-01	-1
GGTGG->GGAGG + CCACC->CCTCC vs GGAGG->GGTGG + CCTCC->CCACC	802	1,450	1,132	2,086	1.02	8.19E-01	0
CACTA->CATTA + TAGTG->TAATG vs CATTA->CACTA + TAATG->TAGTG	4,081	7,524	5,353	10,058	1.02	5.37E-01	1
GTCGA->GTTGA + TCGAC->TCAAC vs GTTGA->GTCGA + TCAAC->TCGAC	2,091	3,564	1,823	3,165	1.02	7.32E-01	1
CTCAC->CTGAC + GTGAG->GTCAG vs CTGAC->CTCAC + GTCAG->GTGAG	1,107	2,045	1,399	2,631	1.02	8.00E-01	0
GACTA->GATTA + TAGTC->TAATC vs GATTA->GACTA + TAATC->TAGTC	2,373	4,560	3,137	6,132	1.02	6.94E-01	1
ATTTC->ATATC + GAAAT->GATAT vs ATATC->ATTTC + GATAT->GAAAT	1,547	3,034	729	1,454	1.02	8.46E-01	-2
CCTTT->CCCTT + AAAGG->AAGGG vs CCCTT->CCTTT + AAGGG->AAAGG	4,613	8,175	7,854	14,151	1.02	5.54E-01	0

Table A.1 (continued)

GTCAA->GTGAA + TTGAC->TTCAC vs GTGAA->GTCAA + TTCAC->TTGAC	1,587	2,986	1,110	2,122	1.02	8.20E-01	0
ACCGA->ACGGA + TCGGT->TCCGT vs ACGGA->ACCGA + TCCGT->TCGGT	232	493	151	326	1.02	9.84E-01	0
ATCTC->ATGTC + GAGAT->GACAT vs ATGTC->ATCTC + GACAT->GAGAT	2,389	4,546	622	1,202	1.02	8.65E-01	0
TGTGG->TGAGG + CCACA->CCTCA vs TGAGG->TGTGG + CCTCA->CCACA	949	1,733	799	1,481	1.02	8.72E-01	0
TCTGG->TCAGG + CCAGA->CCTGA vs TCAGG->TCTGG + CCTGA->CCAGA	834	1,330	1,052	1,701	1.01	8.93E-01	0
CATGC->CAGGC + GCATG->GCCTG vs CAGGC->CATGC + GCCTG->GCATG	1,694	2,886	895	1,545	1.01	8.74E-01	1
CACAG->CAGAG + CTGTG->CTCTG vs CAGAG->CACAG + CTCTG->CTGTG	3,531	6,486	5,010	9,318	1.01	7.29E-01	0
GGTGA->GGGGA + TCACC->TCCCC vs GGGGA->GGTGA + TCCCC->TCACC	1,068	2,053	1,481	2,879	1.01	8.81E-01	2
AACTG->AAGTG + CAGTT->CACTT vs AAGTG->AACTG + CACTT->CAGTT	2,893	5,238	2,970	5,437	1.01	8.14E-01	0
GTTTC->GTATC + GAAAC->GATAC vs GTATC->GTTTC + GATAC->GAAAC	434	864	557	1,121	1.01	9.51E-01	-2
TACTC->TATTC + GAGTA->GAATA vs TATTC->TACTC + GAATA->GAGTA	5,596	10,396	6,633	12,450	1.01	7.29E-01	1
TATAC->TAAAC + GTATA->GTTTA vs TAAAC->TATAC + GTTTA->GTATA	998	1,828	612	1,132	1.01	9.47E-01	2
TACAG->TAGAG + CTGTA->CTCTA vs TAGAG->TACAG + CTCTA->CTGTA	2,433	4,182	5,215	9,048	1.01	8.34E-01	0
AATGC->AAGGC + GCATT->GCCTT vs AAGGC->AATGC + GCCTT->GCATT	1,575	2,695	1,135	1,958	1.01	9.37E-01	1
CCTTG->CCCTG + CAAGG->CAGGG vs CCCTG->CCTTG + CAGGG->CAAGG	4,224	7,498	6,952	12,440	1.01	8.12E-01	1
CATTC->CAGTC + GAATG->GACTG vs CAGTC->CATTC + GACTG->GAATG	1,153	1,992	1,172	2,041	1.01	9.46E-01	-1
CACAG->CATAG + CTGTG->CTATG vs CATAG->CACAG + CTATG->CTGTG	10,370	18,040	15,319	26,840	1.01	7.30E-01	0
CTTAA->CTAAA + TTAAG->TTTAG vs CTAAA->CTTAA + TTTAG->TTAAG	971	1,493	1,268	1,962	1.01	9.55E-01	1
GATCA->GAGCA + TGATC->TGCTC vs GAGCA->GATCA + TGCTC->TGATC	586	1,001	1,298	2,230	1.01	9.63E-01	0
AATTC->AAATC + GAATT->GATTT vs AAATC->AATTC + GATTT->GAATT	1,388	3,195	1,115	2,581	1.01	9.63E-01	1
GATCC->GAGCC + GGATC->GGCTC vs GAGCC->GATCC + GGCTC->GGATC	594	929	1,511	2,376	1.01	9.84E-01	0
TGCCT->TGGCT + AGGCA->AGCCA vs TGGCT->TGCCT + AGCCA->AGGCA	1,752	3,548	2,217	4,514	1.01	9.43E-01	0
TACTC->TAATC + GAGTA->GATTA vs TAATC->TACTC + GATTA->GAGTA	950	1,609	920	1,566	1.01	9.84E-01	1
AGTGC->AGAGC + GCACT->GCTCT vs AGAGC->AGTGC + GCTCT->GCACT	859	1,653	605	1,170	1.00	9.84E-01	0
AACTA->AATTA + TAGTT->TAATT vs AATTA->AACTA + TAATT->TAGTT	5,193	8,983	7,557	13,122	1.00	9.30E-01	1
CCCTC->CCATC + GAGGG->GATGG vs CCATC->CCCTC + GATGG->GAGGG	1,287	2,158	1,082	1,820	1.00	9.88E-01	-1
TACAT->TATAT + ATGTA->ATATA vs TATAT->TACAT + ATATA->ATGTA	11,912	20,819	19,030	33,348	1.00	9.18E-01	0
CCCAC->CCTAC + GTGGG->GTAGG vs CCTAC->CCCAC + GTAGG->GTGGG	6,749	11,138	2,454	4,059	1.00	9.84E-01	-1
CACAC->CATAC + GTGTG->GTATG vs CATAC->CACAC + GTATG->GTGTG	8,402	14,092	6,108	10,266	1.00	9.63E-01	0
TCCTG->TCTTG + CAGGA->CAAGA vs TCTTG->TCCTG + CAAGA->CAGGA	7,750	13,870	6,133	10,996	1.00	9.77E-01	0
ATCAA->ATGAA + TTGAT->TTCAT vs ATGAA->ATCAA + TTCAT->TTGAT	3,068	5,699	1,863	3,466	1.00	1.00E+00	0
CCTCT->CCACT + AGAGG->AGTGG vs CCACT->CCTCT + AGTGG->AGAGG	1,103	2,125	1,265	2,440	1.00	1.00E+00	0
CTCAA->CTGAA + TTGAG->TTCAG vs CTGAA->CTCAA + TTCAG->TTGAG	3,343	6,049	2,313	4,190	1.00	1.00E+00	0
CATTT->CAGTT + AAATG->AACTG vs CAGTT->CATTT + AACTG->AAATG	2,272	4,054	2,413	4,310	1.00	1.00E+00	-1