UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA
DISSERTAÇÃO DE MESTRADO

SNP LANE, uma base de dados de vizinhança de SNPs e análises do efeito dos nucleotídeos vizinhos na probabilidade de substituição de nucleotídeos em mamíferos

Fernanda Stussi Duarte Lage

Belo Horizonte
Julho 2019

FERNANDA STUSSI DUARTE LAGE

SNP LANE, uma base de dados de vizinhança de SNPs e análises do efeito dos nucleotídeos vizinhos na probabilidade de substituição de nucleotídeos em mamíferos

Dissertação de Mestrado apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais como requisito para obtenção do título de Mestre em Bioinformática.

Orientador: Prof. Dr. José Miguel Ortega
Co-orientador: Prof. Dr. Tetsu Sakamoto

Belo Horizonte
Julho 2019

043 Lage, Fernanda Stussi Duarte.
SNP LANE, uma base de dados de vizinhança de SNPs e análises do efeito dos nucleotídeos vizinhos na probabilidade de substituição de nucleotídeos em mamíferos [manuscrito] / Fernanda Stussi Duarte Lage. – 2019.

210 f. : il. ; 29,5 cm.

Orientador: Prof. Dr. José Miguel Ortega. Co-orientador: Prof. Dr. Tetsu Sakamoto.
Dissertação (mestrado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Mutação. 3. Base de Dados. 4. Polimorfismo de Nucleotídeo Único. I. Ortega, José Miguel. II. Sakamoto, Tetsu. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004

Ficha catalográfica elaborada pela Biblioteca do Instituto de Ciências Biológicas da UFMG

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática da UFMG
Avenida Presidente Antônio Carlos, 6627 – Pampulha
31270-901 - Belo Horizonte – MG
Endereço eletrônico: bioinfo@icb.ufmg.br 55 31 3409-2554

"SNP LANE, uma base de dados de vizinhança de SNPs e análises do efeito dos nucleotídeos vizinhos na probabilidade de substituição de nucleotídeos em mamíferos"

Fernanda Stussi Duarte Lage

Dissertação aprovada pela banca examinadora constituída pelos Professores:

Prof. José Miguel Ortega - Orientador
UFMG

Prof. Tetsu Sakamoto - Coorientador
UFRN

Prof. Gerald Weber
UFMG

Prof. Sandro José de Souza
UFRN

Belo Horizonte, 31 de julho de 2019.

AGRADECIMENTOS

Desde que me propus a entrar na trajetória da Biologia, e trilhar o caminho de um mestrado, por mais Fitas-verdes que eu inventasse no cabelo, durante esse caminho louco e longo eu precisei de amigos, de exemplos e de suporte.

Começo agradecendo ao Prof. Miguel Ortega, que me sinto mais confortável em chamá-lo de Miguelito, por me receber desde o primeiro período da graduação e me ensinar - além de todo o conteúdo - como a carreira científica pode ser divertida. Agradeço a todas as correções, todos os projetos, todas as oportunidades, toda a companhia, copos coloridos e tatuagens solares;

Agradeço ao Prof. Tetsu Sakamoto, que além de co-orientador, se tornou um exemplo de cientista a quem me espelho, obrigada por toda a dedicação que teve em me ensinar - desde programar a como formatar uma tabela da ABNT - e me orientar;

Aos outros também co-autores desse trabalho: Carlos e Lissur, por todas as horas que ambos dedicaram em me ensinar Python e MySQL (respectivamente), bem como todo o tempo gasto desenvolvendo ativamente para que este trabalho pudesse acontecer;

A Prof. Glaura por trazer a estatística a uma linguagem acessível e ajudar nas análises e escolhas dos melhores testes;

Ao departamento de Bioinformática; Sheila, Natália e Tiago que tantas vezes me salvaram, foram sempre solícitos em ajudar a resolver qualquer tipo de situação. com documentações e toda infra-estrutura que poderia precisar;

Em ordem cronológica agora, agradeço aos meus pais Chyntia e Joaquim; juntamente com a minha madrinha - que se me referir a ela com qualquer outro nome que não seja *Dim* corro o risco substancial de sofrer uma agressão física - por todo o suporte para que eu pudesse me dedicar à pesquisa e a Ciência;

Às amigas que carrego desde a infância, Dayane e *Boka,* que me viram crescer, me deram motivos para sorrir quando eu mais precisei e trilharam muito do caminho da vida comigo;

Da Biologia, eu imaginei que conheceria muito de Flores nas aulas de Botânica, mas jamais esperaria que as mais incríveis não estariam em livro algum: Annalice, Joyce, Lorenna, Marcella, Priscila e Rayssa: obrigada por todo o apoio, ombros, carinhos e tapas na cara sempre que necessários;

Ao professor Francisco Lobo que foi quem me apresentou o Laboratório de Biodados e a Bioinformática;

A Fenícia, que começou como uma colega de trabalho e se tornou amiga para carregar da Ciência à Bahia;

Aos membros atuais do Laboratório de Biodados; Beatriz, Renato - por estarem sempre dispostos a ajudar, ensinar, e mostrarem que não devemos nos contentar com menos, e àqueles também que por aqui passaram e deixaram uma parte do que sabiam comigo e sorrisos;

Áqueles que me acompanharam em disciplinas - e fora delas também - passando madrugadas em claro estudando, fazendo códigos e slides: Daniel, Herbert, Heron, Nayara e Stella;

Ao Prof. Raphael Tavares que me trouxe a confiança de programar em Perl;

A Professora Glória por esse exemplo de mulher e cientista que eu quero seguir;

E, aos membros da banca Gerald e Sandro por aceitarem o convite de colaborar com uma melhor versão deste trabalho.

SNP Lane is in my nears and in my why's
There, beneath the true disturbance's guise

(Darren Natale e José Miguel Ortega)

ABREVIATURAS

3'UTR - Região 3' Não traduzida, sequências de mRNA após o códon de terminação

5'UTR - Região 5' Não traduzida, é a região do transcrito de mRNA que está localizada entre o local da CAP e o códon de iniciação

A - Adenina

ApA - Dinucleotídeo ApA, uma Adenina seguida por uma outra Adenina, um nucleotídeo de Adenina ligado a outro por meio de uma ligação fosfodiéster ("p" de *phosphodiester*)

ApT - Dinucleotídeo ApT, uma Adenina seguida por uma Timina, um nucleotídeo de Adenina ligado a um nucleotídeo de Timina por meio de uma ligação fosfodiéster ("p" de *phosphodiester*)

AT - Conteúdo das bases Adenina e Timina no genoma, dado em porcentagem

B - Refere-se aos nucleotídeos "Não A": Citosina, Guanina e Timina

C - Citosina

CAP - *Catabolite activator protein*, proteína que se liga a locais específicos no DNA para promover a ligação e a função da RNA polimerase

CDS - *Coding Sequence*, regiões codificadoras do DNA

CG - Conteúdo das bases Citosina e Guanina no genoma, dado em porcentagem

CpC - Dinucleotídeo CpC, uma Citosina seguida por uma outra Citosina, um nucleotídeo de Citosina ligado a outro por meio de uma ligação fosfodiéster ("p" de *phosphodiester*)

CpG - Dinucleotídeo CpG, uma Citosina seguida por uma Guanina, um nucleotídeo de Citosina ligado a um nucleotídeo de Guanina por meio de uma ligação fosfodiéster ("p" de *phosphodiester*)

D - Refere-se aos nucleotídeos "Não C": Adenina, Guanina e Timina

dbSNP - Base de dados de Polimorfismos de Nucleotídeos Únicos do NCBI que contém variações humanas de nucleotídeo único, microssatélites e inserções e deleções de pequena escala, juntamente com publicação, frequência de população, consequência molecular e informações de mapeamento genômico e para variações comuns e mutações clínicas

DNA - *deoxyribonucleic acid* de ácido desoxirribonucleico

EVA - *European Variation Archive*, banco de dados de acesso aberto de todos os tipos de dados de variação genética de todas as espécies

G - Guanina

GpC - Dinucleotídeo GpC, uma Guanina seguida por uma Citosina, um nucleotídeo de Guanina ligado a um nucleotídeo de Citosina por meio de uma ligação fosfodiéster ("p" de *phosphodiester*)

GpG - Dinucleotídeo GpG, uma Guanina seguida por uma outra Guanina, um nucleotídeo de Guanina ligado a outro por meio de uma ligação fosfodiéster ("p" de *phosphodiester*)

H - Refere-se aos nucleotídeos "Não G": Adenina, Citosina e Timina.

K - Refere-se às bases que apresentam um grupamento cetona (*Ketone*): Guanina e Timina.

M - Refere-se às bases que apresentam um grupo aMino: Adenina e Citosina.

mRNA - RNA (ácido ribonucleico) mensageiro.

N - Refere-se a qualquer nucleotídeo, Adenina, Timina, Citosina ou Guanina.

NCBI - *National Center for Biotechnology Information,* base de dados que fornece acesso a informações biomédicas e genômicas

R - Refere-se às bases púRicas: Adenina e Guanina

RNA - Ácido ribonucleico, *ribonucleic acid*

S - Refere-se às bases que realizam três ligações de hidrogênio entre si (*Strong*): Citosina e Guanina

SNP - *Single Nucleotide Polymorphism*, Polimorfismos de Nucleotídeo Único

SNV - *Single Nucleotide Variation*, Variação de Nucleotídeo Único

T - Timina

TpA - Dinucleotídeo TpA, uma Timina seguida por uma Adenina, um nucleotídeo de Timina ligado a nucleotídeo de Adenina por meio de uma ligação fosfodiéster ("p" de *phosphodiester*)

TpG - Dinucleotídeo TpG, uma Timina seguida por uma Guanina, um nucleotídeo de Timina ligado a um nucleotídeo de Guanina por meio de uma ligação fosfodiéster ("p" de *phosphodiester*)

TpT - Dinucleotídeo TpT, uma Timina seguida por uma outra Timina, um nucleotídeo de Timina ligado a outro por meio de uma ligação fosfodiéster ("p" de *phosphodiester*)

V - Refere-se aos nucleotídeos "Não T": Adenina, Citosina e Guanina

W - Refere-se às bases que realizam duas ligações de hidrogênio entre si (*Weak*): Adenina e Timina

Y - Refere-se às bases pirimídicas (*pYrimidines*): Citosina e Timina

LISTA DE FIGURAS

Figura 1: Código de uma letra para ambiguidade dupla de bases do DNA. 3
Figura 2: Alterações envolvendo Citosina que podem levar ao erro de pareamento. 7
Figura 3: Exemplo de arquivo ASN de *Sus scrofa*.....11
Figura 4: Exemplo de entrada no formato ASN no banco de dados para *Sus scrofa*.....11
Figura 5: Exemplo de entrada no formato FASTA no banco de dados.. 13
Figura 6: Exemplo de tabela com as bases adjacentes e região a qual pertence o SNP.....14
Figura 7: Exemplo de tabela de saída do banco de dados. 15
Figura 8: Diagrama entidade-relacionamento para o banco de dados.. 15
Figura 9: Frequência relativa de Adenina em torno de um SNP simulado W em *Homo sapiens*, com diferentes quantidades de posições sorteadas e repetições do sorteio. 19
Figura 10: Resumo simplificado da metodologia para o estudo dos efeitos de vizinhança em SNPs. 21
Figura 11: Frequência relativa de cada substituição conforme organismo.. 25
Figura 12: Interface web da base de dados SNP Lane. 27
Figura 13: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões 3'UTR em camundongos. 29
Figura 14: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões 5'UTR em camundongos. 30
Figura 15: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões sinônimas em camundongos.. 31
Figura 16: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões intrônicas em camundongos.. 32
Figura 17: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões intrônicas em humanos.....33
Figura 18: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões intrônicas em *Rattus norvegicus*.. 34
Figura 19: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões intrônicas em *Sus scrofa*.. 35
Figura 20: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões intrônicas de bovinos. 36
Figura 21: Representação dos locais vizinhos em transições complementares. 38

Figura 22: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões 3'UTR em camundongos.. 39
Figura 23: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões 5'UTR em camundongos.. 40
Figura 24: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões 5'UTR em bovinos. 41
Figura 25: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões sinônimas em camundongos.. 42
Figura 26: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões intrônicas em camundongos.. 43
Figura 27: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões intrônicas em *Sus scrofa*. 44
Figura 28: Ocorrência de bases nas proximidades do SNP M (A/C) e a frequência basal esperada no genoma, nas regiões sinônimas em *Rattus norvegicus*. 48
Figura 29: Ocorrência de bases nas proximidades do SNP M (A/C) e a frequência basal esperada no genoma, nas regiões intrônicas em humanos. 49
Figura 30: Representação dos locais vizinhos em transversões complementares. 52
Figura 31: Ocorrência de bases nas proximidades do SNP K (G/T) e a frequência basal esperada no genoma, nas regiões intrônicas em humanos. 53
Figura 32: Representação dos locais vizinhos em mutações autocomplementares. 55
Figura 33: Ocorrência de bases nas proximidades do SNP W (A/T) e a frequência basal esperada no genoma, nas regiões intrônicas em camundongos. 56
Figura 34: Ocorrência de bases nas proximidades do SNP W (A/T) e a frequência basal esperada no genoma, nas regiões sinônimas em humanos. 57
Figura 35: Ocorrência de bases nas proximidades do SNP S (C/G) e a frequência basal esperada no genoma, nas regiões intrônicas em camundongos. 60
Figura 36: Frequência relativa de cada dinucleotídeo conforme organismo nas regiões 3' e 5' UTR. 63
Figura 37: Frequência relativa de cada dinucleotídeo conforme organismo nas regiões 3' e 5' UTR. 63
Figura 38: Frequência relativa de cada dinucleotídeo conforme organismo nas regiões intrônicas e codificadoras sinônimas. 64
Figura 39: Frequência relativa de cada dinucleotídeo conforme em diferentes simulações de aminação nas regiões intrônicas de *Mus musculus*. 66

Figura 40: Distância Euclidiana entre as médias das posições 5 e 1 em diferentes regiões em *Mus musculus*.. 67
Figura 41: Ocorrência de bases nas proximidades do SNP W (A/T) e a frequência basal esperada no genoma, nas regiões intrônicas de *Homo sapiens* em duas situações de aminação distintas.. 68
Figura 42: Distância Euclidiana entre cada posição e as médias das posições +5 e -5 em diferentes regiões em *Mus musculus*. 72
Figura 43: Ocorrência de bases nas proximidades do SNP W (A/T) em simulações de desaminação do dinucleotídeo CpG em *Mus musculus* em regiões intrônicas. 75

LISTA DE TABELAS

Tabela 1: Na Tabela 1 temos um resumo da codificação para ambiguidade conforme a *International Union of Pure and Applied Chemistry (IUPAC*, União Internacional de Química Pura e Aplicada). Preferimos utilizar essas letras para abreviar os SNPs, por exemplo W para o SNP A/T ou A←→T. 12

Tabela 2: Cobertura de cada sorteio de posições e repetições conforme o tamanho do menor e do maior genoma analisado. 18

Tabela 3: Média dos desvios-padrões das frequências simuladas esperadas em diferentes sorteios de posições e repetições em um estudo de caso de regiões intrônicas de *Homo sapiens*. 18

Tabela 4: Frequência relativa de cada substituição conforme o organismo.. 24

Tabela 5: Comparativo da frequência de cada substituição em *H. sapiens* em diferentes estudos. 71

Tabela 6: Comparativo da frequência de cada substituição em *B. taurus* em diferentes estudos. 71

RESUMO

A dinâmica da composição do DNA dos genomas de vários táxons é uma das principais questões na análise genômica. Essas alterações podem ser consequências de erros de máquinas de replicação e reparo. A substituição de nucleotídeos, por exemplo, é um dos combustíveis na produção de variação genética e direciona a evolução. Essas substituições devem ocorrer aleatoriamente ao longo do genoma ao longo do tempo. Entretanto, análises recentes de dados genômicos, especialmente dados de SNP (*Single Nucleotide Polymorphism*), demonstraram várias características que sustentam a natureza não aleatória desse evento, como a alta variação na taxa de mutação ao longo do genoma. Estudos recentes mostraram que os componentes do genoma afetam os padrões de mutação até certo ponto, isto é, vieses de vizinhança de nucleotídeos intensos foram observados em posições imediatamente adjacentes a mutações e vieses menos pronunciados se estendem a regiões distantes do local de substituição. Esse fenômeno pode ser atribuído principalmente à enzima que modifica ou modifica o material genético, uma vez que a maioria das enzimas tende a ter contextos específicos de sequências que determinam sua atividade. Assim, a identificação de efeitos de contexto pode levar à descoberta de outros sites de edição ou fatores enzimáticos desconhecidos.

Com o objetivo de investigar de forma abrangente esse evento, construímos um banco de dados on-line para mostrar o padrão de bases na vizinhança do SNP, disponível em: http://bioinfo.icb.ufmg.br/snplane/ utilizando os últimos conjuntos de dados SNP de 5 espécies diferentes de mamíferos (*Mus musculus, Homo sapiens, Bos taurus, Rattus norvegicus e Sus scrofa*) foram baixados e então analisados de acordo com a região genômica onde o SNP pertencia (intrón, exon, 5'UTR, 3'UTR e CDS) e classificados conforme o tipo de substituição: K, M, R, Y, W ou S. Para cada classe de SNP, as frequências de nucleotídeos foram calculadas para as primeiras cinco posições a montante e a jusante em torno do SNP. As frequências esperadas de nucleotídeos basais para as posições vizinhas ao SNP foram estimadas aleatoriamente, escolhendo posições no genoma e recuperando os nucleotídeos que o flanqueiam. Dois gráficos são apresentados para cada uma das 1200 situações distintas.

Na maioria dos casos, a frequência basal não foi significativamente diferente dos dados observados, indicando que o efeito do nucleotídeo vizinho observado não influenciou a mutação, mas se T ou A forem mais frequentes a jusante de C, pareceria que C poderia estar influenciando a mutação. transição T / A, mas a frequência de referência indica que este é apenas um efeito de não-aleatoriedade do genoma. Quando desaminamos todo o C restante no

CpG, houve um pequeno aumento no viés. Simulando diferentes porcentagens de aminação de "CpA" e "TpG" de volta para dinucleotídeos CpG foi notável que o viés é completamente apagado com 25% a 35% de aminação. Nós não vemos o efeito de nucleotídeo vizinho nessas condições. As substituições de R e Y não responderam à aminação, provavelmente porque a aminação já causa R e Y.

Sugere-se que a composição de dinucleotídeos produza o viés de vizinhança reportado anteriormente sobre a probabilidade do SNP. A maior parte deste efeito pode ser explicada pela desaminação de C em CpG e sugerimos que originalmente o genoma humano teria 25% a 35% da presente CpA e TpG na forma de CpG.

ABSTRACT

The dynamics of DNA composition of the genomes of several taxa is one of the main issues in genomic analysis. These changes might be consequences of errors of both replication and repair machinery. Nucleotide substitution, for example, is one of the fuels in the production of genetic variation and directs evolution. These substitutions are supposed to occur randomly throughout the genome through the time. However, recent analysis on genomic data, especially SNP (single nucleotide polymorphism) data, demonstrated several characteristics supporting the non-randomness nature of this event, such as the high variation on the mutation rate along the genome. Recent studies have shown that the components of the genome affect mutation patterns to some extent, that is, intense nucleotide neighborhood biases were observed at positions immediately adjacent to mutations and less pronounced bias extend to regions distant from the replacement site. This phenomenon may be mainly attributed to the enzyme that modifies or mutates the genetic material, since most enzymes tend to have specific sequence contexts that dictate their activity. Thus, identification of context effects may lead to the discovery of additional editing sites or unknown enzymatic factors.

Aiming to investigate comprehensively this event, we built an online database to show the pattern of bases in SNP neighborhood, available at: http://bioinfo.icb.ufmg.br/snplane/ using the latest SNP datasets from 5 different species of mammals (*Mus musculus*, *Homo Sapiens*, *Bos taurus*, *Rattus Norvegicus* and *Sus scrofa*) were downloaded and then parsed according to the genomic region where the SNP belonged to (intron, exon, 5' upstream, 3' downstream and coding sequences) and classified by substitution type: K, M, R, Y, W or S. For each SNP class, nucleotide frequencies were calculated for the first five positions upstream and downstream surrounding the SNP. Expected baseline nucleotide frequencies for positions neighbouring the SNP were estimated by randomly choosing positions in the genome and retrieving nucleotides flanking it. Two graphics are presented for each of 1200 distinct situations.

In the majority of cases, baseline frequency was not significantly different from the observed data, indicating that the observed neighboring effect was not an influence on the mutation, but rather if T or A are more frequent downstream of C, it would seem C might be influencing the transition T/A but baseline frequency indicates that this is just an effect of non-randomness of the genome.

When we deaminated all remaining C in CpG, was a small increase in bias. Simulating different percentages of amination of "CpA" and "TpG" back to CpG dinucleotides was noteworthy that the bias is completely erased with 25% to 35% of amination. We do not see the neighboring nucleotide effect on these conditions. R and Y substitutions did not respond to amination, probably because amination already causes R and Y.

It is suggested that dinucleotide composition produces the previously reported neighborhood bias on SNP probability. Most of this effect might be explained by deamination of C in CpG and we suggest that originally human genome would have 25% to 35% of the present CpA and TpG in the form of CpG.

SUMÁRIO

AGRADECIMENTOS I

ABREVIATURAS IV

LISTA DE FIGURAS VI

LISTA DE TABELAS IX

RESUMO X

ABSTRACT XII

SUMÁRIO XIV

1. INTRODUÇÃO 1

1.1 Mutações e polimorfismos 1

1.1.1 SNPs 2

1.1.1.1 dbSNP 4

1.1.1.2 EVA 5

1.2 Efeitos de Vizinhança 5

1.2.1 Hipermutabilidade dos dinucleotídeos CpG 6

2. OBJETIVOS 9

2.1. Objetivo Geral 9

2.2 Objetivos Específicos 9

3. MATERIAL E MÉTODOS 10

3.1 Coleção de dados dos SNPs 10

3.1.1 Coleção de dados a partir dos arquivos ASN 10

3.1.2 Coleção de dados a partir dos arquivos FASTA 12

3.2 Estruturação do banco de dados em torno de um SNP 14

3.3 Determinação das frequências em torno de um SNP 14

3.4 Cálculo da frequência basal do genoma 16

3.5 Comparação das Frequências Basais do Genoma e na vizinhança do SNP 20

3.6 Determinação da frequência de dinucleotídeos por região 20
3.7 Experimentos de Aminação dos Genomas 20
4. RESULTADOS 23
4.1 CONSTRUÇÃO DA BASE DE VIZINHANÇA DE SNPS: SNP LANE 23
4.1.1 Obtenção e processamento de dados do dbSNP 23
4.2 Transições e transversões nas subdivisões do dbSNP 24
4.2.1 3'UTR 24
4.2.2 5'UTR 25
4.2.3 CDS-Sinônimas 26
4.2.4 Intron 26
4.3 Determinação da frequência de base na vizinhança de cada tipo de substituição 27
4.3.1 Frequência de base na vizinhança de R 28
4.3.1.1 Frequência de base na vizinhança de R em regiões 3'UTR 28
4.3.1.2 Frequência de base na vizinhança de R em regiões 5'UTR 37
4.3.1.3 Frequência de base na vizinhança de R em regiões CDS-Sinônimas 37
4.3.1.4 Frequência de base na vizinhança de R em regiões intrônicas 38
4.3.2 Frequência de base na vizinhança de Y 38
4.3.2.1 Frequência de base na vizinhança de Y em regiões 3'UTR 45
4.3.2.2 Frequência de base na vizinhança de Y em regiões 5'UTR 45
4.3.2.3 Frequência de base na vizinhança de Y em regiões CDS-Sinônimas 46
4.3.2.4 Frequência de base na vizinhança de Y em regiões intrônicas 46
4.3.3 Frequência de base na vizinhança de M 47
4.3.3.1 Frequência de base na vizinhança de M em regiões 3'UTR 47
4.3.3.2 Frequência de base na vizinhança de M em regiões 5'UTR 50
4.3.3.3 Frequência de base na vizinhança de M em regiões CDS-Sinônimas 50
4.3.3.4 Frequência de base na vizinhança de M em regiões intrônicas 51
4.3.4 Frequência de base na vizinhança de K 51

4.3.4.1 Frequência de base na vizinhança de K em regiões 3'UTR 51
4.3.4.2 Frequência de base na vizinhança de K em regiões 5'UTR 52
4.3.4.3 Frequência de base na vizinhança de K em regiões CDS-Sinônimas 54
4.3.4.4 Frequência de base na vizinhança de K em regiões intrônicas 54
4.3.5 Frequência de base na vizinhança de W 55
4.3.5.1 Frequência de base na vizinhança de W em regiões não codificadoras 55
4.3.5.2 Frequência de base na vizinhança de W em regiões codificadoras 58
4.3.6 Frequência de base na vizinhança de S 59
4.3.6.1 Frequência de base na vizinhança de S em regiões 3'UTR 59
4.3.6.2 Frequência de base na vizinhança de S em regiões 5'UTR 59
4.3.6.3 Frequência de base na vizinhança de S em regiões CDS-Sinônimas 61
4.3.6.4 Frequência de base na vizinhança de S em regiões intrônicas 62
4.4 Análise da frequência relativa de cada dinucleotídeo 62
4.5 Análise do efeito da aminação dos dinucleotídeos TpG e CpA 65
5. DISCUSSÃO 70
5.1 Frequência de transições e transversões em diferentes mamíferos 70
5.2 Efeito dos nucleotídeos vizinhos para a ocorrência de uma mutação 73
5.3 Desaminação *in silico* das citosinas em CpG 76
6. CONCLUSÕES 78
7. PERSPECTIVAS 79
8. REFERÊNCIAS BIBLIOGRÁFICAS 80
9. APENDICE A 84

1. INTRODUÇÃO

Sabe-se há muitos anos que diferentes regiões do genoma possuem diferentes comportamentos quanto aos padrões de mutação. No entanto, somente com o advento das tecnologias de sequenciamento e com o acúmulo de grandes conjuntos de dados genômicos é que a extensão total dessa variação se tornou aparente. A taxa de mutação varia em muitas escalas diferentes, desde o nível de nucleotídeo único até a variação entre cromossomos inteiros. Essa variação tem implicações importantes na biologia evolutiva e implicações subexploradas para a nossa compreensão como por exemplo em doenças hereditárias e câncer (HODGKINSON; EYRE-WALKER, 2011). Essas mutações são majoritariamente atribuídas a erros tanto da replicação quanto da maquinaria de reparo. Uma das abordagens para entender a dinâmica composicional do DNA é analisar as possíveis sequências marcadoras de polimorfismos bem como suas regras (ZHAO et al., 2005).

1.1 Mutações e polimorfismos

Indivíduos distintos de uma dada espécie não são geneticamente idênticos. Suas sequências de DNA diferem até certo ponto e essas diferenças que formam a diversidade genética é conhecida como polimorfismo de uma espécie. A diversidade genética contribui para a habilidade de uma espécie em responder a mudanças ambientais e possui implicações na área da saúde (BODMER, 2015; QUINTANA-MURCI; CLARK, 2013), nas estratégias de melhoramento em cultivos e animais de criação (HAKE; ROSS-IBARRA, 2015), no manejo de doenças infecciosas (SOARES; WEISS, 2015) e até mesmo na conservação de espécies ameaçadas (FORCADA; HOFFMAN, 2014; VANDER WAL et al., 2012). Do ponto de vista teórico, a diversidade genética pode ser pensada como reflexo do equilíbrio entre o surgimento e o desaparecimento de variantes genéticas (alelos) - um dos combustíveis que direciona a evolução (ZHAO et al., 2005).

Novos alelos podem surgir a cada geração por mutação espontânea devido a erros de replicação do DNA, recombinação gênica, inserções, deleções, substituições ou por danos no DNA induzidos por mutagênese; que podem ser causados por fatores exógenos, como a radiação.

Ainda que exista uma maquinaria de reparo, nem todos os erros de pareamento são corrigidos. Mais do que isso, muitas vezes ao invés de reparar o erro, o próprio mecanismo de

correção pode inserir erros - como em casos onde ocorre pareamentos imperfeitos. Esses erros de pareamentos podem acontecer em regiões que não são reguladoras e nem codificadoras ou, se acontecem em uma região codificadora podem ainda ser: (1) mutações sinônimas (quando o aminoácido não é alterado); (2) mutações *missense* conservativas (quando o caráter físico-químico do aminoácido mutado permanece o mesmo); (3) mutações *missense* não conservativas (quando a substituição resulta na alteração do aminoácido); e (4) *non-sense* (quando a substituição resulta em um códon de parada). As mutações não estão necessariamente associadas a um prejuízo para o indivíduo, mas podem inclusive resultar no ganho de uma determinada característica que seja importante para a sobrevivência de um organismo (OLIVEIRA, 2011).

A taxa de mutação não é constante ao longo do genoma (HODGKINSON; EYRE-WALKER, 2011) e nem entre as espécies (LYNCH, 2010), o que pode explicar parte da variação observada na diversidade genética (ELLEGREN; GALTIER, 2016). Para muitos genes, há um único alelo predominante, em geral presente em mais da metade dos indivíduos em uma população - o alelo selvagem ou alelo comum. As outras versões do gene são alelos variantes (ou mutantes) que diferem do alelo selvagem, devido à presença de uma mutação, uma alteração permanente na sequência de nucleotídeos ou na disposição do DNA. A frequência de variantes diferentes pode variar amplamente em populações distintas - os estudos de evolução molecular tratam as mutações como documentos históricos que nos permitem entender o processo evolucionário de um dado organismo. A partir da identificação da ocorrência de diferentes versões da sequência de DNA em um mesmo locus em uma população é possível, por exemplo, estimar a taxa evolutiva e reconstruir a cronologia evolutiva (OLIVEIRA, 2011). Se houver dois ou mais alelos relativamente comuns (definidos por convenção como tendo uma frequência alélica >1%) em um locus na população, diz-se que esse locus apresenta polimorfismo nessa população. A maioria dos alelos variantes, no entanto, não é suficientemente frequente na população para ser considerada como polimorfismo (NUSSBAUM; MCINNES; WILLARD, 2016).

1.1.1 SNPs

Um processo básico na evolução de sequências de DNA é a substituição de um nucleotídeo por outro, e, quando esta variante apresenta uma frequência populacional superior a 1% dos alelos na população, ela é formalmente considerada um polimorfismo de um único

nucleotídeo (cuja sigla SNP vem do inglês de *Single Nucleotide Polymorphism*) (NUSSBAUM; MCINNES; WILLARD, 2016).

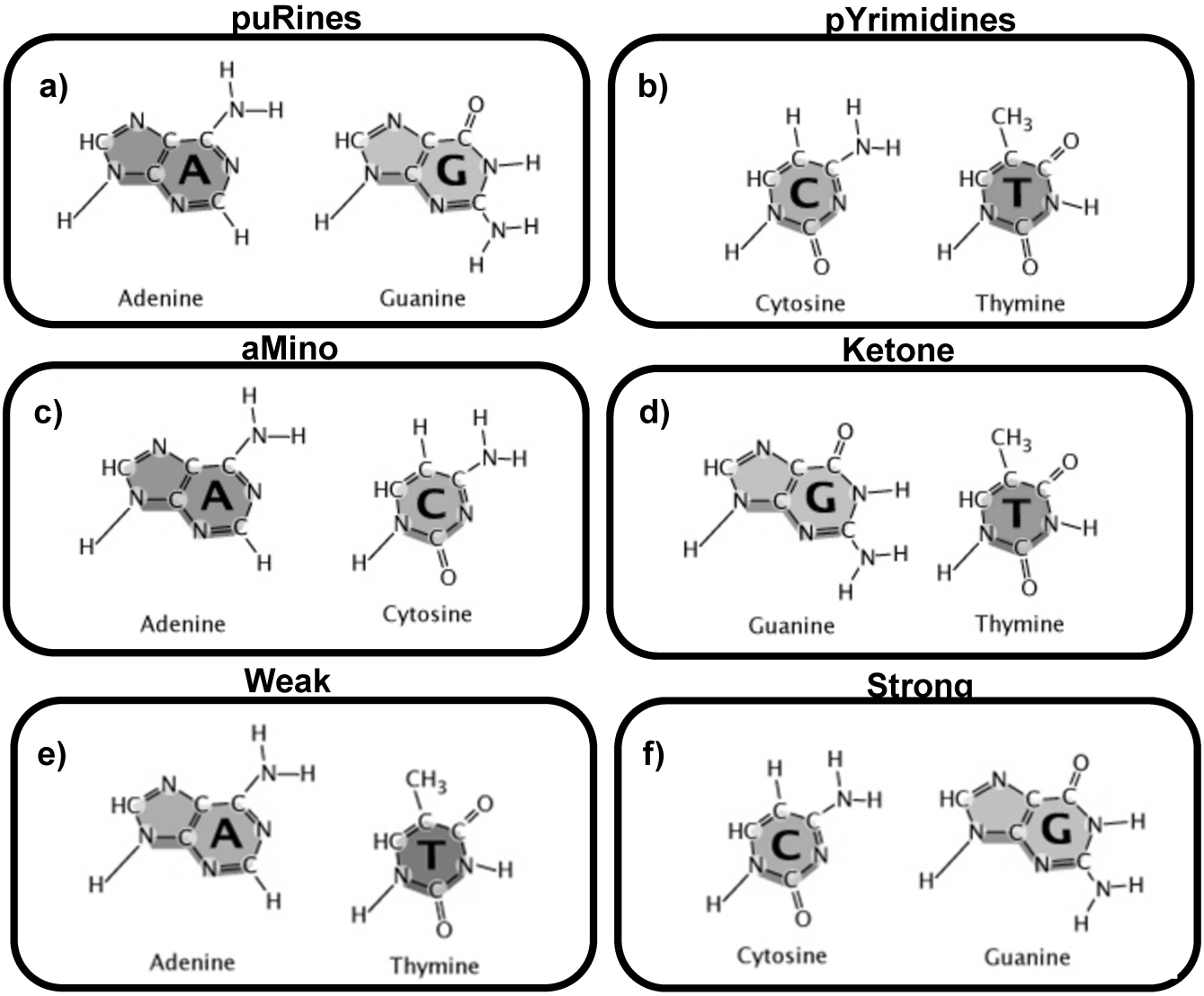


Figura 1: Código de uma letra para ambiguidade dupla de bases do DNA. Alguns sequenciadores de DNA quando em dúvida entre duas bases utilizam essa codificação, utilizada também na especificação de posições degeneradas de iniciadores para PCR. São três maneiras distintas de separar as quatro bases. A primeira é ilustrada na Figura 1a. Duas das bases do DNA possuem dois anéis e duas, um anel. As com dois anéis são as "puRines", assim R designa A ou G. Em contraponto, Y designa as com um anel, as "pYrimidines" C e T Figura 1b. A segunda maneira é ilustrada nas Figuras 1c e 1d. Note que as bases A e C apresentam um grupo químico amino voltado para cima enquanto G e T na mesma posição estão desaminadas, apresentando um grupo que lembra uma cetona. Assim, M de "aMino" designa A ou C conforme ilustrado na Figura 1c, enquanto K de "Ketone" designa "G ou T", ilustrado na Figura 1d. A terceira maneira de separar as quatro bases em grupos de duas é exemplificada nas Figuras 1e e 1f. Em alguns lugares é comum se referir ao pareamento de Watson e Crick com "Agnaldo Timóteo" e "Gal Costa". Como mostrado na Figura 1e A e T formam duas interações de hidrogênio e C e G, três, Figura 1f. Assim, usa-se W de "Weak" para se designar A ou T e por outro lado S de "Strong" para C ou G, em referência à força de interação entre os pares de base.

Fonte: *Adaptado de:* PRAY, L. (2008) Discovery of DNA structure and function: Watson and Crick. **Nature Education** 1(1):100. Disponível em: <https://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397>. Acesso em 19/03/2019

SNPs são o tipo mais simples e comum de variação genética entre genomas de plantas e animais. Cada SNP representa uma diferença em um nucleotídeo. Por exemplo, um SNP pode substituir o nucleotídeo citosina (C) pelo nucleotídeo timina (T) em um certo trecho de DNA (NIH, 2019) - os SNPs que representam trocas entre duas bases podem ser visualizados na Figura 1.

Os SNPs podem ocorrer ao longo de todo o genoma, mas em certas regiões, a ocorrência do SNP pode influenciar no fenótipo do organismo. Caso ele ocorra nos limites intrón-éxon, ele pode inserir/retirar os motivos 'GU – AG' conservados e modificar o peptídeo resultante. Alterações nas regiões promotoras e reguladoras podem causar ganhos ou perdas de sítios de ligação específicos e resultar em regulação diferencial da transcrição. Mesmo SNPs sinônimos, desconsiderados em muitos estudos com base na suposição de que estes são silenciosos, podem alterar a estrutura secundária do RNA (WANG et al., 2005) e afetar a conformação e a função de proteínas (SAUNA et al., 2007). Os SNPs podem ainda ser divididos em dois grandes grupos: transições e transversões. As transições representam as substituições de uma base por outra da mesma categoria química, ou seja, substituições de uma purina por outra purina (A ←→ G, representadas por R conforme visto na Figura 1) ou de uma pirimidina por outra pirimidina (Citosina ←→ Timina, representadas por Y). Já as transversões representam substituições de bases de categoria química diferente. Tanto uma pirimidina pode ser substituída por uma purina (C → A, C → G, T → A ou T → G) quanto uma purina pode ser substituída por uma pirimidina (A → C, A → T, G → C, ou G → T) (OLIVEIRA, 2011).

Além disso, os SNPs são importantes marcadores que ligam os genes a alterações fisiológicas normais, doenças e respostas a patógenos, produtos químicos, drogas, vacinas e outros agentes em humanos (JIANG et al., 2008, 2007; KIM; MISRA, 2007; RILEY et al., 2000).

1.1.1.1 dbSNP

Em uma iniciativa do NCBI (*The National Center for Biotechnology Information*) em colaboração com o *National Human Genome Research Institute* (NHGRI), o dbSNP (*Single Nucleotide Polymorphism Database* (SHERRY, 2001) foi criado em 1998 como um repositório que arquiva, processa, exibe e relata informações relacionadas a variações germinativas e somáticas de várias espécies. Esse banco de dados cresceu rapidamente conforme o sequenciamento evoluiu, e agora contém quase dois bilhões de variantes de mais de 360 espécies.

Dentro do universo de bancos de dados de domínio público, o dbSNP é de particular interesse no contexto deste trabalho por conta de em seu penúltimo *release* - *release* 151 - ainda ser o repositório central para SNPs disponível publicamente. Ainda que, com base no crescimento projetado e nos recursos necessários para arquivar e distribuir os dados, o suporte contínuo a todos os organismos se tornou insustentável para o NCBI e, dessa forma, no *release* mais recente disponível, o *release* 152 - o NCBI eliminou o suporte para todos os organismos não humanos, passando a suportar apenas variações humanas (NCBI INSIGHTS, 2017).

1.1.1.2 EVA

O *European Variation Archive* (EVA) é um repositório aberto para arquivar, acessar e disponibilizar variações genômicas em qualquer espécie. Desde o seu lançamento em 2014, o EVA arquivou aproximadamente 700 milhões de variantes únicas em 351 estudos e 46 espécies.

Uma função chave do EVA é fornecer identificadores estáveis e de longo prazo para que estudos e variantes possam ser referenciados em publicações, interligados entre bancos de dados e integrados com sucessivas compilações de genoma de referência. Desde 1º de setembro de 2017 - data que o dbSNP foi descontinuado para espécies não-humanas - é de responsabilidade do EVA de emitir e manter identificadores de referência para variantes (os indicadores rs - de *Reference SNP*) para essas espécies e importá-las progressivamente do dbSNP. (NCBI INSIGHTS, 2017; EUROPEAN MOLECULAR BIOLOGY LABORATORY, [s.d.]).

1.2 Efeitos de Vizinhança

Estudos recentes mostraram que alguns componentes ou características do genoma podem afetar nas probabilidades de um tipo de mutação ocorrer. Uma das formas de observar estas alterações nas probabilidades de uma mutação ocorrer é verificando a composição dos nucleotídeos vizinhos ao sítio onde a mutação aconteceu. Muitos trabalhos encontraram vieses intensos na frequência dos nucleotídeos em posições imediatamente adjacentes a mutações e vieses menos pronunciados nas regiões distantes do sítio de substituição. Sabe-se ainda que, a substituição de nucleotídeos não ocorre aleatoriamente nas regiões não-codificadoras e codificadoras - sua taxa depende do contexto da sequência e as probabilidades de uma mutação ocorrer são específicas para cada tipo de substituição (GOJOBORI; LI; GRAUR, 1982; MAJEWSKI, 2003; SIEPEL, 2003; ZHAO et al., 2005).

Um exemplo disso é o conteúdo AT nos dois sítios imediatamente adjacentes aos locais de mutação está associado a um aumento da taxa de transversão no genomas nucleares de humanos, camundongos e *Arabidopsis* (ZHANG; ZHAO, 2004). Esses dois sítios de nucleotídeos adjacentes que flanqueiam as mutações também mostram os maiores vieses em comparação com a média específica do genoma e do cromossomo (JIANG et al., 2007; ZHANG; ZHAO, 2004; ZHAO, 2002).

Um exame detalhado da não-aleatoriedade de onde o polimorfismo genético ocorre é importante para a compreensão do mecanismo de mutação (por exemplo, *hotspots*), interação proteína-DNA e evolução da sequência do genoma. Esse fenômeno pode ser atribuído principalmente às enzimas que modificam o material genético, uma vez que a maioria das enzimas tende a ter contextos específicos de sequências que determinam sua atividade. Outro fator que influencia a ocorrência de polimorfismos é a energia livre associada à flexibilidade da microrregião do DNA que ele ocorre, ou seja, existe uma razão física para a predominância de certas bases vizinhas sobre outras na frequência de SNPs (OLIVEIRA, 2011). Assim, a identificação de efeitos de contexto pode levar à descoberta de outros sítios de edição ou fatores enzimáticos desconhecidos (LING et al., 2019; ZHAO et al., 2005).

1.2.1 Hipermutabilidade dos dinucleotídeos CpG

Um dos padrões comumente relatados para SNPs é a ocorrência de uma maior frequência de transição em regiões que apresenta sequência de dinucleotídeo CpG (PLYLER et al., 2015). A notação CpG vem em contraste com a notação CG, que se refere ao conteúdo das bases Citosina e Guanina; quando nos referimos a CpG estamos nos referindo há uma sequência em que a Citosina está sucedida por uma Guanina, onde a letra "p" dessa notação provém de *phosphodiester bond,* que não se trata de um pareamento entre Citosina e Guanina, mas refere-se a uma sequência em que a citosina está ligada a Guanina por meio de uma ligação fosfodiéster - esse mesmo raciocínio pode ser aplicado a descrição dos demais nucleotídeos neste estudo, como TpG e CpA. Os efeitos de hipermutabilidade em CpG resultam da desaminação de citosinas metiladas (JIANG et al., 2007; ZHAO et al., 2005).

Uma base pode ser alterada por processos que danificam o DNA (como, por exemplo, desaminação - em *H. sapiens*, como exemplo, pode acontecer pela metilação de resíduos CpG pela enzima DNA (citosina-5)-metiltransferase 1 (Ensembl ID: ENSG00000130816) - que se associa em sítios de replicação de DNA na fase S, mantendo o padrão de metilação na cadeia recém-sintetizada, que é essencial para a herança epigenética (UNIPROT CONSORTIUM,

[s.d.]) . Em mamíferos, a maioria dos pares de CpG são quimicamente modificados pela ligação covalente de um grupo metil à posição C5 do anel de citosina. Este resíduo modificado é distribuído em todo o genoma, incluindo corpos genéticos, repetições endógenas e elementos transponíveis, e sua presença próximos a genes pode reprimir a transcrição. Metilcitosina desamina espontaneamente a timina, resultando na sub-representação de CpG (21% do esperado no genoma humano) (ILLINGWORTH; BIRD, 2009; INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM, 2001).

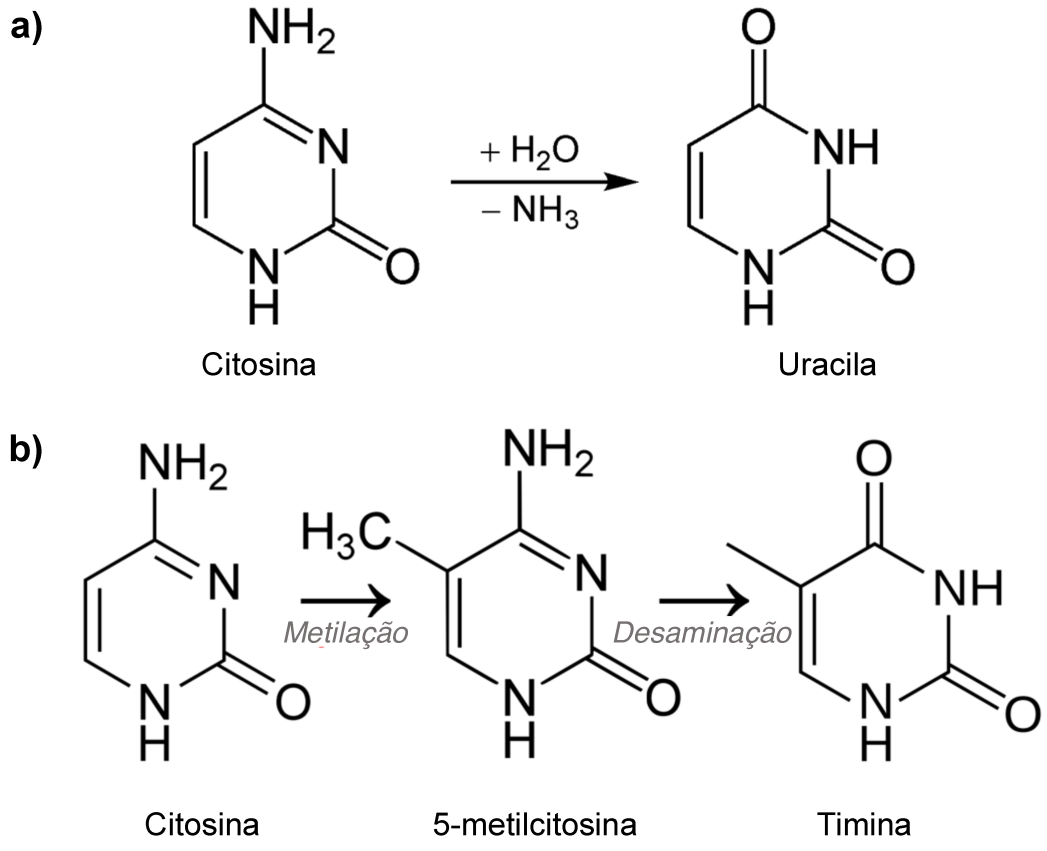


Figura 2: Alterações envolvendo Citosina que podem levar ao erro de pareamento.

Fonte: *Adaptado de:*

<https://en.wikipedia.org/wiki/Deamination#/media/File:DesaminierungCtoU.png>, <https://en.wikipedia.org/wiki/CpG_site#/media/File:Cytosine_becomes_thymine.png> Acesso em 21/03/2019.

Se a base danificada for corretamente reparada por sistemas de reparo de DNA, ou se a base incompatível for resolvida corretamente pelo reparo de incompatibilidade, nenhuma mutação será detectada. Como podemos notar na Figura 2a que se o grupo amino da base C for

desaminado, resultando em cetona, ela se transformará em Uracila (U), se não for corrigido, o futuro pareamento de bases no local afetado resultará em uma mudança de GC para AT quando os pares de bases de Uracila fizerem ligações de hidrogênio com Adenina em vez da Guanina original.

No entanto, o aparecimento no DNA de timina não metilada (isto é, Uracila) é reconhecido como estranho pela uracil-DNA glicosidase e, há um mecanismo de reparo por excisão de base que retira a base U e recoloca a base C. Todavia se a base C estiver metilada na posição 5, a desaminação gerará exatamente a base T, que não é detectada como um erro por essa maquinaria de reparo (PELLEY, 2012).

É importante notar que as taxas de mutações para *forward* e *reverse* geralmente não são iguais; em mamíferos, o C:G →T:A é uma transição de pares de bases que é usualmente mais frequente, e sua frequência é significativamente maior do que a da transição T:A → C: G reversa. Assim, o acúmulo de substituições ainda que neutras, resulta em uma sub-representação significativa de CpG em genomas de mamíferos (MAJEWSKI, 2003).

2. OBJETIVOS

2.1. Objetivo Geral

Este projeto teve como objetivo verificar a existência de vieses na frequência dos nucleotídeos nas posições vizinhas a mutações do tipo SNP. Para isso foram utilizados dados de genomas e de variantes de mamíferos considerados modelo.

2.2 Objetivos Específicos

- Observar frequências de mutações de cada substituição e observar quanto essas frequências desviam do modelo de Kimura (3:1:1);

- Obter dados de genoma e de variantes nos bancos de dados do NCBI (Genome e dbSNP);

- Desenvolver uma metodologia que facilite a análise de dados do dbSNP em larga escala - integrando os dados de diferentes formatos de arquivos;

- Identificar as frequências de cada nucleotídeo nas regiões que flanqueiam cada transição (C/T e A/G) e cada transversão (A/C, A/T, C/G e G/T);

- Estimar a frequência esperada de cada nucleotídeo flanqueando um par de bases específicos, a *baseline frequency*;

- Comparar estatisticamente as frequências relativas de cada nucleotídeo em torno de um SNP para com o que seria esperado no genoma - *baseline frequency*;

- Disponibilizar em uma plataforma web uma visualização gráfica e interativa dos dados resultantes deste projeto;

- Analisar a extensão da influência dos dinucleotídeos CpG sobre os efeitos de vizinhança em torno de um SNP e em diferentes regiões genômicas.

3. MATERIAL E MÉTODOS

3.1 Coleção de dados dos SNPs

Para o cálculo da frequência de cada nucleotídeo em torno de um SNP, utilizamos como base dois formatos de arquivo: ASN e FASTA (PEARSON; LIPMAN, 1988). Esses arquivos são provenientes da base de dados dbSNP (KITTS et al., [s.d.]) do NCBI - que era o banco de dados público referência para submissão de dados de polimorfismos (OLIVEIRA, 2011) até 01 de Setembro de 2017 - *release* 151, data em que o dbSNP fora descontinuado para qualquer outra espécie além de *Homo sapiens* (NCBI INSIGHTS, 2017) e, cuja migração dos demais organismos acontece de forma progressiva para a base de dados dbVAR (EUROPEAN MOLECULAR BIOLOGY LABORATORY, [s.d.]). Para *Mus musculus, Rattus Norvegicus, Sus scrofa* e *Bos taurus*, todos os dados de SNPs disponíveis foram chamados utilizando o mesmo genoma de referência de cada organismo; para *Homo sapiens* utilizamos dados de SNPs que foram chamados utilizando o "*Genome Reference Consortium Human Build 38 patch release 7*", a versão mais atualizada do genoma humano disponível.

A escolha dos organismos a ser analisado neste trabalho foi feita em função da quantidade de SNPs e das informações disponíveis no *release* 150 do dbSNP (Tabela 4). Destacamos que, estudos prévios em espécies de camundongos, humanos, bovinos e outras tiraram conclusões a respeito da frequência de nucleotídeos na vizinha de um SNP usando menores quantidades de dados (BILLINGS et al., 2010; JIANG et al., 2008; MILLER et al., 2012; PLYLER et al., 2015; XUE et al., 2009; ZHANG, 2005; ZHAO; ZHANG, 2006).

3.1.1 Coleção de dados a partir dos arquivos ASN

Os arquivos em formato ASN (Figura 3) consistem de um identificador de referência que começa com as iniciais "rs" - de *reference* SNP - no início da linha, seguido do nome do organismo, número taxonômico separados por "|" ("*pipes*"). Nas linhas seguintes, iniciadas por "ss" - de *submitted* SNP - são as diferentes submissões para aquele SNP. Por sua vez, da linha iniciada por "SNP" extraímos a informação de qual substituição se trata. Na linha que se inicia por "CTG" - de *Contig* - em qual fita o SNP fora declarado e, nas linhas iniciadas por LOC - de locus gênico - a informação de "*fxn-class*" para a região genômica daquele SNP. Nota-se que para uma mesma referência, as informações podem ser distintas por se apresentarem em mais

de um mRNA, tendo mais de uma entrada para informações como orientação de fita e região genômica.

```
rs45430677 | pig | 9823 | snp | genotype=NO | submitterlink=YES | updated 2017-06-23 11:46
ss23131619 | MARC | MARC 23853-23856:1025289152:1:367 | orient=+ | ss_pick=NO
ss23131621 | MARC | MARC 23854-23856:1025288324:1:185 | orient=- | ss_pick=YES
ss48397704 | MARC | MARC 33810-33811:1053549243:1:80 | orient=+ | ss_pick=NO
ss722578209 | WU_ABGC | WU_10.2_10_47447325 | orient=- | ss_pick=NO
ss1487221011 | CRAG | CRAGCAT_rs45430677 | orient=- | ss_pick=NO
ss1719537751 | CAUMQG | 13pigSNP3421335 | orient=- | ss_pick=NO
ss1766568962 | NIAS_AGBSGL | NIAS_AGBSGL-2015.1:SNV:10:g.47447325G>A | orient=- | ss_pick=NO
ss1777770569 | IGA-TECHNOLOGY | SNV:chr10:47447325 | orient=- | ss_pick=NO
ss2491583228 | KEEL-USMARC | NC_010452.3:g47447325G>A | orient=- | ss_pick=NO
ss2594635066 | FUNCTIONALGENOMICS | SNV3967424 | orient=- | ss_pick=NO
SNP | alleles='C/T' | het=? | se(het)=?
VAL | validated=NO | min_prob=? | max_prob=? | notwithdrawn
CTG | assembly=Sscrofa11.1 | chr=10 | chr-pos=42851286 | NW_018084690.1 | ctg-start=17824722 | ctg-end=17824722 | loctype=2 |
 orient=-
LOC | EPC1 | locus_id=100521573 | fxn-class=intron-variant | mrna_acc=XM_003130718.6
LOC | EPC1 | locus_id=100521573 | fxn-class=intron-variant | mrna_acc=XM_005668099.3
LOC | EPC1 | locus_id=100521573 | fxn-class=intron-variant | mrna_acc=XM_005668100.3
LOC | EPC1 | locus_id=100521573 | fxn-class=intron-variant | mrna_acc=XM_013980224.2
LOC | EPC1 | locus_id=100521573 | fxn-class=intron-variant | mrna_acc=XM_021064807.1
```

Figura 3: Exemplo de arquivo ASN de *Sus scrofa*.

Essas informações foram parseadas com o script *projeto_asn_v6_regioesUteis.pl* - que juntamente com os demais scripts utilizados neste estudo podem ser acessados em: https://github.com/fernandastussi/codes4neighbouringnucleotides - juntamente com comandos em linguagem SHELL, realizamos a filtragem de cada identificador conforme a sua região. Para essa análise, utilizamos os SNPs localizados nas regiões 3'UTR, 5'UTR, CDS-sinônimas e intrônicas. Ressalta-se ainda que dados de *indels*, que também estão presentes no dbSNP, não foram considerados neste trabalho. A saída parseada foi então depositada em uma base de dados local MySQL conforme pode ser visualizada na Figura 4.

```
MariaDB [snps]> select * from pig_observed_asn limit 15;
+-------------+-------+------------+------+------+--------+----------+-------------+--------------+
| id          | snp   | synonymous | utr3 | utr5 | intron | missense | orient_plus | orient_minus |
+-------------+-------+------------+------+------+--------+----------+-------------+--------------+
| rs10719553  | 'C/T' | 1.0        | 2.0  | 3.0  | 4.1    | 5.0      | 1           | 0            |
| rs10719554  | 'A/G' | 1.0        | 2.0  | 3.0  | 4.1    | 5.0      | 1           | 0            |
| rs10719555  | 'A/G' | 1.0        | 2.0  | 3.0  | 4.1    | 5.0      | 1           | 0            |
| rs10719556  | 'C/T' | 1.0        | 2.0  | 3.0  | 4.1    | 5.0      | 0           | 1            |
| rs10719557  | 'A/G' | 1.0        | 2.0  | 3.0  | 4.1    | 5.0      | 0           | 1            |
| rs10719558  | 'A/G' | 1.0        | 2.0  | 3.0  | 4.1    | 5.0      | 0           | 1            |
| rs10719559  | 'A/G' | 1.1        | 2.0  | 3.0  | 4.0    | 5.0      | 0           | 1            |
| rs10719560  | 'A/G' | 1.0        | 2.0  | 3.0  | 4.1    | 5.0      | 1           | 0            |
| rs10719561  | 'A/G' | 1.0        | 2.0  | 3.0  | 4.1    | 5.0      | 1           | 0            |
| rs10719562  | 'C/T' | 1.0        | 2.0  | 3.0  | 4.1    | 5.0      | 0           | 1            |
| rs10719563  | 'A/T' | 1.0        | 2.0  | 3.0  | 4.1    | 5.0      | 1           | 0            |
| rs10719564  | 'G/T' | 1.0        | 2.1  | 3.0  | 4.0    | 5.0      | 1           | 0            |
| rs10719565  | 'A/G' | 1.0        | 2.1  | 3.0  | 4.0    | 5.0      | 1           | 0            |
| rs10719566  | 'A/C' | 1.0        | 2.1  | 3.0  | 4.0    | 5.0      | 1           | 0            |
| rs10719567  | 'C/T' | 1.0        | 2.1  | 3.0  | 4.0    | 5.0      | 1           | 0            |
+-------------+-------+------------+------+------+--------+----------+-------------+--------------+
15 rows in set (0.07 sec)
```

Figura 4: Exemplo de entrada no formato ASN no banco de dados para *Sus scrofa*. Os dados obtidos do download do dbSNP são parseados de forma que temos um indicador numérico para cada região (1 para regiões codificadoras sinônimas, 2 para regiões 3'UTR, 3 para regiões de 5'UTR , 4 para regiões intrônicas e 5 para regiões

codificadoras em que há perda de sentido ("*missense*"), usamos um ponto como separador e em seguida temos um indicador de ausência ou presença para essa região (respectivamente 0 para presença e um para ausência – mesmo conceito aplicado a orientação das fitas em que a mutação foi declarada, sendo "*orient_plus*" a fita contínua e "*orient_minus*" a fita descontínua). No formato ASN removemos as entradas de SNPs *indels* do dbSNP.

3.1.2 Coleção de dados a partir dos arquivos FASTA

Para obtermos a sequência em torno de um SNP utilizamos o formato de representação conhecido como FASTA - que é composto por uma linha descritiva que começa com o sinal de maior (>) seguida por linhas consecutivas que representam os nucleotídeos da sequência. Quando se trata de um arquivo FASTA que descreve um SNP, na posição da sequência onde está um SNP temos um caractere que representa um código de ambiguidade estabelecido pela *International Union of Pure and Applied Chemistry* (IUPAC) (Oliveira, 2011) que pode ser visualizado na Tabela 1. Uma das características dessa anotação é ressaltar o caráter bidirecional da mutação (por exemplo, C → A ou A → C que, são ambos representados por M) uma vez que uma limitação deste estudo inclui a incapacidade de determinar a direção da mutação (ZHAO, 2002).

Tabela 1: Na Tabela 1 temos um resumo da codificação para ambiguidade conforme a *International Union of Pure and Applied Chemistry (IUPAC*, União Internacional de Química Pura e Aplicada). Preferimos utilizar essas letras para abreviar os SNPs, por exemplo W para o SNP A/T ou A←→T.

Fonte: *Adaptado de:* <https://www.bioinformatics.org/sms/iupac.html>. Acesso em 21/03/2019

Representação:	Base:
A	Adenina
B	C, G ou T
C	Citosina
D	A, G ou T
G	Guanina
H	A, C ou T
K	G ou T
M	A ou C
N	A, C, G ou T
R	A ou G
S	G ou C
T	Timina
U	Uracila
V	A, C ou G
W	A ou T
Y	C ou T

Para a nossa análise, classificamos cada sequência de SNP - conforme seu indicador "rs" - para cada categoria de substituição R, Y, M, K, W e S, utilizando o script *banco_fasta.pl*, e os seis nucleotídeos a jusante e a montante aos SNPs foram rotulados com números de -1 a -6 e de 1 a 6, respectivamente. Um exemplo de como as sequências foram parseadas e depositadas no banco de dados pode ser visualizado na Figura 5. Com o objetivo de otimizar o tempo das consultas, as tabelas foram fracionadas de acordo com as posições dos SNPs, de tal forma que foram geradas uma tabela para as posições de -3 a +3 e uma outra tabela para as posições de -6 a -4 e +4 à +6. Estudos prévios já haviam demonstrado que as proporções nas duas posições mais próximas em cada lado mostraram um grande viés em relação à média do genoma humano (ZHAO et al., 2005; ZHAO, 2002), então, em nossa análise investigamos a extensão desse viés até as cinco posições mais próximas em cada lado do SNP.

Através de comandos em Shell no próprio terminal Linux, selecionamos apenas os SNPs que possuem dois nucleotídeos diferentes no sítio polimórfico, excluindo as substituições B, D, H, V e N. O número de transições foi pontuado a partir das substituições R e Y, enquanto o número de transversões foi pontuado a partir das substituições M, K, W e S.

```
MariaDB [snps]> select * from  pig_observed_fasta_123 limit 15;
+-------------+-------+------+------+------+------+------+------+------+
| id          | taxid | m3   | m2   | m1   | snp  | p1   | p2   | p3   |
+-------------+-------+------+------+------+------+------+------+------+
| rs10719553 |  9823 |  T   |  C   |  A   |  Y   |  G   |  G   |  C   |
| rs10719554 |  9823 |  A   |  C   |  C   |  R   |  A   |  G   |  C   |
| rs10719555 |  9823 |  G   |  T   |  G   |  R   |  A   |  T   |  C   |
| rs10719556 |  9823 |  G   |  G   |  G   |  Y   |  A   |  G   |  G   |
| rs10719557 |  9823 |  G   |  G   |  G   |  R   |  T   |  G   |  G   |
| rs10719558 |  9823 |  G   |  G   |  G   |  R   |  T   |  G   |  G   |
| rs10719559 |  9823 |  G   |  G   |  A   |  R   |  T   |  T   |  C   |
| rs10719560 |  9823 |  C   |  C   |  T   |  R   |  A   |  G   |  C   |
| rs10719561 |  9823 |  W   |  C   |  A   |  R   |  A   |  G   |  T   |
| rs10719562 |  9823 |  C   |  T   |  T   |  Y   |  T   |  G   |  G   |
| rs10719563 |  9823 |  A   |  C   |  T   |  W   |  C   |  A   |  R   |
| rs10719564 |  9823 |  G   |  C   |  C   |  K   |  C   |  C   |  T   |
| rs10719565 |  9823 |  G   |  A   |  C   |  R   |  C   |  A   |  A   |
| rs10719566 |  9823 |  T   |  C   |  A   |  M   |  G   |  G   |  C   |
| rs10719567 |  9823 |  A   |  T   |  T   |  Y   |  C   |  T   |  A   |
+-------------+-------+------+------+------+------+------+------+------+
15 rows in set (0.02 sec)
```

Figura 5: Exemplo de entrada no formato FASTA no banco de dados. Os dados obtidos do download do dbSNP são parseados de forma que armazenamos as cinco posições mais próximas do SNP - na Figura 5 observamos apenas as 3 posições mais adjacentes á mutação, as posições 4 e 5 se encontram em uma tabela distinta. Tal separação ocorreu em função de uma otimização no tempo de consulta das tabelas. No formato fasta desconsideramos as ambiguidades de mais de dois nucleotídeos, ou seja: B, D, H, N e V.

3.2 Estruturação do banco de dados em torno de um SNP

Para possuirmos um banco de dados em que tivéssemos a informação da região genômica de um SNP juntamente com a informação da sequência na vizinhança dessa mutação, se fez necessário concatenar a informação dos arquivos FASTA e ASN. Como os indicadores "rs" são únicos nas duas tabelas, estes foram usados como chave primária para conciliarmos essas informações em uma nova tabela (Figura 6).

Através de um comando "*select count(*)*" do próprio MySQL geramos a saída final do banco que consiste em uma tabela com três colunas para cada posição e para cada região. Nesta tabela, as colunas correspondem (1) ao tipo de SNP, (2) à base encontrada em uma das posição 10 posições (-5 a -1 ou 1 a 5) ao redor do SNP, e (3) ao número de vezes que aquela base (na segunda coluna) apareceu ao redor daquela posição visto na Figura 7. Um diagrama de entidade-relacionamento de todas as relações do banco de dados, pode ser observado na Figura 8. O script *metodologia.sql* contém todos os comandos em sua respectiva ordem que foram usados para construção e montagem do banco.

```
MariaDB [snps]> select * from pig_new_observed_intron_p123 limit 15;
+------------+-------+------+------+------+------+------+------+------+-------------+--------------+
| id         | taxid | m3   | m2   | m1   | snp  | p1   | p2   | p3   | orient_plus | orient_minus |
+------------+-------+------+------+------+------+------+------+------+-------------+--------------+
| rs10719553 |  9823 |  T   |  C   |  A   |  Y   |  G   |  G   |  C   | 1           | 0            |
| rs10719554 |  9823 |  A   |  C   |  C   |  R   |  A   |  G   |  C   | 1           | 0            |
| rs10719555 |  9823 |  G   |  T   |  G   |  R   |  A   |  T   |  C   | 1           | 0            |
| rs10719556 |  9823 |  G   |  G   |  G   |  Y   |  A   |  G   |  G   | 0           | 1            |
| rs10719557 |  9823 |  G   |  G   |  G   |  R   |  T   |  G   |  G   | 0           | 1            |
| rs10719558 |  9823 |  G   |  G   |  G   |  R   |  T   |  G   |  G   | 0           | 1            |
| rs10719560 |  9823 |  C   |  C   |  T   |  R   |  A   |  G   |  C   | 1           | 0            |
| rs10719561 |  9823 |  W   |  C   |  A   |  R   |  A   |  G   |  T   | 1           | 0            |
| rs10719562 |  9823 |  C   |  T   |  T   |  Y   |  T   |  G   |  G   | 0           | 1            |
| rs10719563 |  9823 |  A   |  C   |  T   |  W   |  C   |  A   |  R   | 1           | 0            |
| rs10719571 |  9823 |  G   |  A   |  A   |  Y   |  G   |  G   |  C   | 0           | 1            |
| rs10719572 |  9823 |  C   |  A   |  T   |  R   |  G   |  G   |  G   | 1           | 0            |
| rs10719573 |  9823 |  G   |  G   |  C   |  Y   |  G   |  G   |  T   | 1           | 0            |
| rs10719574 |  9823 |  C   |  A   |  C   |  S   |  T   |  C   |  T   | 1           | 0            |
| rs10719575 |  9823 |  C   |  C   |  C   |  R   |  C   |  C   |  C   | 1           | 0            |
+------------+-------+------+------+------+------+------+------+------+-------------+--------------+
15 rows in set (0.04 sec)
```

Figura 6: Exemplo de tabela com as bases adjacentes e região a qual pertence o SNP. A informação da região do SNP proveniente do formato ASN é concatenada com a informação de sequência do formato fasta através do indicador – coluna "id".

3.3 Determinação das frequências em torno de um SNP

Uma vez que temos os dados pré-processados no banco de dados, utilizamos o script *formatacao_de_saidas_singlets_usage_p5a1_m1a5_v3.py* para: unificar os arquivos de saída do banco de dados de cada posição em um único arquivo - separados conforme organismo e

```
MariaDB [snps]> select * from pig_new_observed_intron_p1_singlet limit 15;
+------+------+----------+
| snp  | p1   | count(*) |
+------+------+----------+
|  K   |  A   |   681361 |
|  K   |  C   |   659499 |
|  K   |  G   |   770211 |
|  K   |  K   |      142 |
|  K   |  M   |       15 |
|  K   |  R   |       56 |
|  K   |  S   |       24 |
|  K   |  T   |   820797 |
|  K   |  W   |       64 |
|  K   |  Y   |       87 |
|  M   |  A   |  1066893 |
|  M   |  C   |   658396 |
|  M   |  G   |   435111 |
|  M   |  K   |       17 |
|  M   |  M   |      171 |
+------+------+----------+
15 rows in set (0.00 sec)
```

Figura 7: Exemplo de tabela de saída do banco de dados. Na figura observamos um exemplo de saída para a posição "+1" de *Sus scrofa* nas regiões intrônicas. Esse modelo de tabela é extraído para cada posição de uma respectiva região de cada organismo.

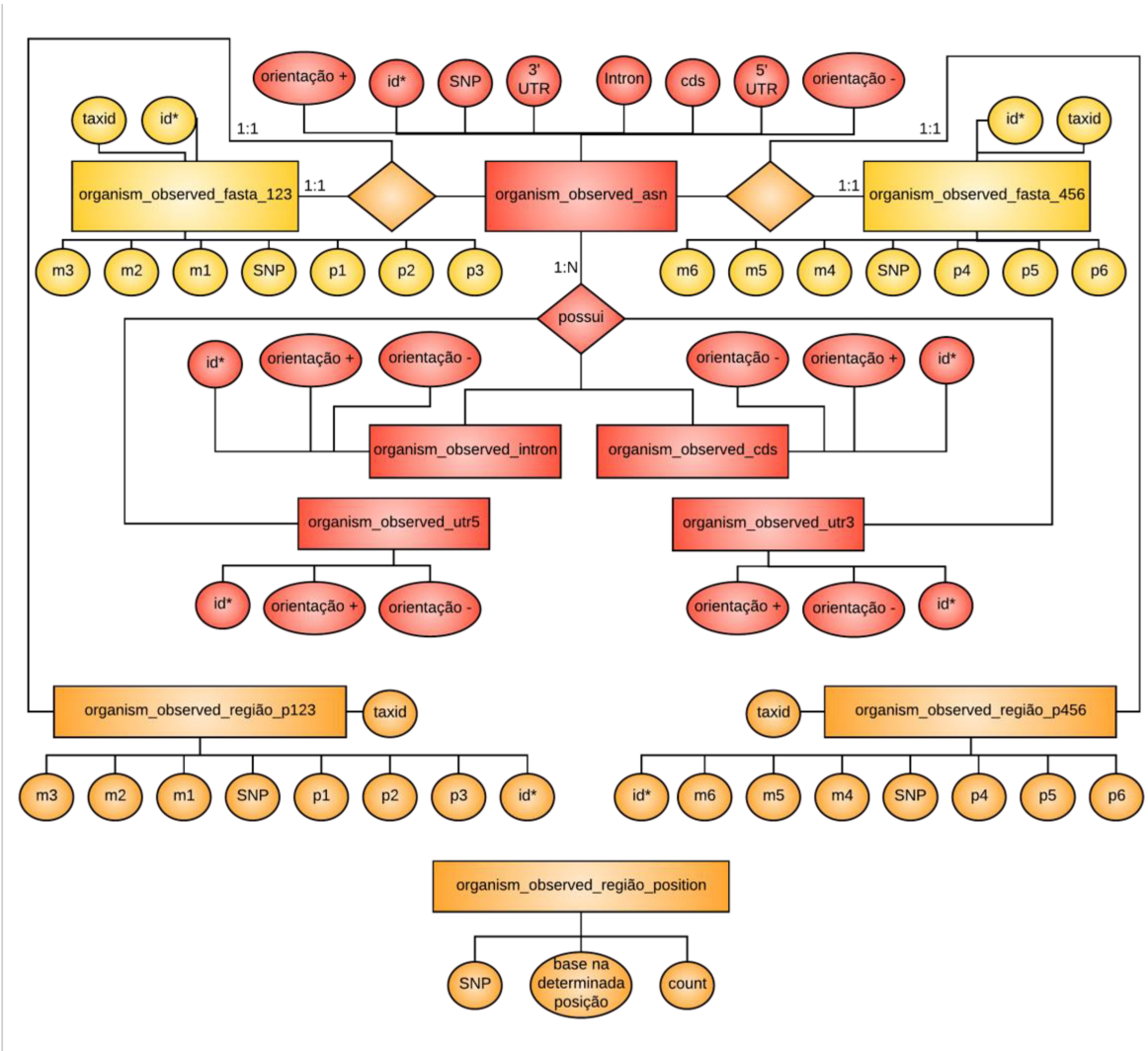


Figura 8: Diagrama entidade-relacionamento para o banco de dados. A construção do banco tem como primeira entrada as seguintes tabelas (as entidades do banco, representadas por retângulos) para cada organismo: asn, fasta_123 e fasta_456. Todas elas possuem o id como chave primária (o atributo-chave para cada tabela se encontra marcado com um asterisco), concatenamos a informação da região presente no ASN com a informação

da sequência presente nas tabelas FASTA (cada relacionamento é representado por um losango), com isso, geramos uma nova tabela com todas essas informações indexadas pelo id. Com o comando "*select count(*),position,snp from organism_observed_region_456 group by position,snp order by snp,position;* " geramos a saída final do banco que consiste em uma tabela para cada posição, para cada região com uma coluna para o SNP, a base que flanqueia esse SNP e outra coluna que conta o número de vezes que aquela base apareceu ao redor daquela posição.

região genômica; verificar e eliminar sequências com caracteres inválidos - tais como a presença de N flanqueando os códigos bialélicos e sequências incompletas; e somar para cada posição os valores absolutos de A, C, G e T encontrados para o cálculo da frequência relativa de cada nucleotídeo.

3.4 Cálculo da frequência basal do genoma

Estudos prévios, para verificar se existe um viés na vizinhança do SNP, usaram de comparações com a composição global de nucleotídeos no genoma (ZHAO, 2002), no entanto, sabe-se que dentro de um mesmo genoma, grandes regiões cromossômicas também podem diferir grandemente em sua composição nucleotídica (ROMIGUIER; ROUX, 2017). Para uma comparação precisa entre a região que contêm o SNP e o que seria esperado naquela região genômica na ausência de um SNP, propomos nesse trabalho uma nova maneira de calcular a frequência basal do genoma para cada região específica de nossa análise.

Para isso, criamos primeiramente uma base de dados de sequência em um de nossos servidores utilizando a ferramenta *makeblastdb*, fornecendo como *input* as sequências genômicas em formato fasta de cada organismo analisado neste trabalho. O *makeblastdb* faz parte do pacote do BLAST e gera um banco de dados de sequência que permite tanto a realização de busca por similaridade quanto o rápido acesso a sequência em um determinado intervalo de posições. A anotação do genoma, no formato ".gff", também foram obtidos no site do NCBI, para identificar as posições do genoma que correspondem às quatro regiões analisadas neste trabalho (3'UTR, 5'UTR, intrônicas e CDS-sinônimas). As posições correspondentes a cada uma das regiões genômicas em análise foram recuperadas com o auxílio de um script em Perl (*get_pos_gff_plus5.pl*). Estas posições foram, posteriormente, submetidas ao programa *blastdbcmd*, que também faz parte do pacote BLAST, para acessar as sequências que estejam dentro do intervalo dessas posições. O formato padrão do arquivo de saída do programa *blastdbcmd* é o formato multiFasta.

Esse arquivo multiFasta contendo as sequências específicas de uma determinada região genômica, serve como entrada para o script *join_seq.pl*. Este script sorteia n sítios aleatórios e

determina a frequência das bases nas posições vizinhas ao sítio sorteado. O script permite que o usuário restrinja o sorteio em posições onde ocorra um determinado conjunto de tipos de nucleotídeos, permitindo assim a simulação da ocorrência de determinados SNPs ao acaso. Este procedimento pode ser repetido m vezes, e após todo o procedimento o algoritmo calcula a média e o desvio padrão das frequências de cada nucleotídeo em cada uma das posições vizinhas ao sítio sorteado.

O comando para a execução desse script, pode ser descrito na seguinte forma:
"perl join_seq.pl <fasta_file> <snp> <número de sítios aleatórios> <número de iterações>".
Por exemplo, quando simulamos a substituição W numa dada região genômica, o fizemos conforme declarado abaixo:
"perl join_seq.pl multiFasta_daRegiãoGenomica_organismo.fas AT 100.000 500".
Neste exemplo, o programa sorteará 100.000 sítios aleatórios e determinará a frequência de cada nucleotídeo em cada posição vizinha ao sítio sorteado. Este procedimento é repetido 500 vezes, e no final das 500 repetições, o algoritmo determinará a média e o desvio padrão das frequências obtidas em cada repetição. O valor "AT" fornecido ao programa restringe o sorteio apenas para aquelas posições que contiver um A ou um T. Isto nos permite verificar os valores das frequências nucleotídicas vizinhas esperado caso a mutação do tipo W (A/T) ocorra de forma aleatória. As frequências obtidas a partir deste procedimento são denominadas de **frequências basais**.

Para estabelecer os parâmetros do sorteio, levamos em consideração o tamanho dos genomas aqui analisados, sendo o menor deles o genoma de *Sus scrofa* - que possui 2,478,444,698 pares de bases (ENSEMBL, [s.d.]) - e o maior dele de *Mus musculus* - possuindo 3,486,944,526 pares de bases (ENSEMBL, [s.d.]); e construímos uma série temporal variando a quantidade de posições sorteadas entre 100 e 10.000.000; e a quantidade de iterações realizadas: entre 500 e 50.000. Estima-se que para a cobertura total de um genoma com boa qualidade estima-se que este deva ser seqüenciado ao equivalente a dez vezes seu tamanho em pares de bases - sendo que, nas primeiras versões do genoma humano, cerca de 80% do conteúdo apresentava uma cobertura de cinco vezes (SANTOS; ORTEGA, 2007). Como não estamos montando um genoma, procuramos delimitar parâmetros em que cada base fosse amostrada ao menos uma vez na cobertura analisada - dados que posem ser visualizados na Tabela 2.

Tabela 2: Cobertura de cada sorteio de posições e repetições conforme o tamanho do menor e do maior genoma analisado.

Quantidade de Posições Sorteadas	Quantidade de Repetições dos Sorteios	Cobertura para com o genoma de *Sus scrofa*	Cobertura para com o genoma de *Mus musculus*
10 000 000	500	2,0173942	1,4339201
1 000 000	500	0,2017394	0,1433920
100 000	500	0,0201739	0,0143392
10 000	500	0,0020174	0,0014339
1 000	500	0,0002017	0,0001434
100	500	0,0000202	0,0000143
100000	5 000	0,2017394	0,1433920
100000	50 000	2,0173942	1,4339201

Observamos que, mantendo o número de repetições dos sorteios e alterando a quantidade de posições em cada sorteio o desvio-padrão (da frequência média de cada nucleotídeo em cada posição), diminui progressivamente conforme aumentamos a quantidade de posições sorteadas, e, quanto menor o desvio-padrão, é um indício que os pontos dos dados tendem a estar próximos da média. No entanto, conforme aumentamos o número de sorteios, os valores de desvio-padrão também aumentaram - para uma análise comparativa das variações de desvio-padrão, temos na Tabela 3 a média dos desvios-padrões dos diferentes elementos da em um estudo de caso da série temporal.

Tabela 3: Média dos desvios-padrões das frequências simuladas esperadas em diferentes sorteios de posições e repetições - um estudo de caso de regiões intrônicas de *Homo sapiens*.

	100 P x 500 R	1.000 P x 500 R	10.000 P x 500 R	100.000 P x 500 R
Média dos desvios-padrões	0,042721085	0,013507705	0,004327903	0,00133325

	1.000.000 P x 500 R	10.000.000 P x 500 R	100.000 P x 5.000 R
Média dos desvios-padrões	0,000429626	0,000134946	0,001356983

Legenda:

P = Quantidade de Posições selecionadas em 1 sorteio.

R = Quantidade de Repetições dos sorteios realizados.

Uma análise estatística par a par, através do teste de qui-quadrado entre as séries foi realizada para verificar se há diferença signifcativa entre as proporções de cada nucleotídeo flanqueando a substituição simulada - em todos os casos analisados, o *p-value* ficou entre 0.99346676979491511 e 1, dessa forma, não havendo diferença significativa entre as frequências das diferentes quantidades de posições sorteadas e iterações de sorteio - um exemplo da variação da frequência de um nucleotídeo conforme a série temporal é mostrado na Figura 9. Com esse resultado, optamos por um valor intermediário de posições sorteadas bem como quantidade de repetições de sorteio, de forma ter um menor custo computacional. Nesse estudo, para cada região genômica, em cada organismo, sorteamos 100.000 sítios aleatórios e repetimos esse sorteio 500 vezes.

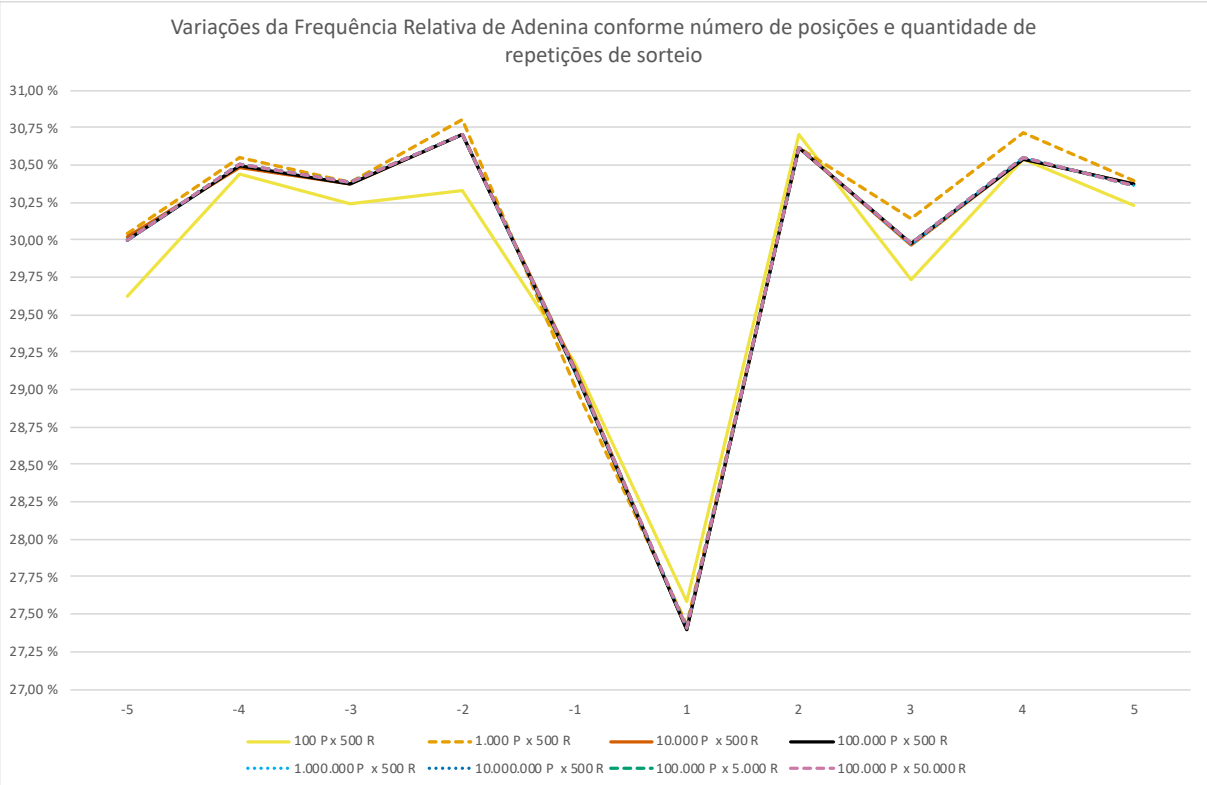


Figura 9: Frequência relativa de Adenina em torno de um SNP simulado W em *Homo sapiens*, com diferentes quantidades de posições sorteadas e repetições do sorteio. No eixo X temos a frequência relativa dessa base em comparação com as demais (Contagem absoluta de Adeninas / Conteúdo absoluto de todos os nucleotídeos (A, T, C e G)) para cada respectiva distância do SNP simulado, visualizado no eixo Y.

Legenda:

P = Quantidade de Posições selecionadas em 1 sorteio.

R = Quantidade de Repetições dos sorteios realizados.

3.5 Comparação das Frequências Basais do Genoma e na vizinhança do SNP

Estudos prévios que envolviam a comparação entre as frequências observadas e esperadas de um SNP mostraram que o teste qui-quadrado - uma vez que nenhum valor seja inferior a 5 - é o mais satisfatório (PLYLER et al., 2015). Tendo as saídas dos dados observados (frequência dos nucleotídeos na vizinha dos SNPs) e esperados (frequências basais) em um formato padrão tabular contendo os dados da posição, da base e da frequência relativa, foi criado um script em Python (*iterador_singlets_sinteticosXobservados_v2.py*), que recebe esta tabela como entrada e que realiza a comparação dos dados observados e esperados através do teste qui-quadrado para cada posição.

Um resumo de todas as etapas da metodologia até a comparação das frequências basais e do genoma na vizinhança de um SNP, pode ser visualizado na Figura 9.

3.6 Determinação da frequência de dinucleotídeos por região

A frequência de cada um dos 16 dinucleotídeos foi estimado a partir do arquivo multiFasta de cada região e organismo gerado na saída do *blastdbcmd*. Esse arquivo é utilizado como entrada para o script *contador_bidi_nucs.pl* que percorre as sequências de cada arquivo FASTA com uma janela deslizante de tamanho 2 e realiza a contagem de cada dinucleotídeo. Depois de percorrer todo o arquivo, os dinucleotídeos que contém bases ambíguas foram descartados da análise, já que eles são pouco informativos e também pouco em quantidades. Esse arquivo de texto final, tabular, contendo em uma coluna os dinucleotídeos e em outra as respectivas contagens é aberto no Excel onde realizamos os cálculos para a frequência relativa.

3.7 Experimentos de Aminação dos Genomas

Para visualizar os efeitos das ilhas CpG no efeito dos nucleotídeos vizinhos, revertemos a metilação da Citosina em Timina nas sequências do genoma, em diferentes proporções que vão de 5% a 95%, os dinucleotídeos CpA e TpG encontrados são substituídos por CpG. Esse procedimento foi realizado através do script *both_percentis_animados.pl*. Esse script recebe como entrada o arquivo multiFasta gerado na saída do *blastdbcmd*, percorre todo o arquivo e conta o conteúdo dos dinucleotídeos CpA e TpG.

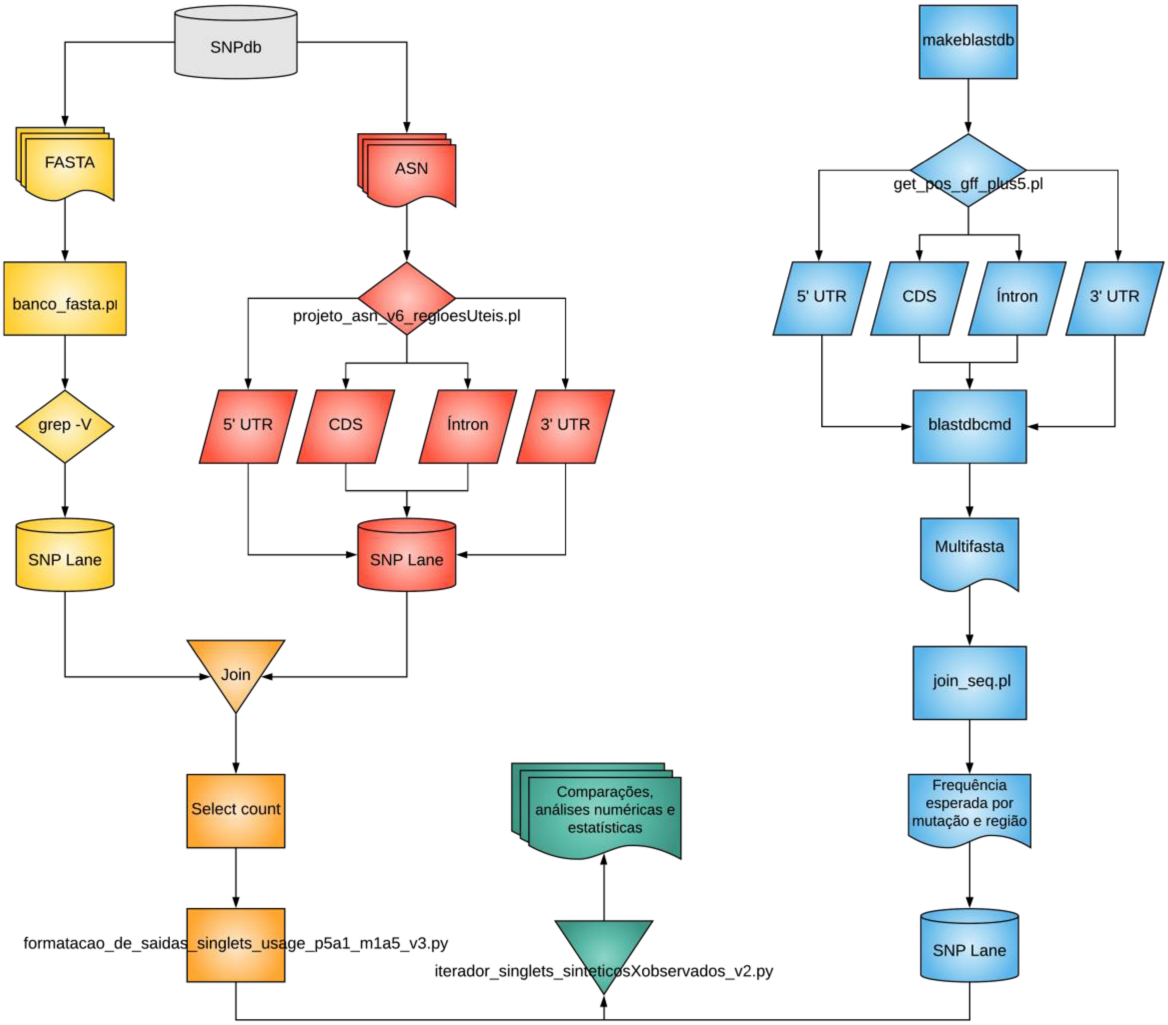


Figura 10: Resumo simplificado da metodologia para o estudo dos efeitos de vizinhança em SNPs. Para a obtenção dos dados de frequência observada - que se inicia a esquerda, para cada organismo da base de dados do dbSNP realizamos o download dos arquivos nos formatos FASTA e ASN. Parseamos os arquivos FASTA com o auxílio do script banco_fasta.pl, através dele, selecionamos da sequência a mutação e as 5 posições que a flanqueiam de cada lado. Dessa saída, no próprio terminal Linux," removemos os SNPs que são ambíguos para mais que duas bases e depositamos essas tabelas em nossa base de dados local. Para os arquivos em formato ASN, utilizamos o script projeto_asn_v6_regioesUteis.pl para filtrar os identificadores de cada mutação conforme a sua região de ocorrência: intron, cds, 3'utr e 5'utr. Também filtramos e removemos os identificadores de indels e depositamos essas tabelas em nossa base de dados local. No próprio banco, através de um comando join, concatenamos a informação da sequência presente no arquivo FASTA com a informação da região que se encontra presente no arquivo ASN. Realizamos a contagem de cada base em cada posição através de comandos do próprio MySQL e, então calculamos as frequências relativas através de um script.

Para a obtenção dos dados de frequência esperada - que se inicia a direita, criamos uma base de dados local para o organismo com a sequência de seu genoma e, buscamos as posições correspondentes a cada região. Utilizamos o blastdbcmd para receber a lista de posições de determinada região e retornar as sequências em um formato multiFasta. Através do script join_seq.pl sorteamos as posições que correspondem aos nucleotídeos que queremos simular, contamos as respectivas vizinhanças e executamos a repetição deste sorteio. Dessas contagens, calculamos a frequência média de cada base por posição e seu desvio-padrão. Tendo as saídas de dados observados e esperados em um formato padrão tabular que se consiste em posição, base e frequência relativa, utilizamos como entrada para o script iterador_singlets_sinteticosXobservados_v2.py que que realiza a comparação dos dados através de análises numéricas e estatísticas.

Depois extrai a média dentre esses nucleotídeos e multiplica pelo valor de cada percentil. Por exemplo, para 5% de aminação do genoma, a média das contagens de CpA e TpG é multiplicada por 0,05, dessa forma, o programa percorre a sequência substituindo os dinucleotídeos CpA e TpG por CpG encontrados até que aproximadamente 5% da média do conteúdo de CpA e TpG estejam aminados. Esse programa gera uma saída intermediária de arquivo multiFasta para cada percentual de aminação para cada organismo e região genômica. Para cada saída intermediária, esse script recruta o script join_seq.pl para cada uma das 6 substituições desse estudo - R, Y, M, K, W e S - para simular como seria a frequência basal do genoma em diferentes regiões genômicas, em diferentes organismos, com diferentes percentuais de aminação - ou diferentes percentuais de reversão da metilação de citosina.

Objetivando uma análise comparativa entre os diferentes percentuais de aminação, calculamos a distância euclidiana média das frequências dos nucleotídeos nas posições +1 e -1 com a média das frequências dos nucleotídeos nas posições +5 e -5, esse cálculo foi realizado através do script em Python *distEuclidiana_usageKMRYWS.py* que recebe como entrada os arquivos tabulares de frequências de nucleotídeo por posição - as saídas do *join_seq.pl*.

4. RESULTADOS

4.1 CONSTRUÇÃO DA BASE DE VIZINHANÇA DE SNPS: SNP LANE

Uma meta deste trabalho foi ofertar um site em que os dados de padrão de bases nas vizinhanças de SNPs estivessem disponíveis para a visualização, conforme descrito nos itens abaixo.

4.1.1 Obtenção e processamento de dados do dbSNP

De um total de 901.829.443 SNPs baixados, 392.591.139 substituições passaram nos filtros de qualidade: são substituições bialélicas, possuem registros nos arquivos FASTA e ASN e não são indels.

Os dados por organismo podem ser visualizados na Tabela 4, ainda que os filtros de qualidade reduziram em mais de 80% o volume dos dados de *Mus musculus,* o total de dados curados ainda é cerca de 30 vezes maior do que o volume de dados previamente analisado pela literatura.

Desses dados, mensuramos as porcentagens de SNPs nos registros de camundongos, bovinos, *Sus scrofa*, *Rattus norvegicus* e humanos da base de dados que correspondiam a transições (SNPs envolvendo bases púricas ou, bases pirimídicas, ex: R) ou transversões (quando envolvem uma purina e uma pirimidina, ex: M). É esperado que a frequência de transversões seja menor. Na Tabela 4 observa-se a porcentagem de transições e transversões nas subdivisões do dbSNP conforme o organismo. A exceção do genoma bovino, as proporções confirmam a razão esperada de três transições para cada transversão, conhecida como 3:1:1 (ex: de todos os SNPs, cerca de 30% envolvem a transição R ou Y, enquanto cerca de 10% representam cada uma das transversões M, K, W ou S). No genoma bovino tivemos cerca de 25% dos dados para transições, 15% dos dados nas transversões K e M e 10% nas substituições W e S - nos estudos prévios de *Jiang et al*. (2008) com a análise de 15.111 SNPs de *Bos taurus*, encontrou 69,8% de transições indicando um possível viés de não aleatoriedade na obtenção dos dados. No entanto, ainda que com diferentes proporções, nossa análise confirmou que a transição é o tipo mais abundante de SNPs - característica relatada em plantas e outros animais (JIANG et al., 2008).

Tabela 4: Frequência relativa de cada substituição conforme o organismo. É mostrado o total dos dados baixados (sem filtros) e o total utilizado em nossa análise e o quanto tivemos de redução do volume de dados inicial. Essa redução vem da remoção de mutação *indel* e de identificadores que estavam presentes em apenas um dos formatos (somente ASN ou somente FASTA). Como fator de comparação, é mostrado o volume de dados que havia sido analisado previamente pela literatura.

Tipo de Substituição	Mus musculus	Bos taurus	Sus scrofa	Rattus norvegicus	Homo sapiens
R	33,50 %	25,11 %	33,54 %	34,28 %	32,89 %
Y	33,53 %	25,35 %	33,52 %	34,10 %	32,81 %
K	8,85 %	14,45 %	9,64 %	8,45 %	8,86 %
M	8,85 %	14,53 %	9,65 %	8,42 %	8,80 %
S	6,75 %	10,96 %	6,55 %	6,85 %	9,10 %
W	8,52 %	9,59 %	7,09 %	7,90 %	7,55 %
Total:[a]	13 203 254	38 105 942	31 447 096	1 764 841	308 070 006
Sem Filtros: [b]	71 236 064	102 386 869	63 882 271	4 755 623	659 568 616
Redução: [c]	81,47 %	62,78 %	50,77 %	62,89 %	53,29 %
Análises anteriores: [d]	433 192	15 110	-	-	8 249 446
Publicações das análises anteriores:	ZHANG, Fengkai; ZHAO, Zhongming. The influence of neighboring-nucleotide composition on single nucleotide polymorphisms (SNPs) in the mouse genome and its comparison with human SNPs. Genomics, [s.l.], v. 84, n. 5, p.785-795, nov. 2004. Elsevier BV. http://dx.doi.org/10.1016/j.ygeno.2004.06.015.	JIANG, Z. et al. The Complementary Neighborhood Patterns and Methylation-to-Mutation Likelihood Structures of 15,110 Single-Nucleotide Polymorphisms in the Bovine Genome. Genetics, [s.l.], v. 180, n. 1, p.639-647, 24 ago. 2008. Genetics Society of America. http://dx.doi.org/10.1534/genetics.108.090860.	-	-	ZHAO, Zhongming; ZHANG, Fengkai. Sequence context analysis of 8.2 million single nucleotide polymorphisms in the human genome. Gene, [s.l.], v. 366, n. 2, p.316-324, fev. 2006. Elsevier BV. http://dx.doi.org/10.1016/j.gene.2005.08.024.

[a] total utilizado em nossa análise; [b] total dos dados baixados (sem filtros); [c] redução do volume de dados inicial; [d] volume de dados analisado previamente pela literatura.

4.2 Transições e transversões nas subdivisões do dbSNP

A Figura 11 mostra a frequência de transições e transversões nas subdivisões do dbSNP conforme a região genômica: 3'UTR, 5'UTR, íntrons e CDS-sinônimos.

4.2.1 3'UTR

Nas regiões de 3'UTR (Figura 11a), observamos que em *Homo sapiens*, *Mus musculus*, *Rattus norvegicus* e *Sus scrofa* as substituições R e Y representam, cada uma, a cerca de 33% das substituições. Em *Bos taurus* cada transição corresponde a cerca de 19% das substituições bialélicas, valor próximo ao encontrado as substituições K e M (aproximadamente 18%). As transversões W e S em bovinos também apresentam uma maior frequência relativa com valores de 11,37% e 12,82% respectivamente, em relação aos demais organismos, que apresentam valores para W entre 8,18% (*Mus musculus*) e 6,49% (*Sus scrofa*) e valores para S entre 6,30%

(*Rattus norvegicus*) e 9,47% (*Homo sapiens*). Assim, analisando por região, os dados de bovinos aparentemente estariam enviesados.

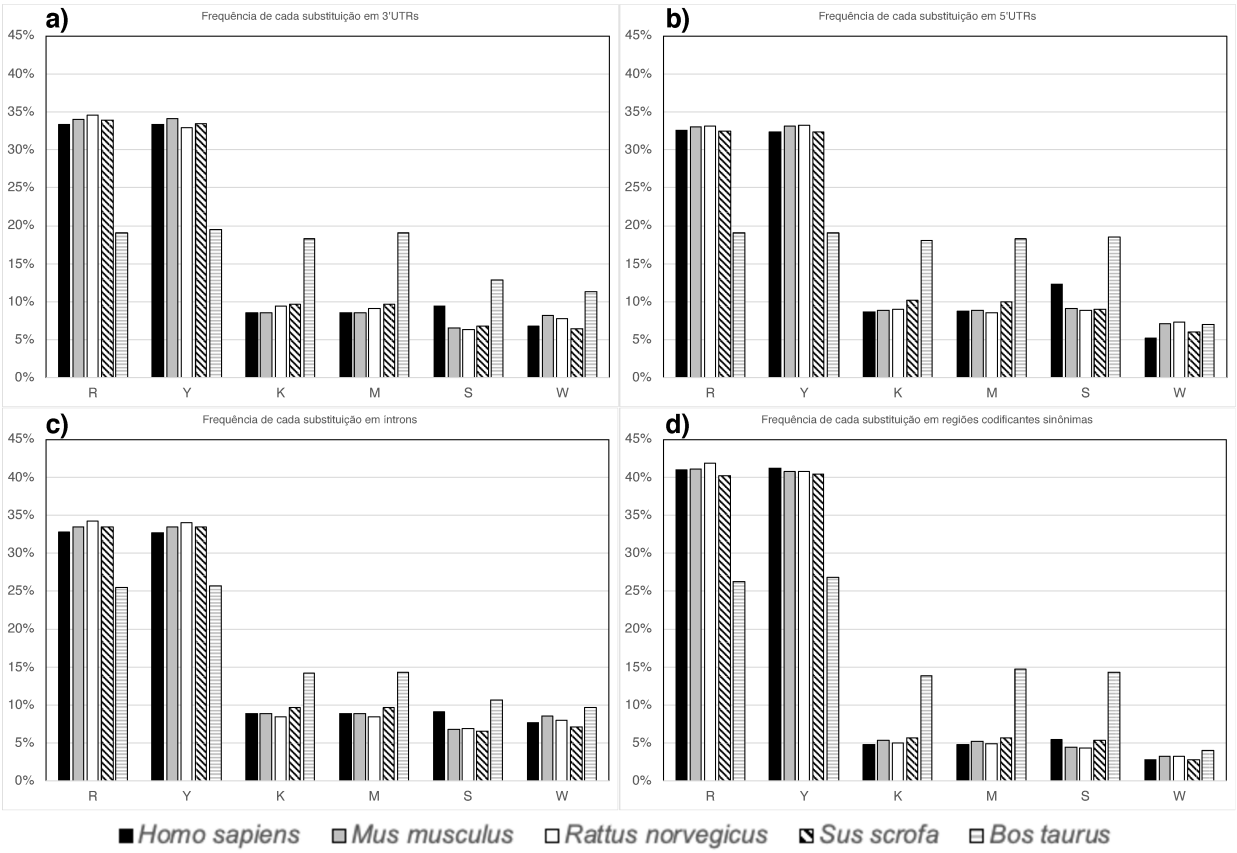


Figura 11: Frequência relativa de cada substituição conforme organismo. Na Figura 10 temos no eixo Y as substituições analisadas (R, Y, M, K, W, S) e, no eixo Y temos a frequência relativa de cada uma dessas substituições nos diferentes mamíferos analisados. Na investigação das proporções entre transições (R e Y) e transversões (K, M, S e W) das entradas filtradas do dbSNP. Para esses organismos, os SNPs foram classificados após a região onde o SNP está localizado: Intron, 5-UTR, 3-UTR, e CDS-sinônimo. Nota-se que o depósito para SNPs de bovinos sofreu de alguma forma um viés que alterou as proporções relativas entre essas categorias - há uma menor frequência de transições e uma maior frequência de transversões em comparação aos demais organismos), sugerindo que a amostragem não é aleatória no projeto da bovinos. A frequência de transições variou em torno de 33% para todas as categorias, a exceção de CDS-sinônimo, em que fora superior a 40%, para o homem, camundongo, rato e porco. Assim, todas as quatro transversões foram menores para o sinônimo de CDS. Transversões S também foram menores para Intron e 3-UTR e especialmente maiores em 5-UTR.

4.2.2 5'UTR

Nota-se nas regiões 5'UTR (Figura 10b), que o genoma bovino apresenta frequências de transição e transversões entre aproximadamente 18% e 19% para R,Y, K, M e S; somente a substituição W apresenta um valor próximo para com os demais organismos (7,01%). Para os demais mamíferos há um pequeno desvio na razão esperada de três transições para cada

transversão, os valores para as transversões ficam entre 32% e 33%. Em camundongos e *Rattus norvegicus*, K e M representam cerca de 8% das substituições, os valores para S ficam próximos de 9% e os valores para W próximos de 7%. *Sus scrofa* apresenta os valores das transições K e M próximos a 10%, cerca de 9% em S e 6% em W. Por sua vez, em *Homo sapiens* temos a maior frequência relativa da substituição S (12,34%) e a menor frequência relativa da substituição W (5,28%) e os valores de K e M são 8,71% e 8,74% respectivamente.

4.2.3 CDS-Sinônimas

De todas as regiões analisadas, as regiões codificadoras sinônimas foram as que apresentaram maior divergência da proporção de três transições para cada transversão em todos os organismos. Em bovinos, a frequência relativa de cada transição é cerca de 26%, as transversões K, M e S tem valores aproximadas de 14% e W, bem como nos demais organismos, apresenta menor frequência relativa dentre as substituições, 4,04%.

Mus musculus, *Sus scrofa*, *Rattus norvegicus* e *Homo sapiens* apresentam uma distribuição semelhante, conforme mencionado. Substituições do tipo W apresentam as menores frequências relativas nessa região variando de 2,85% (*Sus scrofa*) e 3,22% (*Mus musculus*). R e Y têm suas frequências relativas em torno de 40%, e K, M e S apresentam valores entre 4,29% (S em *Rattus norvegicus*) e 5,48% (S em *Homo sapiens*).

4.2.4 Intron

Nas regiões intrônicas (Figura 11c) observa-se que em *Rattus norvegicus*, *Mus musculus* e *Sus scrofa* os valores das frequências relativas para cada substituição são bem próximos. Os valores das transições ficam entre 34,19% (R em *R. norvegicus*) e 32,42% (R em camundongos) e as transversões K e M variaram entre 9,68% (para K e M em *Sus scrofa*) e 8,45% (transição M em *R. norvegicus*). Frequência relativa de S apresentou valores próximos a 6% e de W, entre 8,6% (*Mus musculus*) e 7,15% (*Sus scrofa*). Por sua vez, em humanos, os valores para R,Y, K, M e W são bem próximos - 32,79%, 32,71%, 8,91%, 8,85% e 7,65%, respectivamente - somente com substituição S apresentando uma maior frequência relativa de 9,08%. Novamente, *Bos taurus* se distancia da proporção 3:1:1, com a frequência das transversões em torno de 25%, K e M com aproximadamente 14%, e W e S com cerca de 10%.

4.3 Determinação da frequência de base na vizinhança de cada tipo de substituição

Nessa sessão, para inferir se havia um efeito da vizinhança sobre um dado tipo de SNP nas amostras contidas em nosso banco de dados, determinamos a ocorrência dos nucleotídeos vizinhos ao SNPs, por meio de uma contagem simples determinamos a frequência de cada base (em percentuais) e contrastamos através de um teste qui-quadrado com a frequência basal do genoma, determinada como em "Material e métodos". Resumidamente, quando estudamos a transversão A/T (W), o algoritmo sorteia uma base, se for A ou T continua, e registra as bases vizinhas. Quando alcança um total de sorteios, 500 em nosso caso, calcula as porcentagens de ocupação das posições vizinhas. Representamos graficamente apenas alguns casos como exemplo, os demais em sua totalidade, podem ser visualizados nas Figuras Suplementares 1 - 100 e, através da base de dados SNP Lane (Figura 12), no endereço http://bioinfo.icb.ufmg.br/snplane/. A ferramenta SNP LANE distingue este trabalho dos anteriores ao disponibilizar acesso rápido a todos os gráficos.

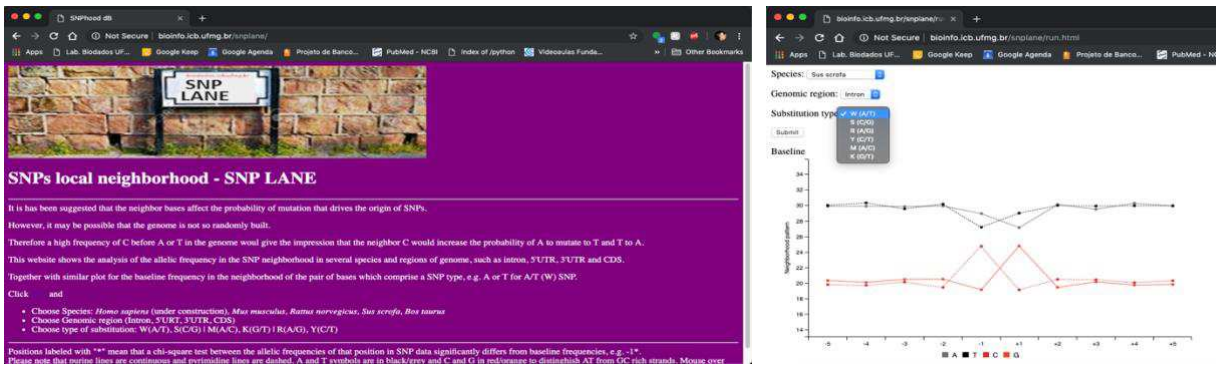


Figura 12: Interface web da base de dados SNP Lane. Na figura a esquerda temos a "*Homepage*" de abertura e, na figura a direita, temos a página de seleção das categorias a serem visualizadas com um menu "*dropdown*" para cada categoria.

Os SNPs foram classificados de acordo com o organismo (5 organismos), tipo de substituição (6 tipos de substituições), região genômica (5'UTR, 3'UTR, intron e CDS) e posição em torno do local da mutação (posições -5 a -1 e 1 a 5 a partir do SNP), totalizando 1200 categorias possíveis. Em 92 dessas categorias (7,6%) a diferença entre a frequência basal e a frequência observada nos locais de mutação foram significativamente distintas de acordo com o teste do qui-quadrado (Tabela 1 do apêndice). A maioria deles 56 casos (4,6% do total) foram encontrados em *Bos taurus*, um organismo que a análise prévia das proporções de

transições e transversões sugere que a amostragem não é aleatória no projeto dos bovinos. Dos 36 casos (3,0%) que pertencem a outros organismos aproximadamente 10 (0,8%) correspondem a mutações K ou M devido às citosinas em CpG dinucleotídeos que podem ser metilados para formar 5-metilcitosinas e então desaminados para Timina.

4.3.1 Frequência de base na vizinhança de R

As frequências basais do genoma nas vizinhanças de R se apresentam de forma similar mesmo em diferentes regiões e em diferentes organismos. Conforme observado nos gráficos inferiores das Figuras 13-20; a frequência relativa de Citosina apresenta um leve declínio das posições -5 a -2, na posição -1 apresenta o menor conteúdo dessa base e esse valor sobe progressivamente da posição +1 até a +5. A frequência relativa de Guanina sobe progressivamente das posições -5 até -1, tendo seu valor mais alto na posição +1, valor este que decresce na posição +2, decresce pouco nas posições +3 e +4, variando pouco até a posição +5. Esse pico de Guanina na posição +1 juntamente com o ponto mais baixo de citosina da posição -1 , pode ser em função do efeito de desaminação de CpG que pode incrementar a frequência do dinucleotídeo TpG e diminuir a frequência de CpG. O conteúdo de Adenina varia pouco nas posições -5 a -3 da frequência basal do genoma, tem um conteúdo gradativamente maior em -2 e atingindo a maior frequência relativa na posição -1, decrescendo na posição +1 a um valor próximo àquele da posição -2 e tendo um segundo pico - porém menor que o da posição -1 - na posição +2. Temos valores menores e similares entre si nas posições +3, +4, +5. Por sua vez, a frequência de Timina varia pouco da posição -5 a -3, tendo um declínio na posição -2 e um aumento da frequência relativa na posição -1 para então ter o seu ponto mais baixo na posição +1. Então, a frequência de Timina aumenta gradativamente nas posições +2 e +3, variando pouco na frequência entre +3, +4 e +5.

Na substituição R, todos os casos em que a frequência observada se difere significativamente da frequência basal do genoma, se encontram na posição -1. Outro fato comum observado é o conteúdo GC ser maior que o conteúdo AT nas regiões 5'UTR e CDS, e o conteúdo AT, por sua vez, ser superior ao conteúdo ao GC nas regiões 3'UTR e CDS.

4.3.1.1 Frequência de base na vizinhança de R em regiões 3'UTR

A região de 3'UTR em *Homo sapiens, Rattus norvegicus* e *Mus musculus* (Figuras Suplementares 4 e 7 do apêndice e Figura 13, respectivamente) apresentaram diferença

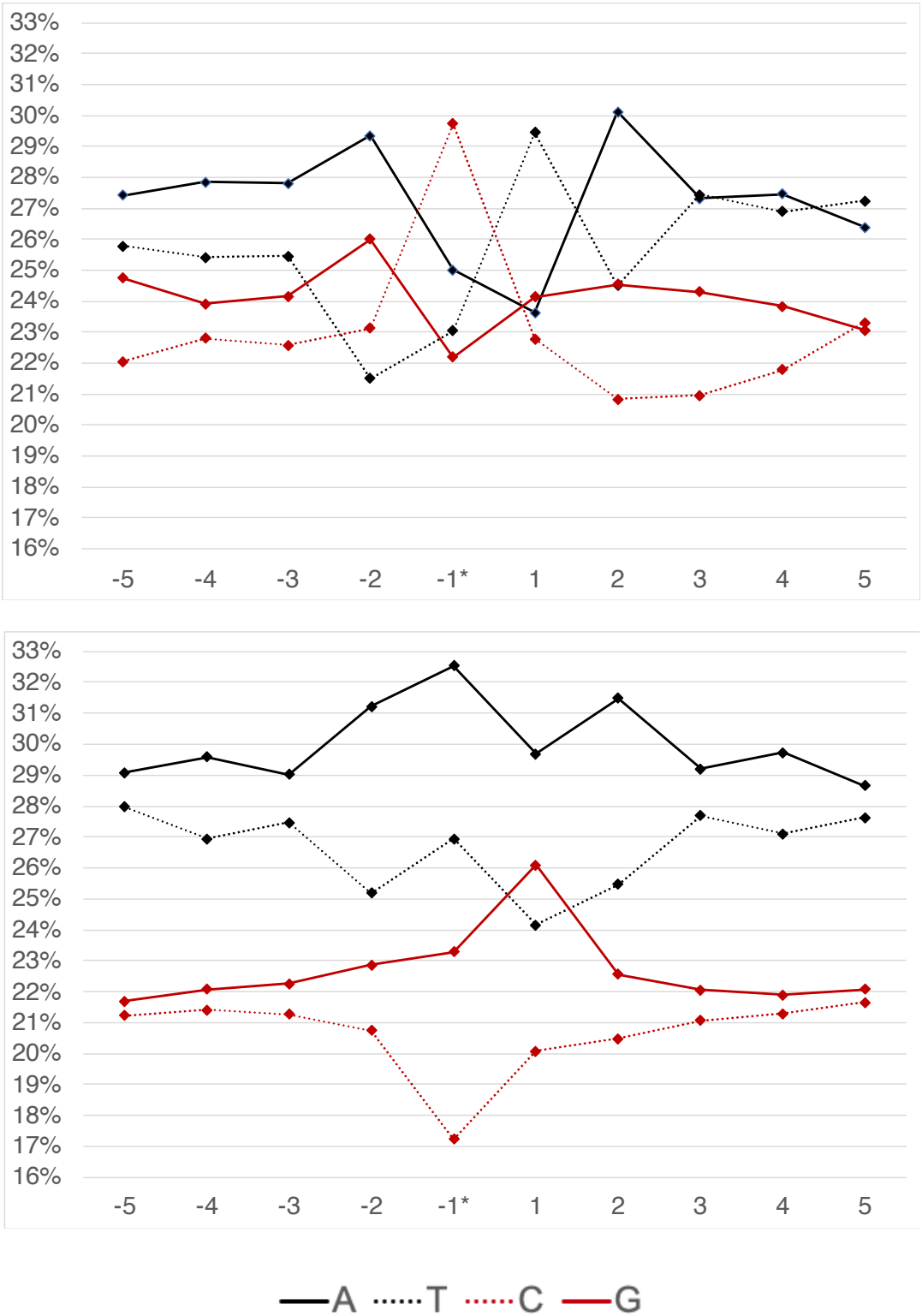


Figura 13: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões 3'UTR em camundongos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma *"baseline frequency"* no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a *"baseline frequency"*.

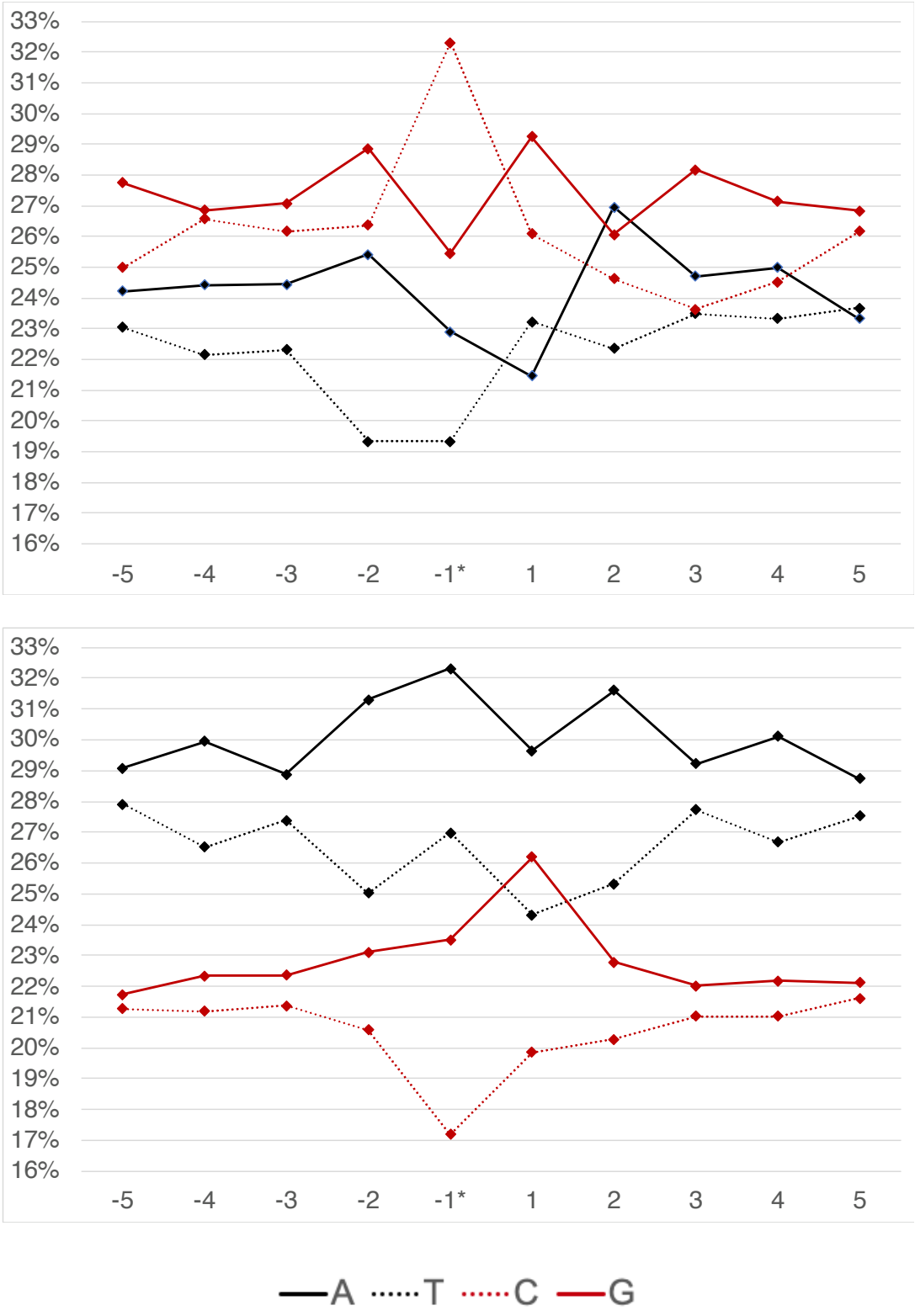


Figura 14: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões 5'UTR em camundongos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma *"baseline frequency"* no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a *"baseline frequency"*.

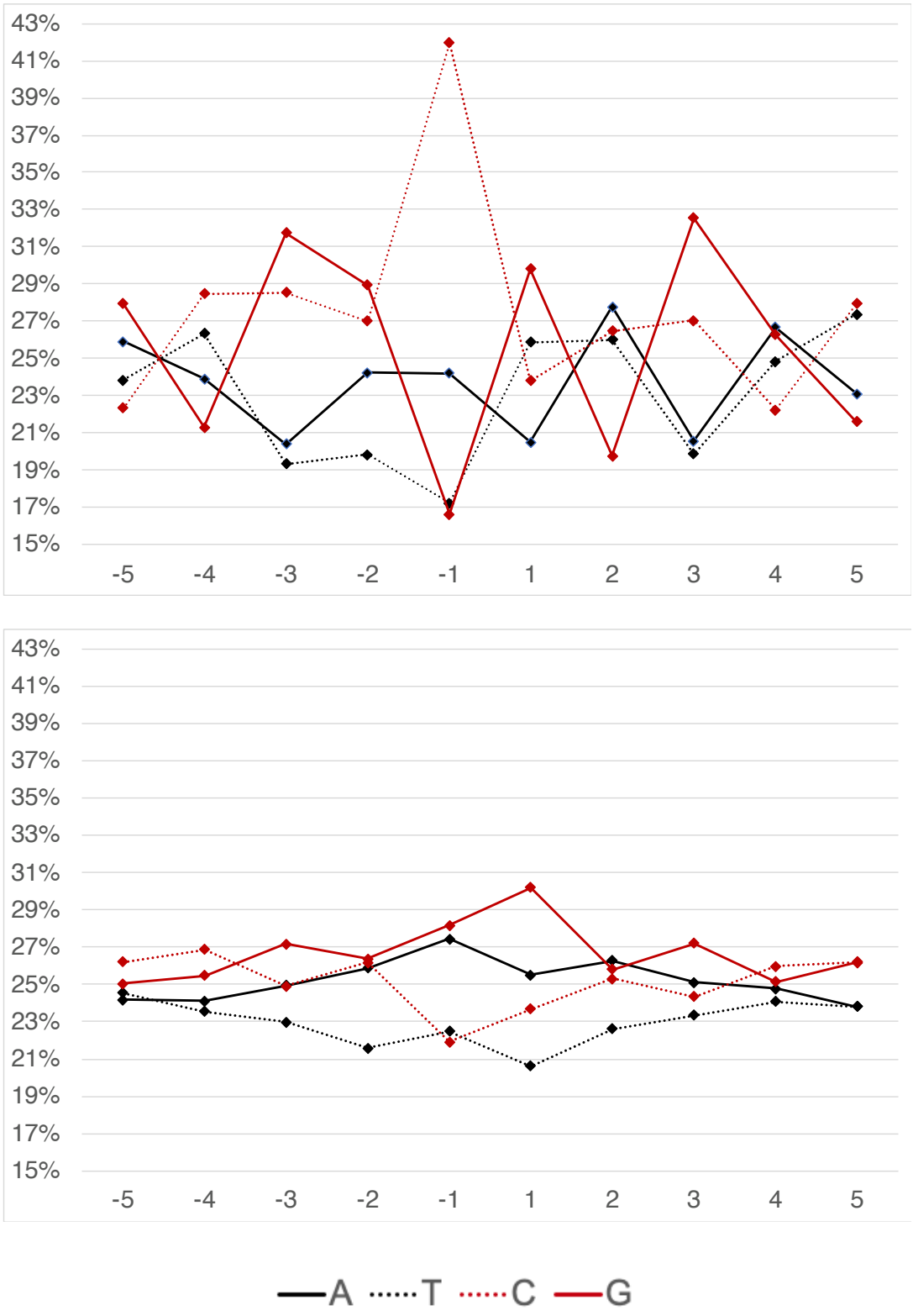


Figura 15: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões sinônimas em camundongos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma *"baseline frequency"* no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

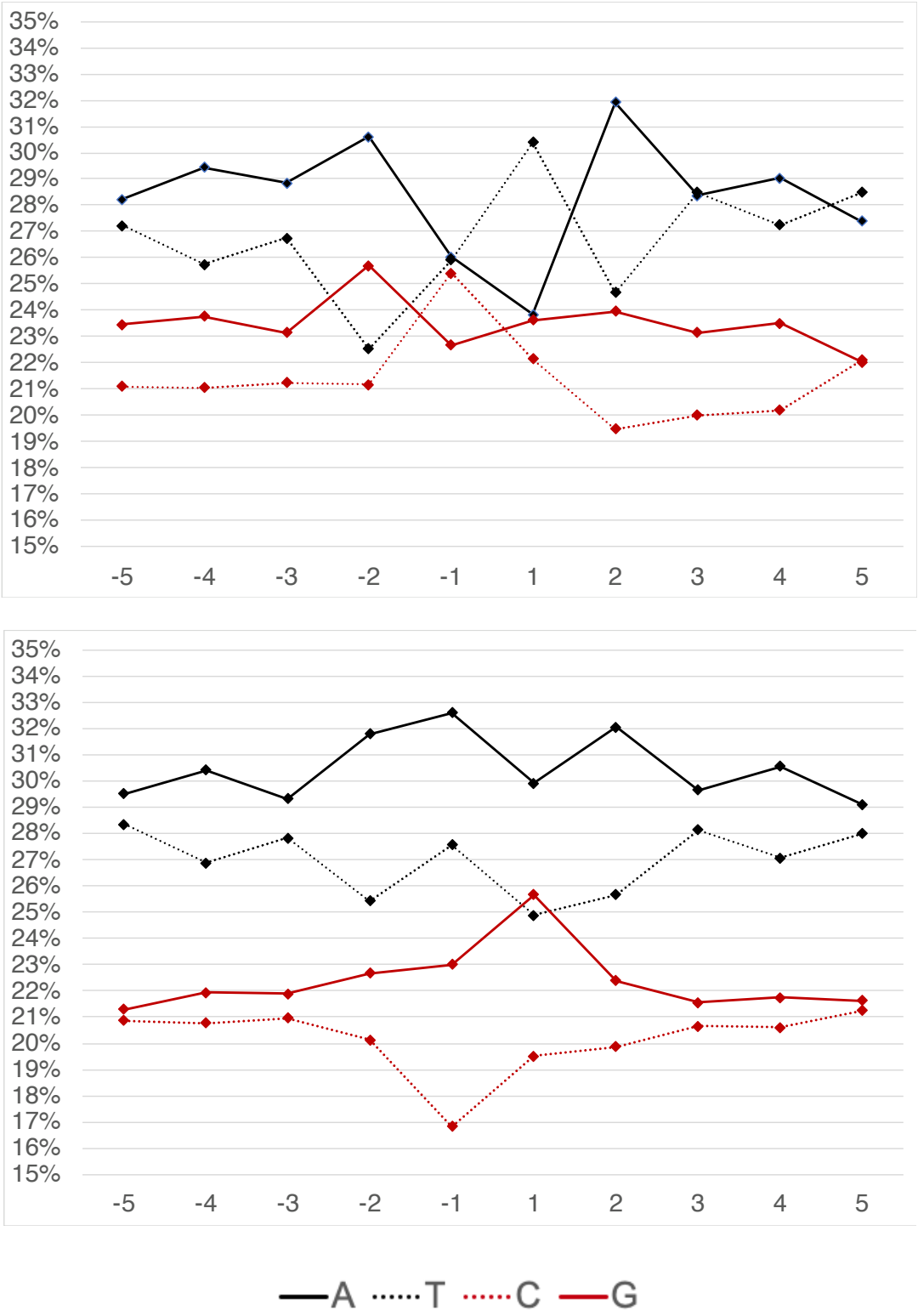


Figura 16: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões intrônicas em camundongos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma *"baseline frequency"* no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

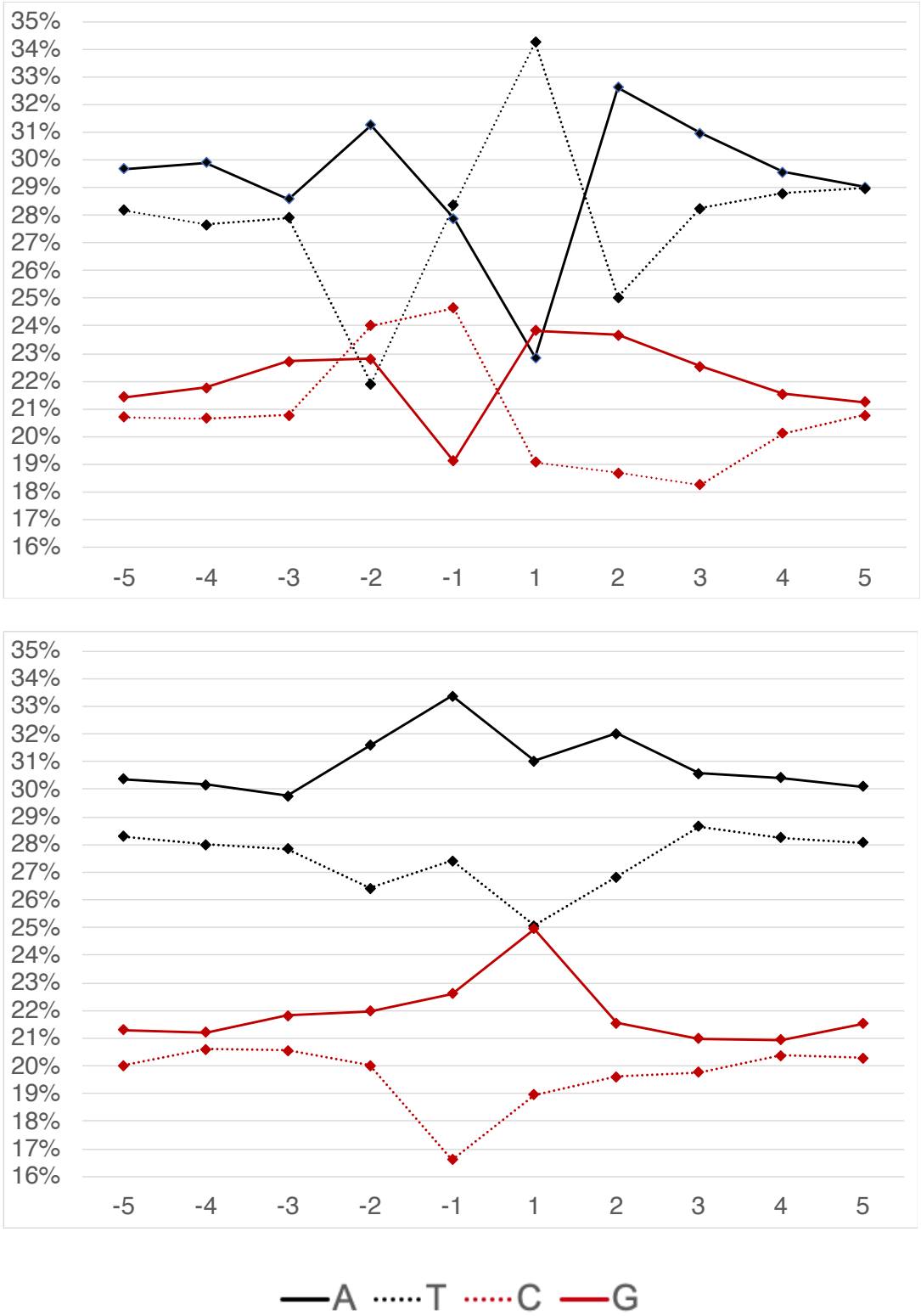


Figura 17: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões intrônicas em humanos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma *"baseline frequency"* no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

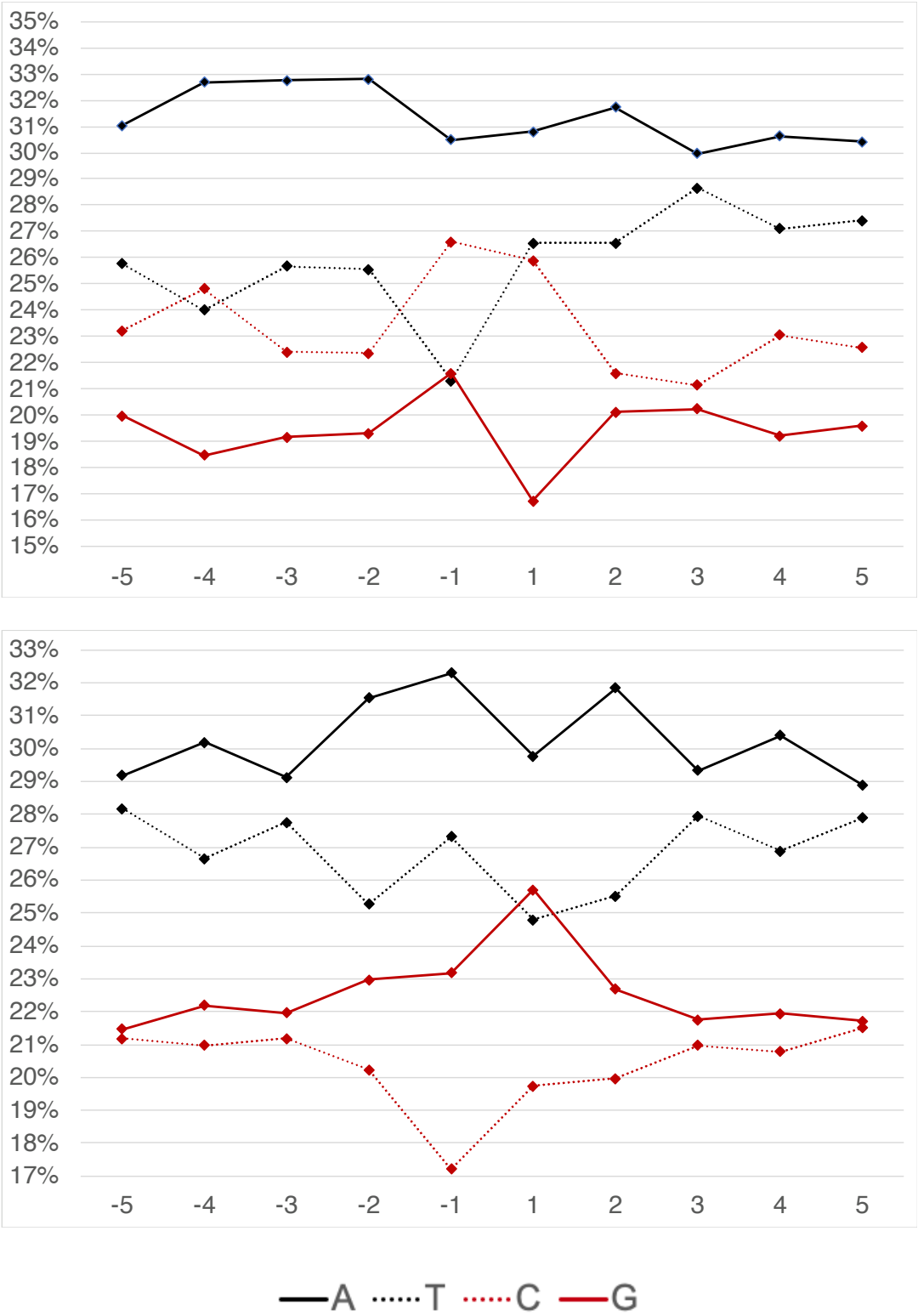


Figura 18: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões intrônicas em *Rattus norvegicus*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma *"baseline frequency"* no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

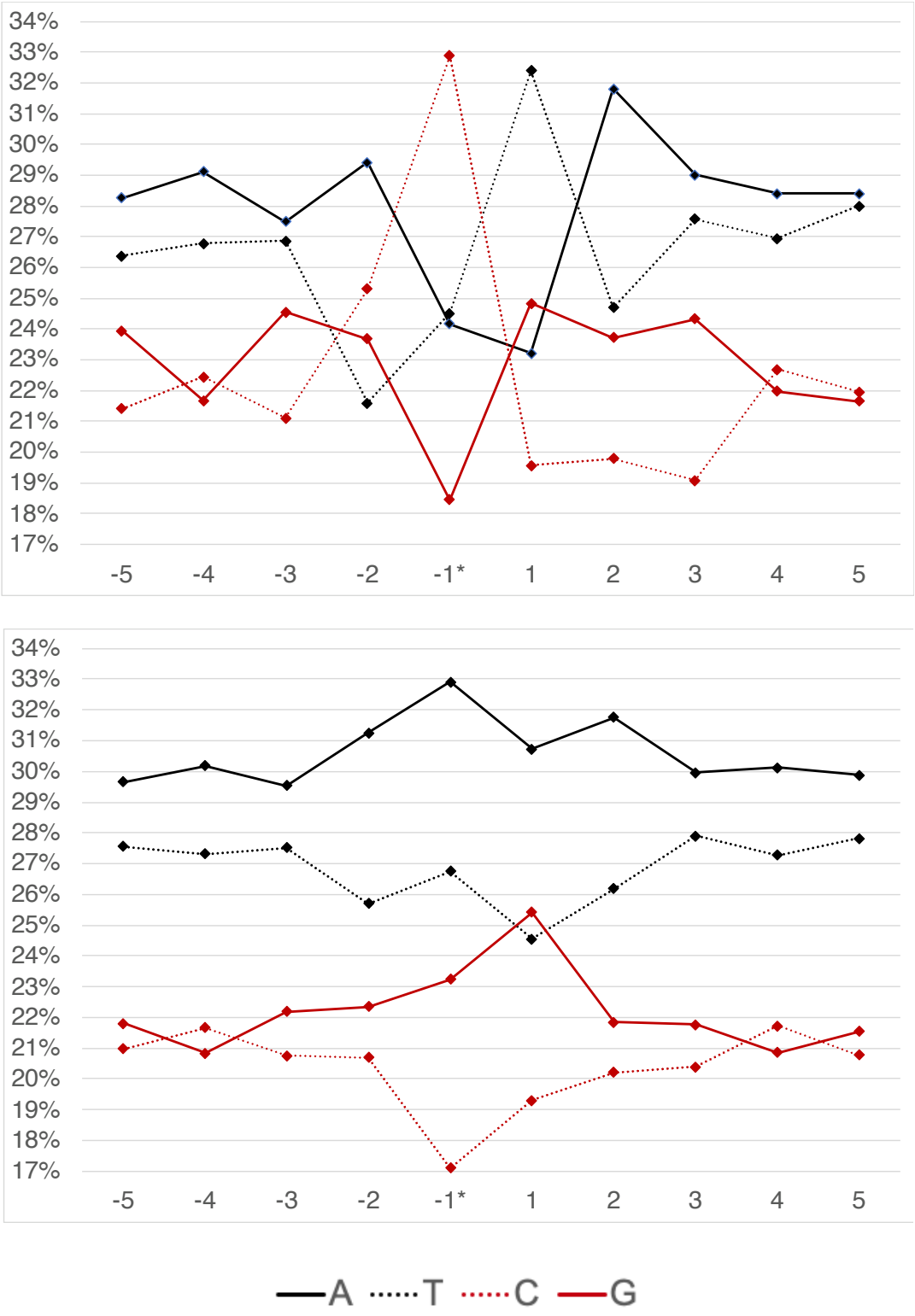


Figura 19: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões intrônicas em *Sus scrofa*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma *"baseline frequency"* no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

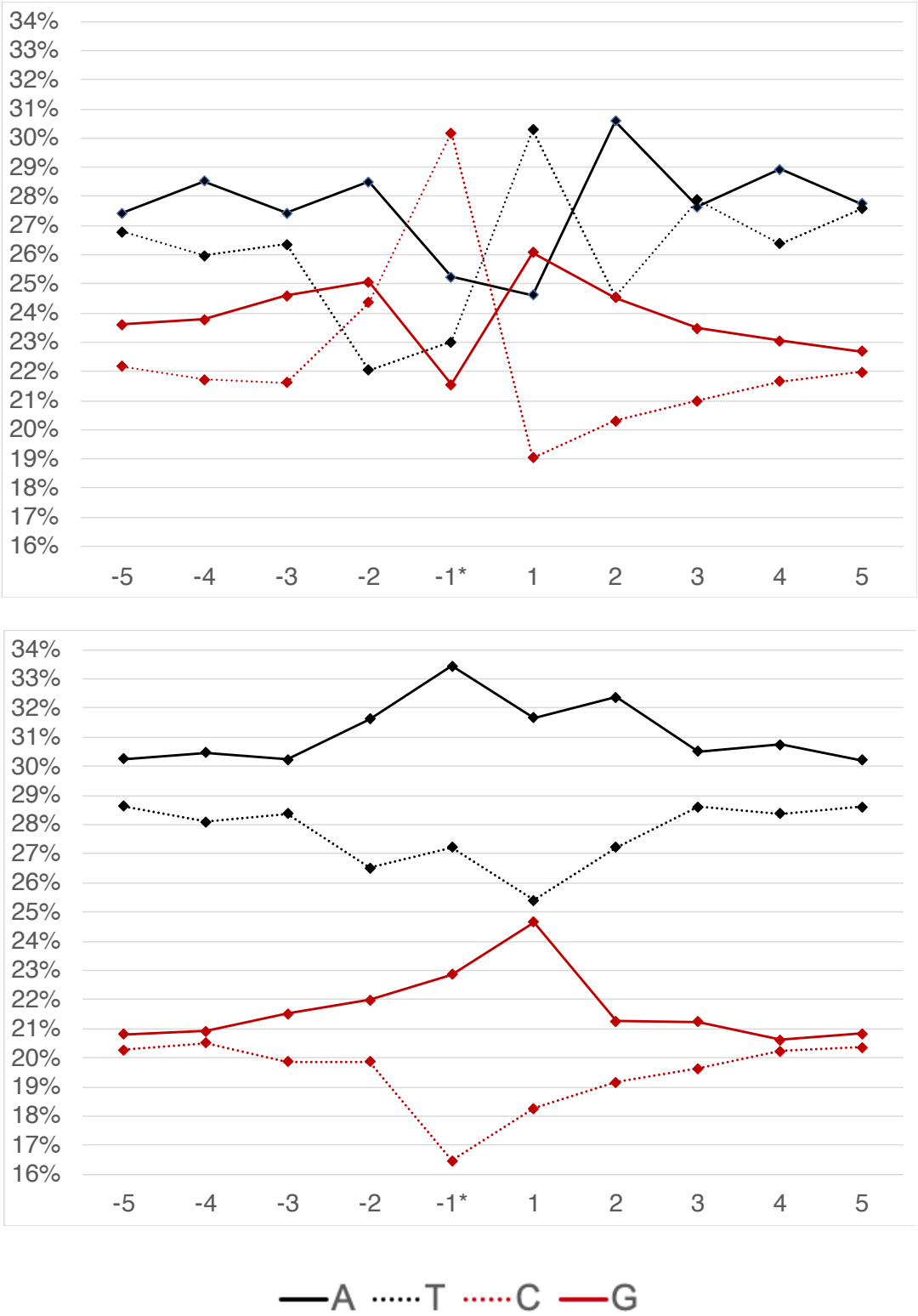


Figura 20: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões intrônicas de bovinos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma *"baseline frequency"* no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a *"baseline frequency*".

significativa entre a frequência basal do genoma e a frequência da base na vizinhança do SNP na posição -1. Nesta posição, observamos algumas características comuns aos cinco organismos como o maior valor para citosina e, o menor valor de Guanina. A curva de Adenina se apresenta de forma semelhante em *Homo sapiens, Sus scrofa* (Figura Suplementar 10), *Bos taurus* (Figura Suplementar 1) e *Mus musculus* tendo em -2 a maior frequência relativa, decrescendo até a menor frequência relativa em +1. Nesses mesmos organismos, Timina apresentou menor frequência relativa na posição -2 e maior na posição +1.

4.3.1.2 Frequência de base na vizinhança de R em regiões 5'UTR

Na região 5`UTR, a comparação das frequências relativas de rato e camundongo na posição -1 apresentou diferença significativa entre a frequência basal do genoma e a frequência da base na vizinhança do SNP. Nessa região, *Bos taurus* (Figura Suplementar 2), se diferencia dos demais organismos, possuindo na posição +5 os maiores valores de frequência relativa para Adenina e Timina, e os menores valores para Citosina e Guanina. Nos demais organismos (Figura 14 e Figuras Suplementares 5, 8, 11) alguns padrões podem ser observados, como uma tendência da menor frequência de Citosina na posição +3, o maior e o menor valor de frequência relativa de Guanina nas posições +1 e -1 respectivamente, e uma menor frequência de Timina na posição -2. Em ratos, camundongos e humanos, nota-se também uma maior frequência de Adenina na posição +2.

4.3.1.3 Frequência de base na vizinhança de R em regiões CDS-Sinônimas

As curvas de frequência observada ao redor do SNP em regiões CDS-sinônimas (Figura 15 e Figuras Suplementares 3. 6. 9 e 12) são as mais distintas em comparação com as demais regiões, podendo admitir uma variação muito grande na frequência entre bases adjacentes. No entanto, os padrões entre os organismos são pronunciáveis, como: os picos de Guanina nas posições -3, +1 e +3 bem como seus pontos mais baixos em -1 e +2. A adenina apresenta um pico na posição +2 e menores valores de frequência relativa nas posições -3 e +3. Temos também, nos cinco organismos analisados, um pronunciado valor de Citosina na posição -1. A exceção de *Rattus Norvegicus* (Figura Suplementar 9), Timina apresenta uma tendência a maior frequência relativa na posição +2.

4.3.1.4 Frequência de base na vizinhança de R em regiões intrônicas

Nas regiões intrônicas, *Sus scrofa* (Figura 19) e *Bos taurus* (Figura 20) apresentaram diferença significativa entre a frequência basal e a frequência observada ao redor da mutação na posição -1. *Rattus norvegicus* (Figura 18) ainda que não apresente uma diferença estatisticamente significativa, apresenta uma curva de Adenina destoante dos demais organismos - que possuem frequência máxima na posição +2 e mínima na posição 1 - possuindo os maiores valores de frequência relativa entre as posições -5 a -2 e os menores entre +3, +4 e +5. A curva de Timina também encontra um consenso entre os organismos, tendo maior frequência em +1 e menor em -2, com exceção em ratos, que apresentam a maior frequência relativa na posição +3 e menor na posição -1. Citosina apresenta uma maior frequência em -1 nos cinco organismos e uma menor na posição +3 para ratos, *Homo sapiens* (Figura 17) e *Sus scrofa*. Para Guanina, *Mus musculus* (Figura 16) que apresenta um padrão divergente dos demais organismos que apresentam um maior valor de frequência relativa na posição +1 e o menor na posição -1.

4.3.2 Frequência de base na vizinhança de Y

Em função da complementaridade das substituições R e Y conforme apresentado na Figura 21, o espelhamento da frequência basal do genoma é bem nítido entre essas duas substituições, e compartilhado entre os cinco organismos e nas quatro regiões analisadas - a única particularidade é vista nas regiões sinônimas em que o conteúdo GC é maior que o

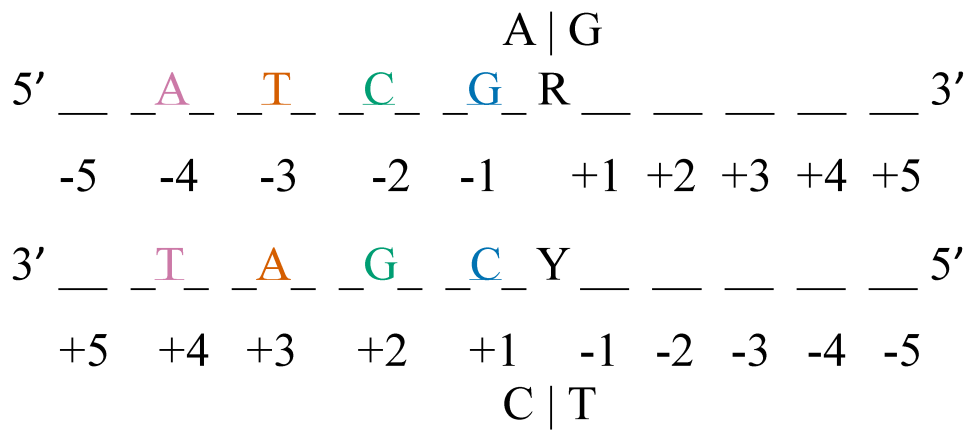


Figura 21: Representação dos locais vizinhos em transições complementares. Em função da complementariedade de bases, temos um espelhamento nos resultados em função da simetria das fitas, de tal forma que, por exemplo, se tivermos uma grande frequência relativa de Guanina na posição -1 em uma substituição R, é esperado que tenhamos essa mesma alta frequência relativa em Citosina na posição +1 em uma substituição Y.

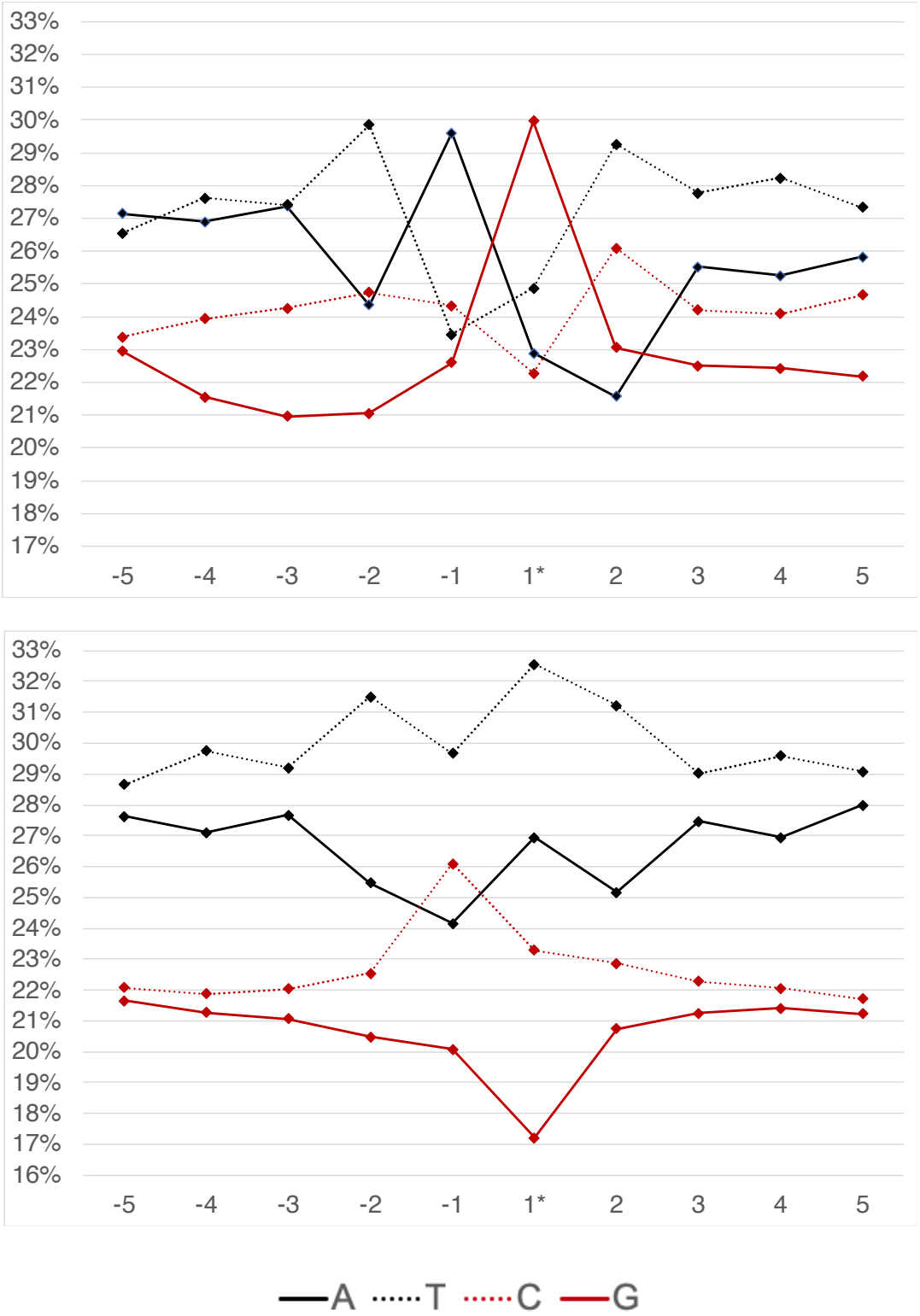


Figura 22: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões 3'UTR em camundongos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma *"baseline frequency"* no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a *"baseline frequency"*.

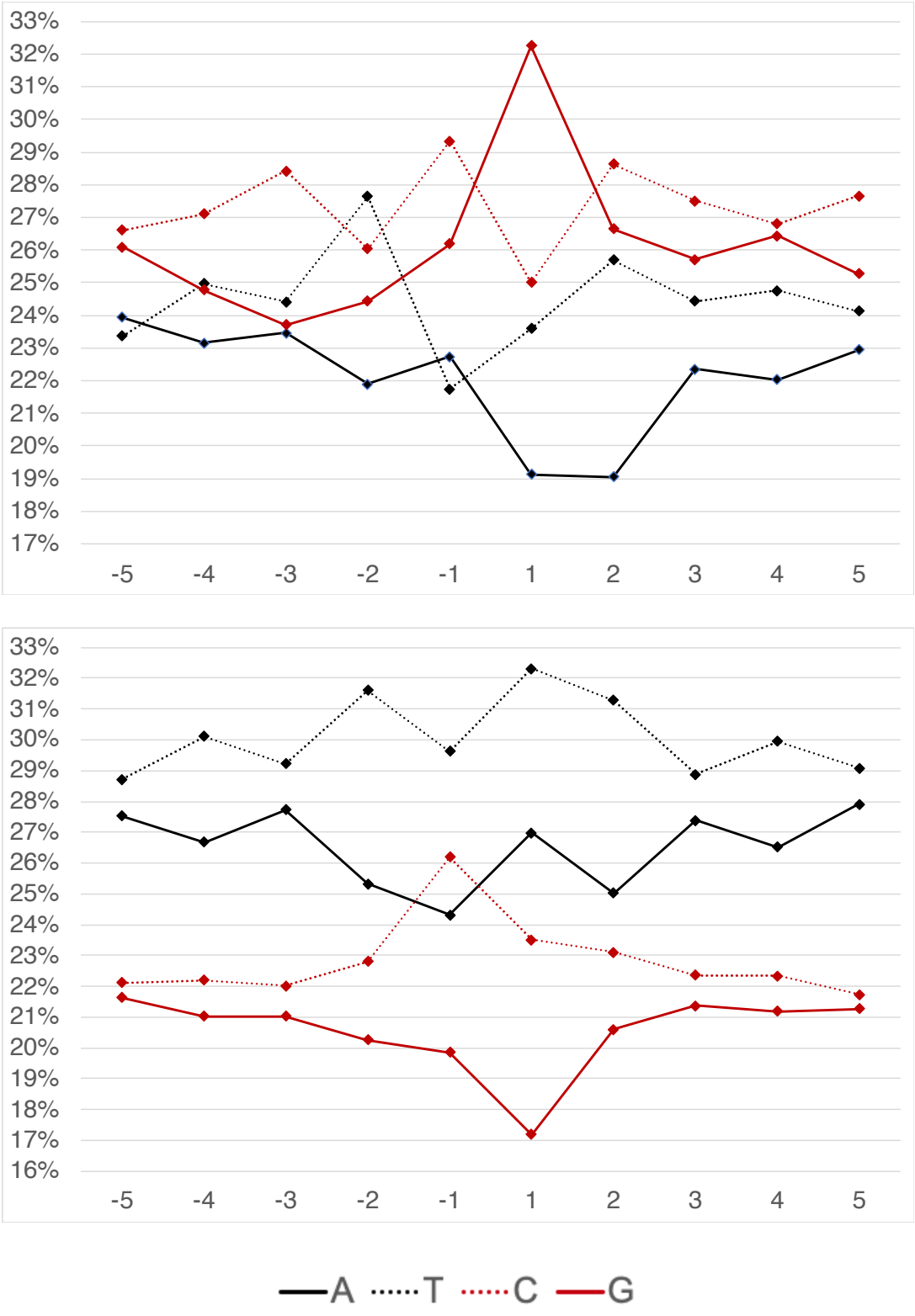


Figura 23: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões 5'UTR em camundongos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma *"baseline frequency"* no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

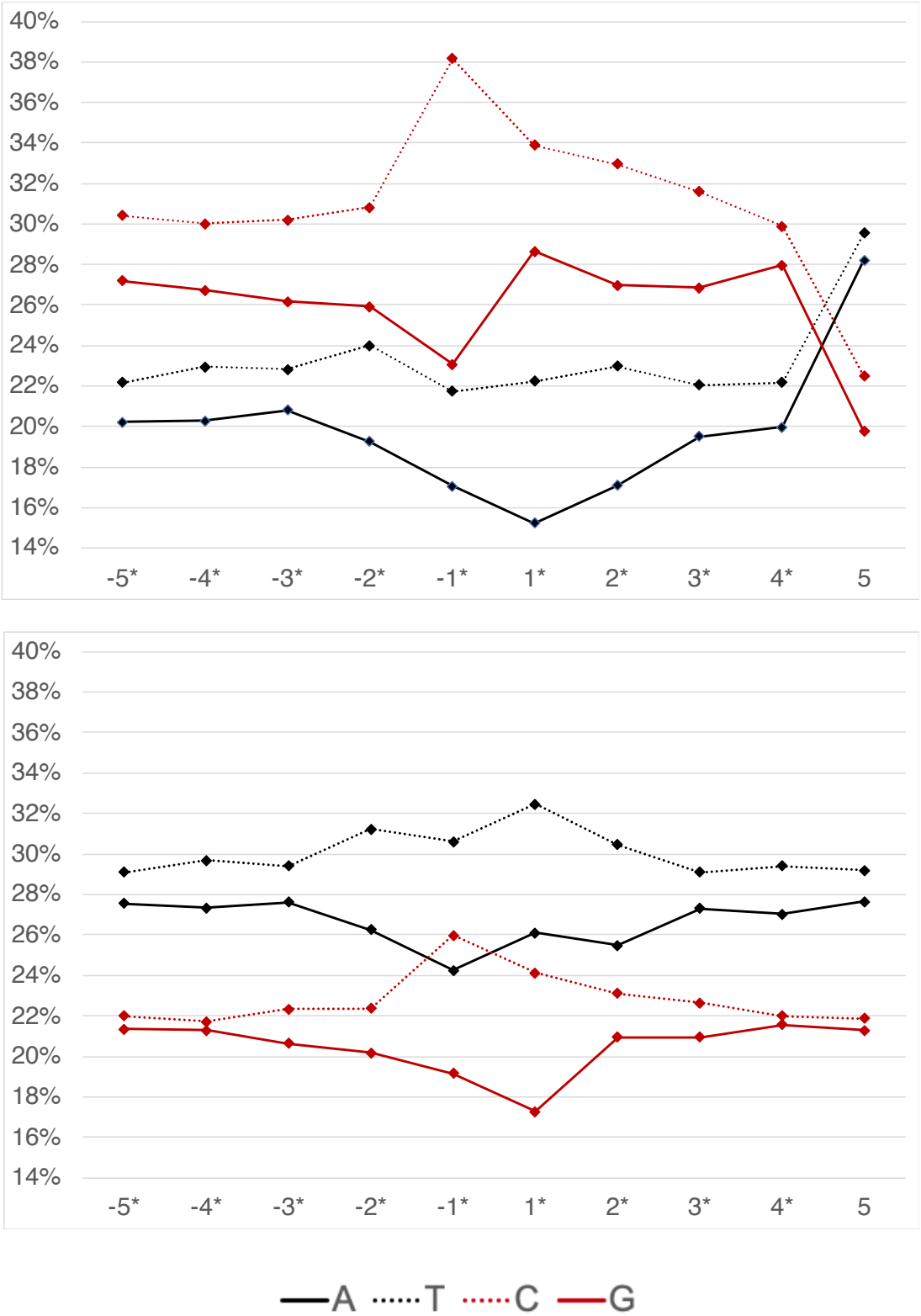


Figura 24: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões 5'UTR em bovinos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma *"baseline frequency"* no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a *"baseline frequency"*.

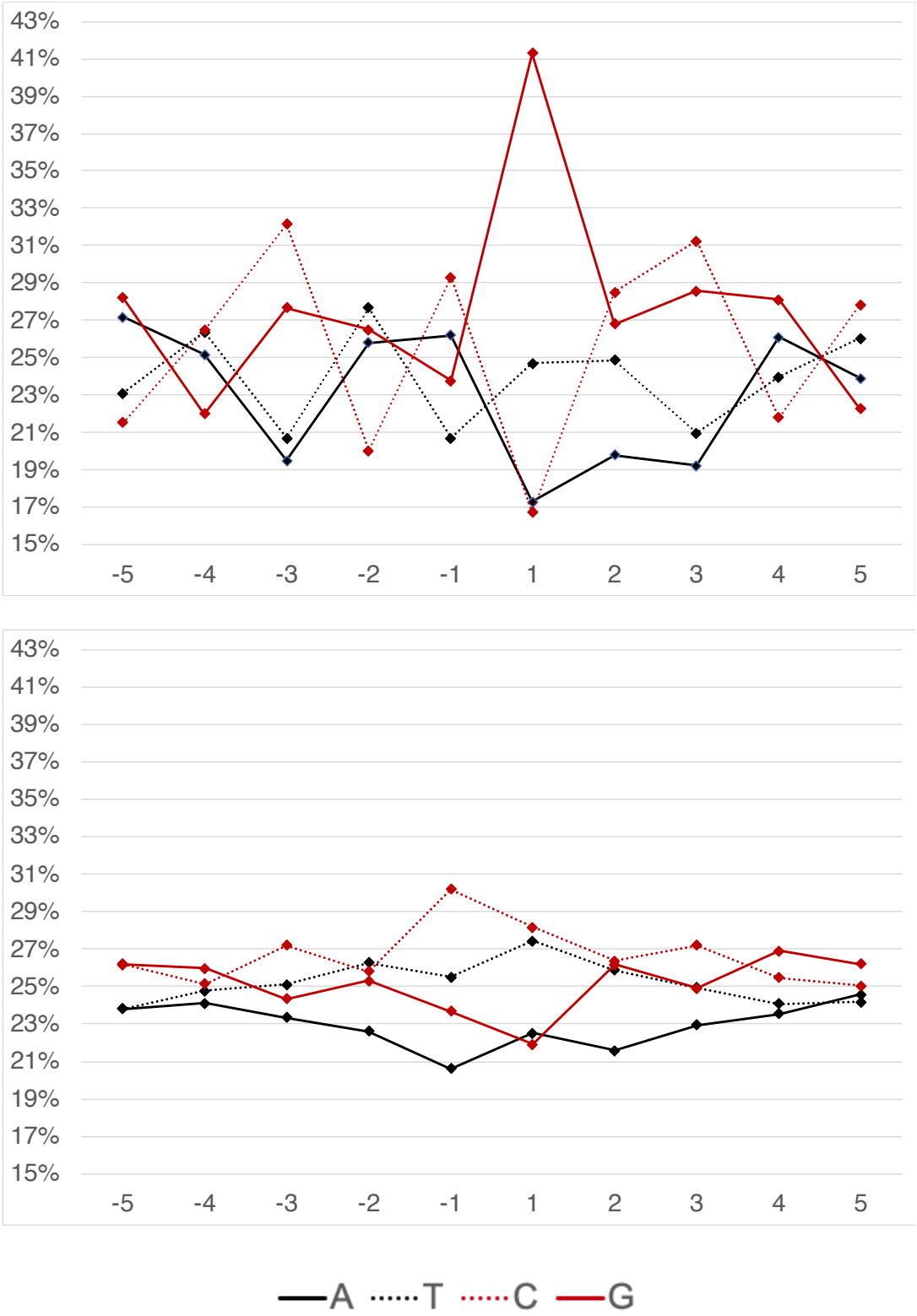


Figura 25: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões sinônimas em camundongos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma *"baseline frequency"* no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

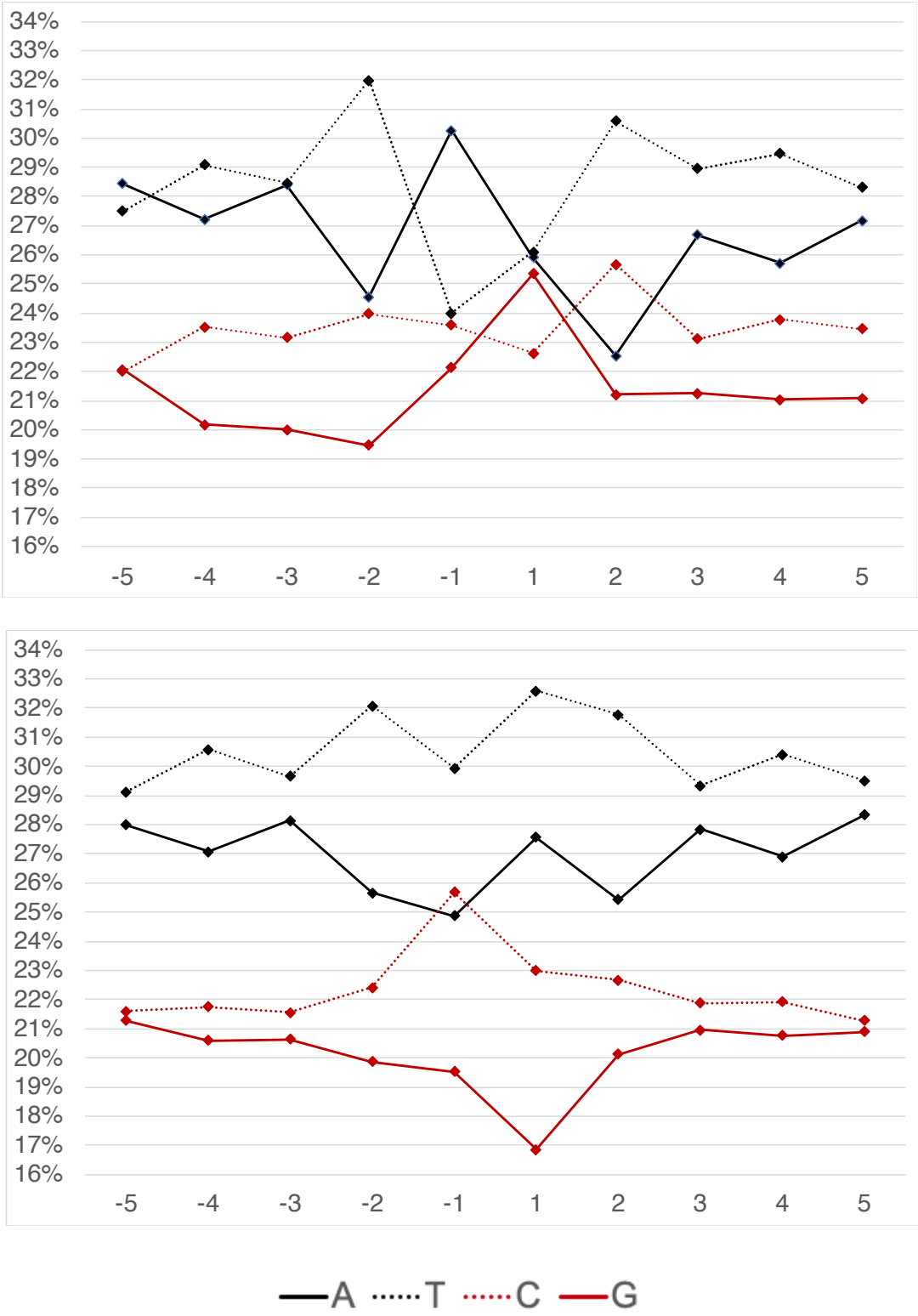


Figura 26: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões intrônicas em camundongos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma *"baseline frequency"* no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

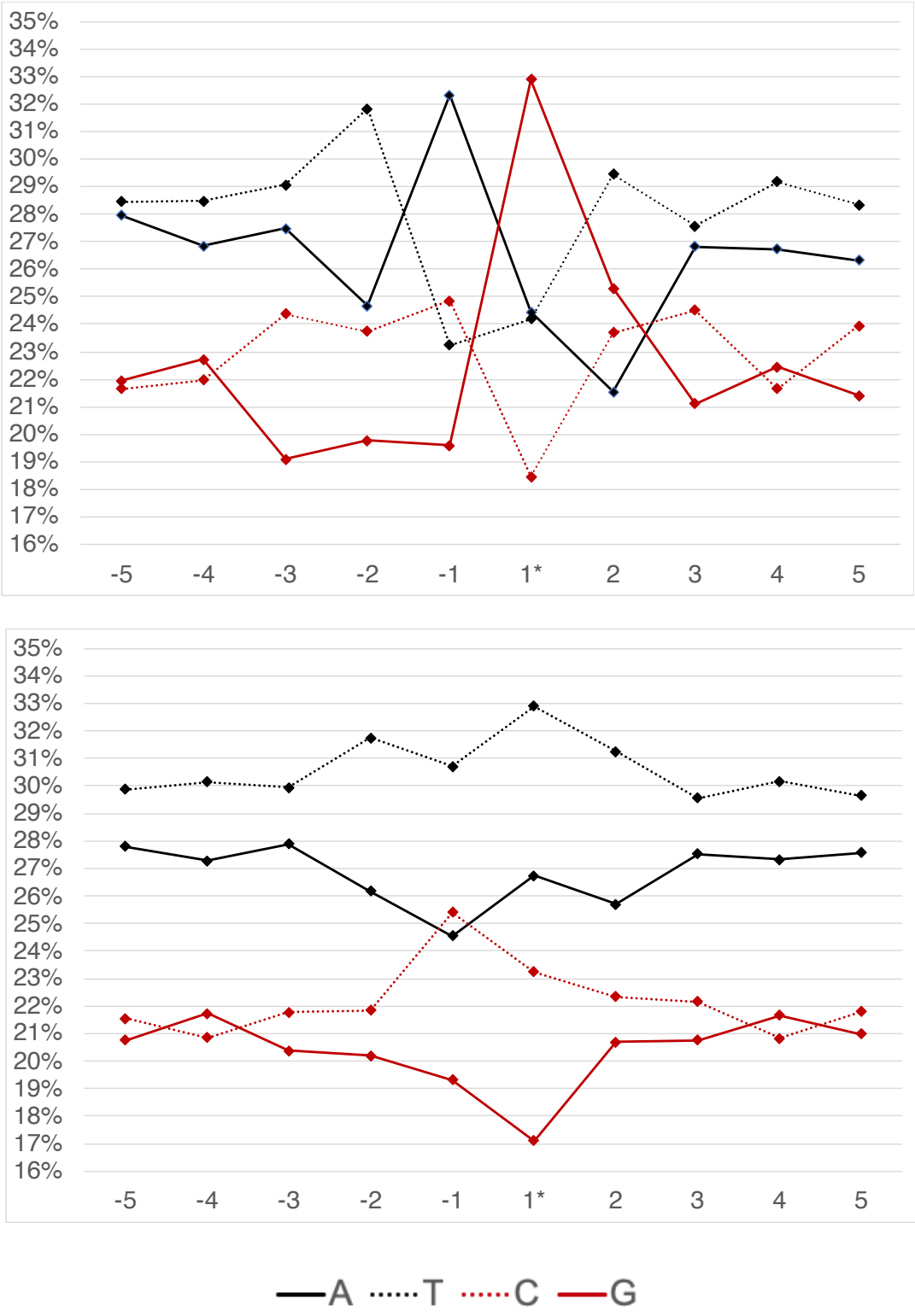


Figura 27: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões intrônicas em *Sus scrofa*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma *"baseline frequency"* no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a *"baseline frequency"*.

conteúdo AT. Enquanto em R (nos gráficos inferiores das Figuras 13-20), Timina tem uma baixa frequência relativa em +1, a Adenina complementar em Y, apresenta baixa frequência relativa na posição -1 - como pode ser visto nos gráficos inferiores das Figuras 22 a 27. Da mesma forma, a menor frequência relativa de Timina na posição +1 em Y coincide com a menor frequência relativa de Adenina em R na posição -1, a alta frequência de Citosina na posição -1 de Y coincide com a alta frequência de Guanina na posição +1 em R e, a baixa frequência de Citosina em R na posição -1 coincide com o ponto mais baixo de Guanina em Y que se encontra na posição +1.

A frequência observada apresenta divergências conforme a região genômica em que está presente, mas em geral, podemos observar o espelhamento entre os gráficos das substituições R e Y de um mesmo organismo para a mesma região.

4.3.2.1 Frequência de base na vizinhança de Y em regiões 3'UTR

Na região de 3'UTR de *Rattus norvegicus*, *Homo sapiens* e *Mus musculus* (Figuras suplementares 20 e 16; e Figura 22 respectivamente) apresentaram diferença significativa entre a frequência observada e a frequência basal do genoma na posição +1. Guanina, nos cinco mamíferos analisados, apresentou a maior frequência relativa na posição +1 e a menor na posição -1 - a exceção de *M. musculus* em que os menores valores de foram encontrados nas posições -4, -3 e -2. Esse organismo também se divergiu quanto a posição do valor máximo de frequência relativa de Citosina, enquanto nos demais organismos foi em -1, em camundongos, foi na posição +2.

Para o menor valor de Citosina, o organismo que apresentou divergência dos demais, foi *B. taurus* (Figura suplementar 13). Ao passo que nos demais esse valor mínimo fora encontrado na posição +1, em bovinos fora na posição +5. Da mesma forma, enquanto nesse organismo o maior valor para a frequência relativa de Adenina se encontra na posição -3, nos demais organismos fora na posição -1. A maior frequência relativa de Timina na posição -2 foi compartilhada por todos os 5 organismos e a menor na posição +1 para bovinos e humanos e, na posição -1 para *Sus scrofa* (Figura suplementar 24), ratos e camundongos.

4.3.2.2 Frequência de base na vizinhança de Y em regiões 5'UTR

Na região de 5'UTR há um maior conteúdo GC do que o AT nas regiões vizinhas aos SNPs nos cinco organismos analisados, ainda que, de acordo com a frequência basal no genoma

fosse esperado um maior conteúdo AT. Nessa região, há um destaque para a distribuição dos nucleotídeos em *Bos taurus* (Figura 24), que apresenta diferenças significativas da posição -5 até a +4, com um forte viés na posição +5, que apresenta os maiores valores para Adenina e Timina e, os menores valores para Citosina e Guanina. Supomos que esses resultados aberrantes em bovinos sejam em função de algum viés nas sequências depositadas de bovinos, uma vez que esse organismo apresentou um desvio das frequências esperadas de transições e transversões em relação ao modelo de Kimura e em relação aos demais organismos. Em ratos (Figura suplementar 21) e camundongos (Figura 23) temos na posição -1 uma diferença significativa entre a frequência basal do genoma e a frequência observada ao redor do SNP. A exceção dos bovinos, nos demais mamíferos analisados, notam-se muitas semelhanças nas posições em que ocorrem os valores máximos e mínimos de frequência relativa. Os maiores valores para Citosina se encontram na posição -1, para Guanina na posição +1 e para Timina na posição -2. Por conseguinte, os menores valores para Adenina estão na posição +2, para Citosina em +1 e, para Timina, em -1.

4.3.2.3 Frequência de base na vizinhança de Y em regiões CDS-Sinônimas

Nas regiões CDS-sinônimas, temos um destaque para o valor máximo de Guanina na posição +1 nos 5 organismos analisados, esse valor, se mostra mais pronunciado dos demais em camundongos (Figura 25), ratos e *Sus scrofa* (Figuras suplementares 22 e 26 respectivamente) do que em *B. taurus* e humanos (Figura suplementares 14 e 18) - que possuem outros valores de frequência relativa próximos ao valor máximo desse nucleotídeo. As diferenças significativas entre a frequência basal do genoma e a frequência ao redor da substituição foram encontradas para a posição +1 em bovinos e *S. scrofa*. A exceção de humanos, observamos uma tendência para os menores valores de frequência relativa de Adenina e Citosina em torno da posição +1. Em bovinos, *H. sapiens* e camundongos, as posições -3, +1 e +3 se sobressaem contendo os valores de frequência relativa máxima para Citosina e, mínimo para Timina (que, nesses organismos apresenta os valores máximos em torno das posições -4 e -2).

4.3.2.4 Frequência de base na vizinhança de Y em regiões intrônicas

Nas regiões intrônicas, os valores máximos e mínimos para Adenina e Timina se comportam de forma semelhante nos organismos, tendo respectivamente os valores máximos

nas posições -1 e -2 e mínimos nas posições +2 e -1. A posição da frequência máxima de Guanina também é comum aos cinco mamíferos em +1. *Rattus norvegicus* (Figura suplementar 23) e *Sus scrofa* (Figura 27), apresentam na posição +1 diferença significativa entre a frequência observada em torno do SNP e a frequência basal, nesses organismos, é interessante notar que os valores das frequências máximas de Adenina e Guanina são bem próximos, o que não ocorre nos demais organismos. Em bovinos, também observamos uma diferença significativa, porém na posição -1. Excetuando camundongos, a Citosina apresentou o valor mínimo de frequência relativa na posição +1.

4.3.3 Frequência de base na vizinhança de M

Na observação da frequência basal do genoma, novamente, como reportado nas transições, um padrão é comum aos cinco mamíferos analisados e às quatro regiões genômicas - a exceção das regiões codificadoras sinônimas que se diferenciam por apresentarem um conteúdo GC superior ao conteúdo AT. Esse padrão consiste na maior frequência de Adenina na posição +1 e a menor frequência de Timina na posição -1; bem como uma maior frequência de Citosina na posição -1 e uma menor frequência de Guanina na posição +1 como pode ser observado nos gráficos inferiores das (Figuras 28-29 e Figuras Suplementares 27-44 do apêndice).

4.3.3.1 Frequência de base na vizinhança de M em regiões 3'UTR

Quando tratamos das frequências ao redor da região do SNP, em 3'UTR, observamos em *Homo sapiens, Mus musculus e Rattus norvegicus* (Figuras Suplementares 31.34 e 38) os valores máximos e mínimos coincidentes com o esperado. Em *Bos taurus* (Figura suplementar 27), temos uma diferença significativa entre observado e esperado na posição +1 e apresenta curvas com grandes discrepâncias nas frequências mínimas observadas. Juntamente com *Mus musculus*, esses dois organismos não apresentam um padrão nítido de vizinhança para Adenina, sendo que nos bovinos, os demais nucleotídeos seguem o padrão da frequência basal do genoma, enquanto que, em camundongos, observamos esse padrão apenas em Timina, tendo citosina sua menor frequência relativa na posição +2 e Guanina sua menor frequência relativa na posição -3.

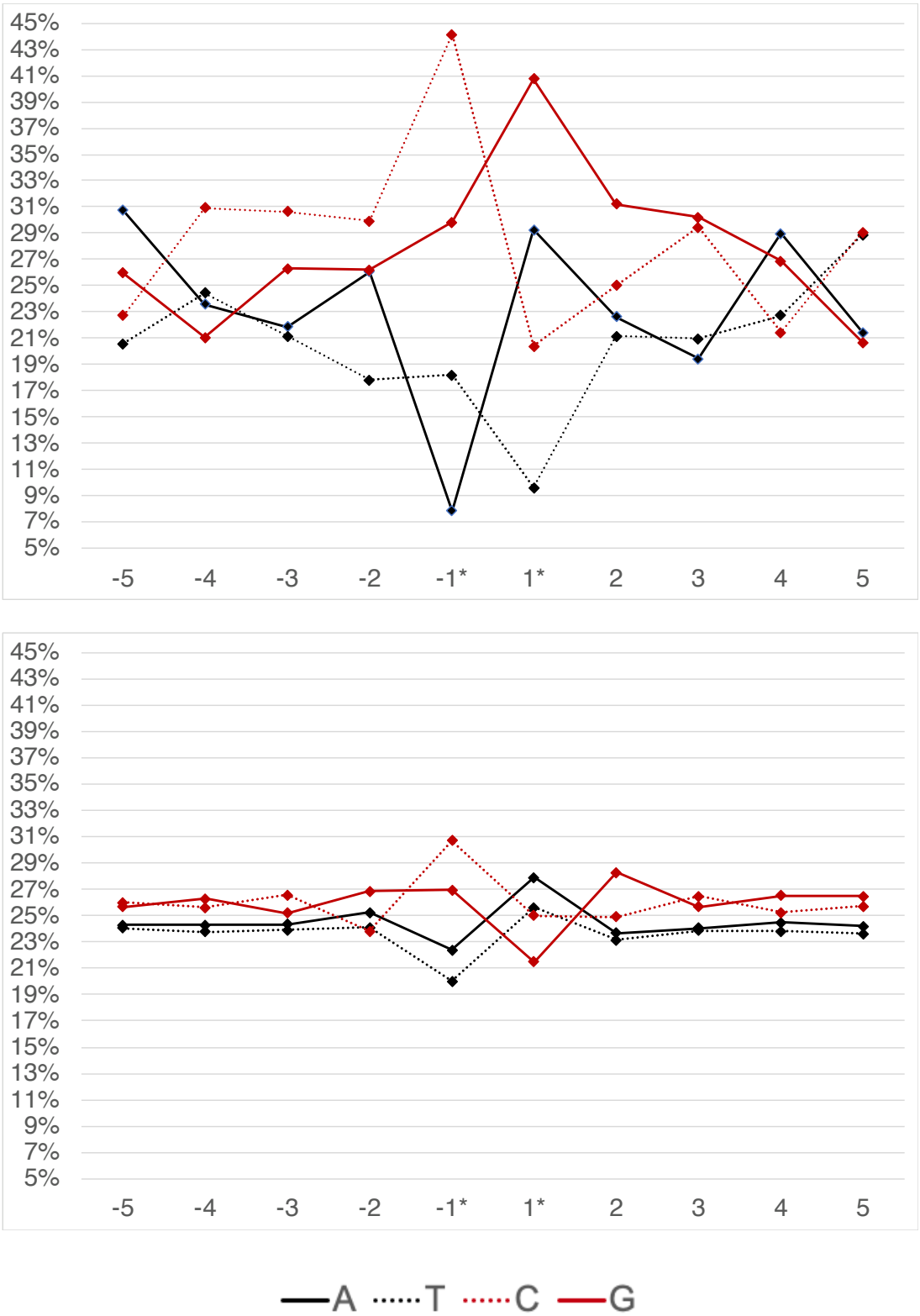


Figura 28: Ocorrência de bases nas proximidades do SNP M (A/C) e a frequência basal esperada no genoma, nas regiões sinônimas em *Rattus norvegicus*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma *"baseline frequency"* no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a *"baseline frequency"*.

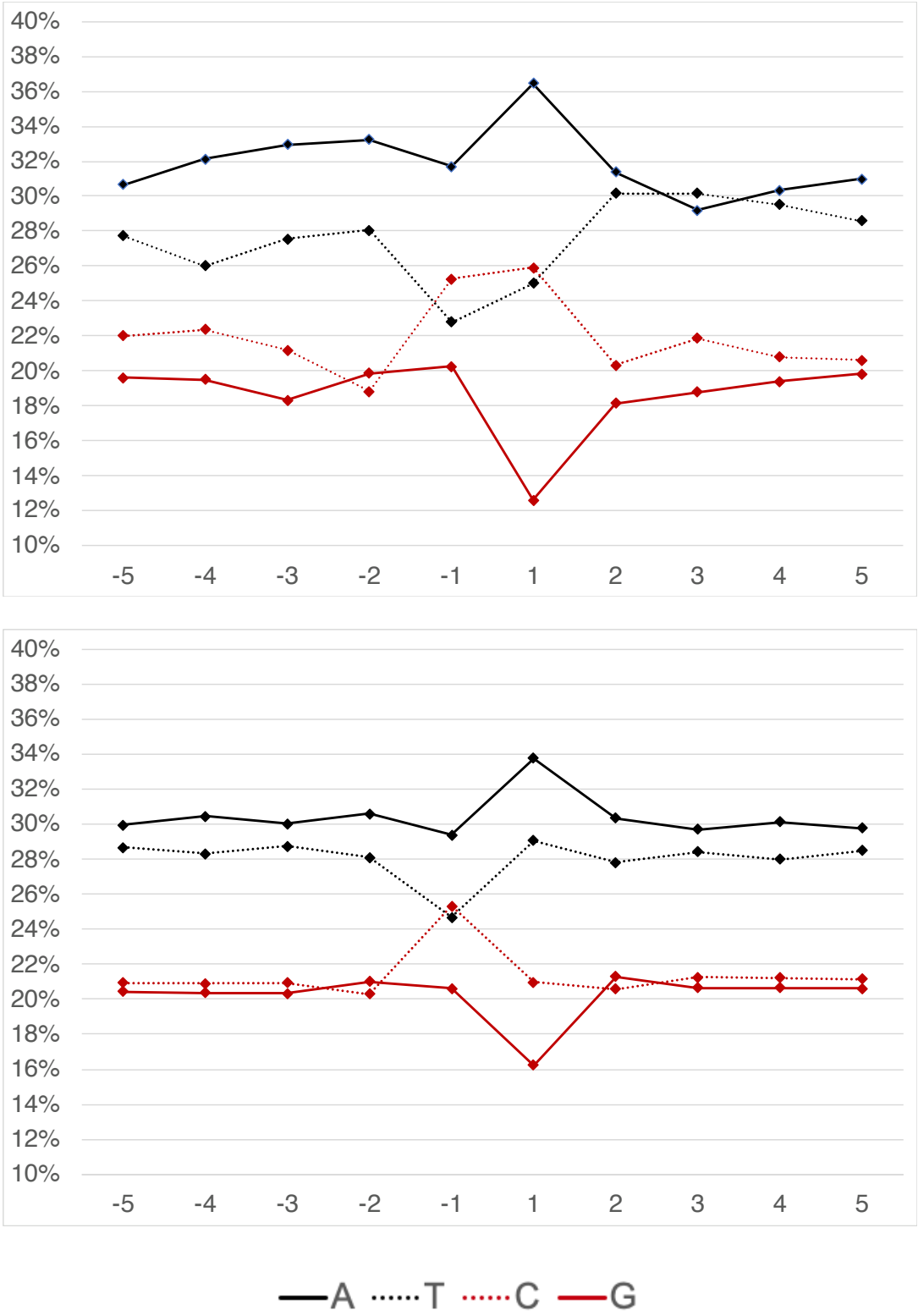


Figura 29: Ocorrência de bases nas proximidades do SNP M (A/C) e a frequência basal esperada no genoma, nas regiões intrônicas em humanos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma *"baseline frequency"* no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

4.3.3.2 Frequência de base na vizinhança de M em regiões 5'UTR

Nas regiões de 5'UTR em que há uma mutação, observamos que o conteúdo GC é maior que o esperado se comparado a frequência basal do genoma em *Mus musculus, Homo sapiens e Bos taurus* - conforme observado nos gráficos superiores das (Figuras suplementares 35,32 e 28). Em bovinos, assim como também ocorreu na região 5'UTR da substituição Y (Figuras 24), apresentou um forte viés na posição +5, onde se encontram as frequências relativas máximas para Adenina e Timina e as mínimas para Citosina e Guanina. Além disso, da posição -5 a +4, todas apresentaram diferença significativa entre a frequência ao redor do SNP e a frequência basal. Os cinco mamíferos compartilham a frequência relativa mínima de Timina, mas se distinguem quanto às frequências máximas e mínimas dos outros nucleotídeos. Em *Homo sapiens*, apesar do gráfico ser visualmente distinto da frequência basal, não houve diferença significativa no teste de qui-quadrado; Adenina apresenta os maiores valores de frequência máxima em -2 e +1, e, mínimo em +3; Citosina apresenta frequência mínima em -2 e Guanina apresenta maior frequência relativa na posição -1 - valor máximo que também é observado em *Rattus norvegicus* e *Sus scrofa*. Este último, apresenta a distribuição das frequências relativas visualmente mais próximo do que seria a frequência basal do genoma - conforme observamos na (Figura suplementar 42), com Adenina tendo os valores máximos e mínimos nas posições +1 e -1 respectivamente e Citosina um valor máximo na posição -1. Em ratos, (Figura suplementar 39), observamos também essa frequência relativa máxima de Citosina na posição -1 e, a maior frequência relativa de Adenina na posição -2.

4.3.3.3 Frequência de base na vizinhança de M em regiões CDS-Sinônimas

Quando analisamos a substituição M numa região codificadora sinônima, temos divergências entre essa frequência e a frequência basal do genoma em bovinos (nas posições -3, -1, +1 e +3 (Figura suplementar 29), em ratos e camundongos (ambos nas posições -1 e +1 Figura 28 e Figura suplementar 36). Ainda que *Sus scrofa* (Figura suplementar 43) não apresente diferenças significativas entre o esperado e o observado, sua distribuição das frequências máximas e mínimas é coincidente com a de *M. musculus* e *R. norvegicus* apresentando os valores máximos de Adenina em -2 e +1, de Citosina em -1 e de Guanina em +1. Para os valores mínimos, nesses organismos, temos a menor frequência de Adenina em -1, de Timina e Citosina em +1 e, de Guanina nas posições -4 e +5. A maior frequência de Citosina em -1 e a menor de Timina em +1 são também presentes em *B. taurus* e *H. sapiens* - Figura

suplementar 33 - que compartilham também a frequência relativa máxima de Adenina em +1 e, menor em -1.

4.3.3.4 Frequência de base na vizinhança de M em regiões intrônicas

A frequência relativa em torno do SNP nas regiões intrônicas não apresenta nenhum caso de divergência significativa entre a frequência esperada e a observada; nos cinco organismos o valor mínimo para Timina se encontra na posição -1 e para Guanina em +1. Em humanos, Figura 29, o padrão das curvas é o mais próximo visualmente da frequência basal e, temos o maior valor de Adenina em +1 - que pode também ser visto em *B. taurus* e *S. scrofa* nas Figuras suplementares 30 e 44 respectivamente - e um maior valor para Citosina em +1 - ainda que o valor da posição -1 seja bem próximo. Nos demais organismos, o menor valor de Citosina está na própria posição -1. *Mus musculus* e *Rattus norvegicus* - Figuras Suplementares 37 e 40 - apresentam pouca definição nas curvas referentes a base Adenina, apresentando frequências máximas em -3 e -4, -3 e -2, respectivamente.

4.3.4 Frequência de base na vizinhança de K

Do mesmo modo que observamos uma complementaridade entre as substituições R e Y, temos uma complementaridade entre as substituições K e M conforme pode ser observado na Figura 30. Em função disso, nota-se - pronunciadamente na frequência basal do genoma - um espelhamento entre os gráficos de frequências de bases dessas duas substituições. Assim nas demais substituições analisadas, a frequência basal apresenta um modelo comum entre os organismos analisados sem grandes divergências entre as regiões. Nesse modelo comum para o esperado - que pode ser observado nos gráficos inferiores das Figura 31 e Figuras suplementares 45-63 - temos na posição +1 o maior valor para a frequência relativa de Guanina e o valor mínimo para Adenina. Por sua vez, na posição -1 temos a maior frequência relativa de Timina e a menor de Citosina.

4.3.4.1 Frequência de base na vizinhança de K em regiões 3'UTR

Para a frequência ao redor de um SNP nas regiões de 3'UTR, temos em *Sus scrofa* (Figura suplementar 60) os valores máximos e mínimos coincidentes com a frequência basal do genoma.

Nos demais organismos, essa concomitância é observada para Adenina e Citosina. Em *H. sapiens* e *B. taurus* (Figuras suplementares 45 e 49 respectivamente), nota-se um maior valor para Timina na posição -1 e, para Guanina, temos nas posições -1 e +2 os valores máximos e mínimos respectivamente. *R. norvegicus* e *M. musculus* (figuras suplementares 52 e 56) apresentam curvas "flat" para Timina que, apesar de serem visualmente diferentes das curvas da frequência basal não houve diferença significativa entre as mesmas de acordo com o teste de qui-quadrado.

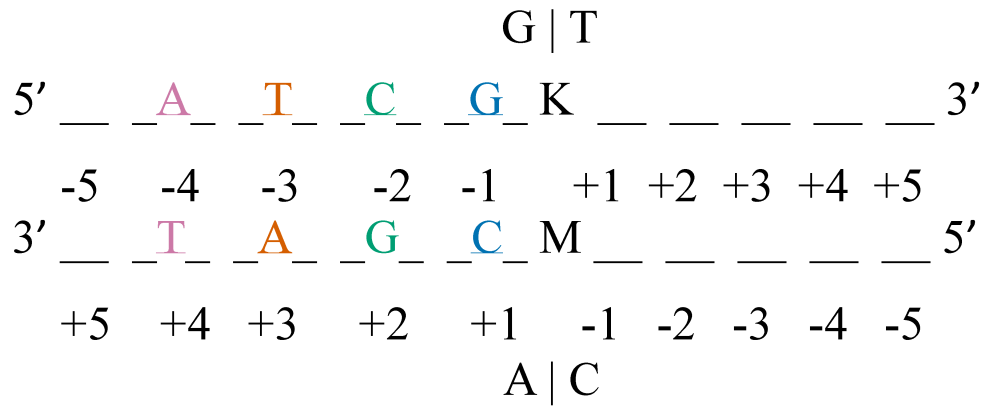


Figura 30: Representação dos locais vizinhos em transversões complementares. Em função da complementariedade de bases, temos um espelhamento nos resultados em função da simetria das fitas, de tal forma que, por exemplo, se tivermos uma grande frequência relativa de Guanina na posição -1 em uma substituição K, é esperado que tenhamos essa mesma alta frequência relativa em Citosina na posição +1 em uma substituição M.

4.3.4.2 Frequência de base na vizinhança de K em regiões 5'UTR

Nas regiões de 5'UTR, em bovinos - (Figura suplementar 46) nota-se bem como nas substituições M e também na Y, um viés na posição 5, apresentando os valores de frequência máxima para Adenina e Timina e mínimos para Citosina e Guanina, além de diferença significativa entre a frequência basal do genoma e a frequência em torno do SNP da posição -5 até a posição 4. Em ratos, os valores máximos e mínimos de Guanina foram nas posições -1 e -3 e em camundongos, nas posições 1 e -2. *Bos taurus, Homo sapiens* e *Mus musculus* (Figuras suplementares 46, 50 e 53), nesta região, apresentaram um maior conteúdo GC do que o esperado. *Mus musculus e Rattus norvegicus* (Figuras suplementares 53 e 56) apresentaram diferença significativa da frequência ao redor do SNP para com a frequência basal na posição +1.

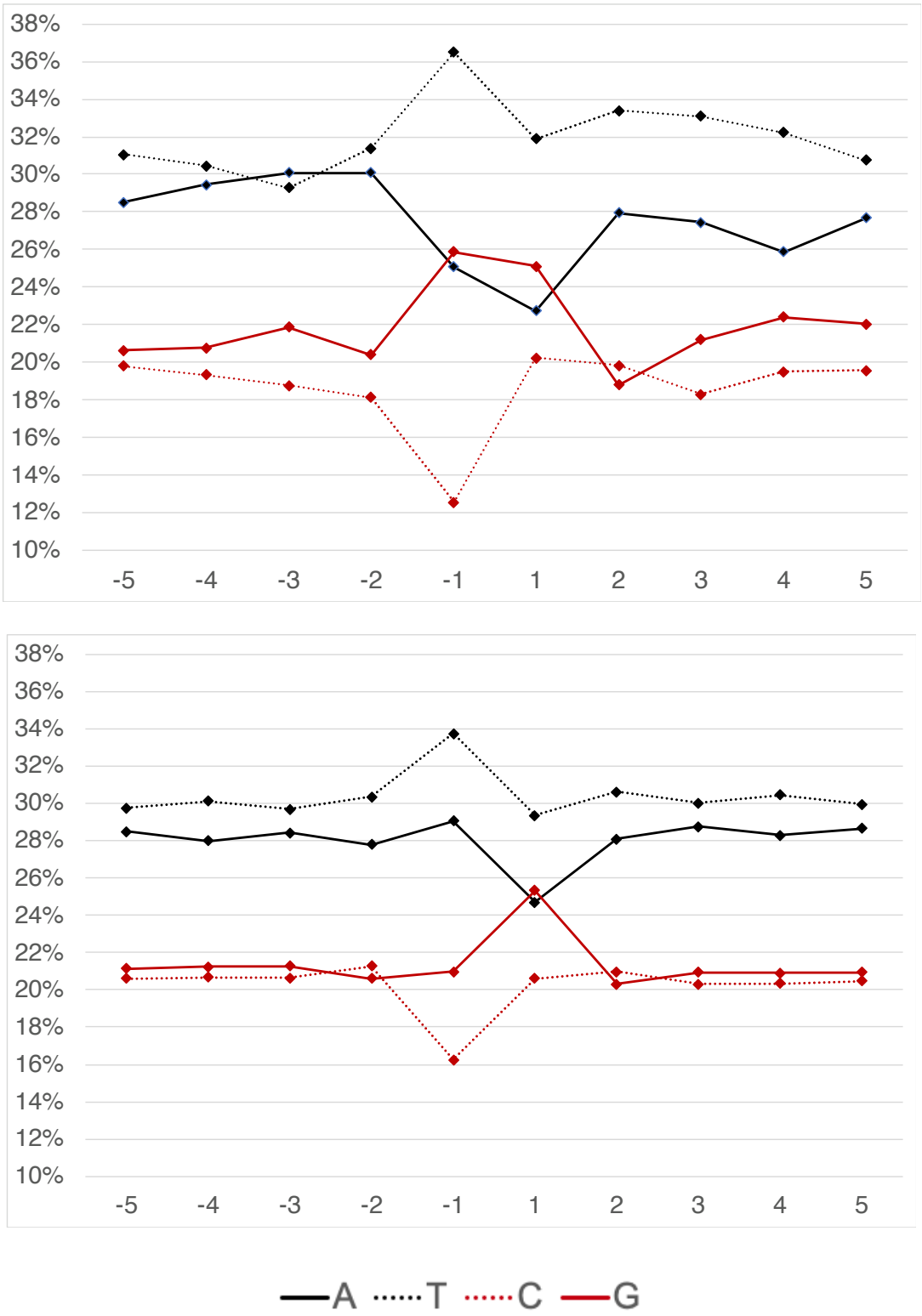


Figura 31: Ocorrência de bases nas proximidades do SNP K (G/T) e a frequência basal esperada no genoma, nas regiões intrônicas em humanos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma *"baseline frequency"* no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

Sus scrofa, por sua vez, fora o organismo que o gráfico das frequências observadas próximas à mutação é visualmente mais próximo do gráfico das frequências simuladas - conforme é possível observar na (Figura suplementar 60).

4.3.4.3 Frequência de base na vizinhança de K em regiões CDS-Sinônimas

Quando a substituição K se encontra numa região codificadora e gera uma mutação sinônima, temos que, para *Mus musculus* e *Rattus norvegicus* há uma diferença significativa das frequências observadas para com frequência basal do genoma nas posições imediatamente adjacentes (-1 e +1, Figuras suplementares 54 e 58). *Bos taurus* (Figura suplementar 47) além de apresentar diferenças significativas nessas posições, apresentou também nas posições -2 e +3. Mesmo com essas diferenças significativas, é possível notar alguns padrões comuns aos cinco organismos - que podem ser vistos nos gráficos superiores das Figuras suplementares 47, 51, 54, 58 e 62 - como valor mínimo para a frequência de Adenina na posição -1, Timina apresentando o maior valor de frequência relativa na posição -1 e o menor na posição +1, Citosina apresentando os menores valores ora na posição -5 ora na posição +4 e, Guanina apresentando um pico de maior valor de frequência relativa na posição 1 - conforme a frequência basal do genoma.

4.3.4.4 Frequência de base na vizinhança de K em regiões intrônicas

Nas regiões intrônicas, *Bos taurus* foi o único organismo que apresentou uma diferença significativa entre a frequência em torno do SNP e a frequência basal do genoma - na posição 1 - no entanto, os valores máximos e mínimos se encontravam em posições coincidentes, o que sugere que essa diferença possa ser em função do maior conteúdo de Guanina que esse organismo apresentou. *Sus scrofa* (Figura Suplementar 63) também apresentou os valores máximos e mínimos coincidentes entre esperado e observado. Em *H. sapiens* Figura 31, embora haja uma alta frequência de Guanina na posição +1, o valor máximo se deu na posição -1. Em ratos e camundongos Figura Suplementares 55 e 59, a curva sem muitos picos de Adenina na substituição M reflete na curva "flat" de Timina. Os valores mínimos de Citosina, Adenina e o valor máximo de Guanina se encontram nas mesmas posições da frequência basal.

4.3.5 Frequência de base na vizinhança de W

Na substituição W, temos a particularidade da auto complementaridade das sequências, como pode ser observado na Figura 32, o que resulta em um espelhamento no próprio gráfico entre as bases anteriores ao SNP - as que são marcadas com indicadores negativos e estão mais próximas da região de 5'- e as bases posteriores ao SNP - marcadas com indicadores positivos e mais próximas da região de 3'.

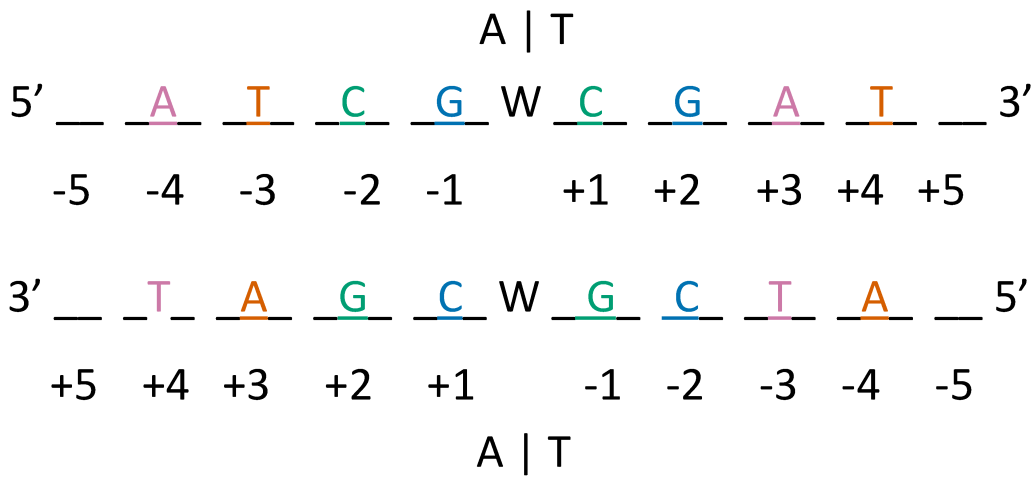


Figura 32: Representação dos locais vizinhos em mutações autocomplementares. Em função da complementariedade de bases, temos um espelhamento nos resultados em função da simetria das fitas, de tal forma que, por exemplo, se tivermos uma grande frequência relativa de Guanina na posição -1 em uma substituição W, é esperado que tenhamos essa mesma alta frequência relativa em Citosina na posição +1.

Notam-se duas tendências distintas para a frequência dos nucleotídeos vizinhos próximos, das regiões não codificadoras (3'UTR, 5'UTR e intrônicas) e as regiões codificadoras sinônimas.

4.3.5.1 Frequência de base na vizinhança de W em regiões não codificadoras

Nessas regiões não codificadoras, a frequência basal do genoma, observado nos gráficos inferiores da Figura 32 e nas Figuras suplementares 64, 65, 67 a 72, 74, 75, 77 a 79 e 81, há um maior conteúdo AT do que o GC, porém, nas adjacências à simulação do SNP, na posição -1, Adenina tem a sua menor frequência relativa, no nucleotídeo Timina, essa menor frequência se dá na posição +1. Por sua vez, as bases Citosina e Guanina, aumentam em sua frequência relativa quando estão imediatamente próximas, nas posições -1 e +1, respectivamente. Na

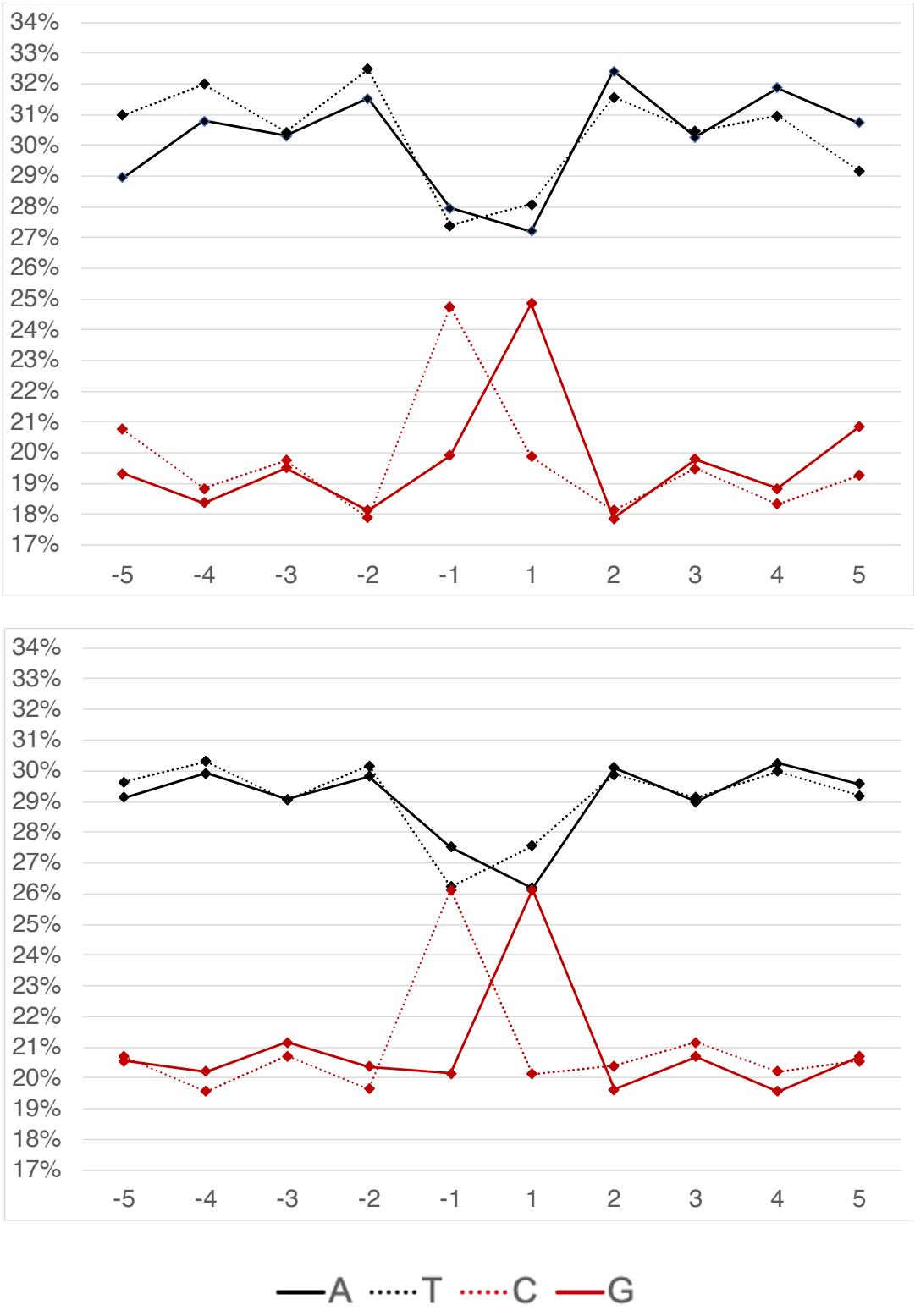


Figura 33: Ocorrência de bases nas proximidades do SNP W (A/T) e a frequência basal esperada no genoma, nas regiões intrônicas em camundongos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma *"baseline frequency"* no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

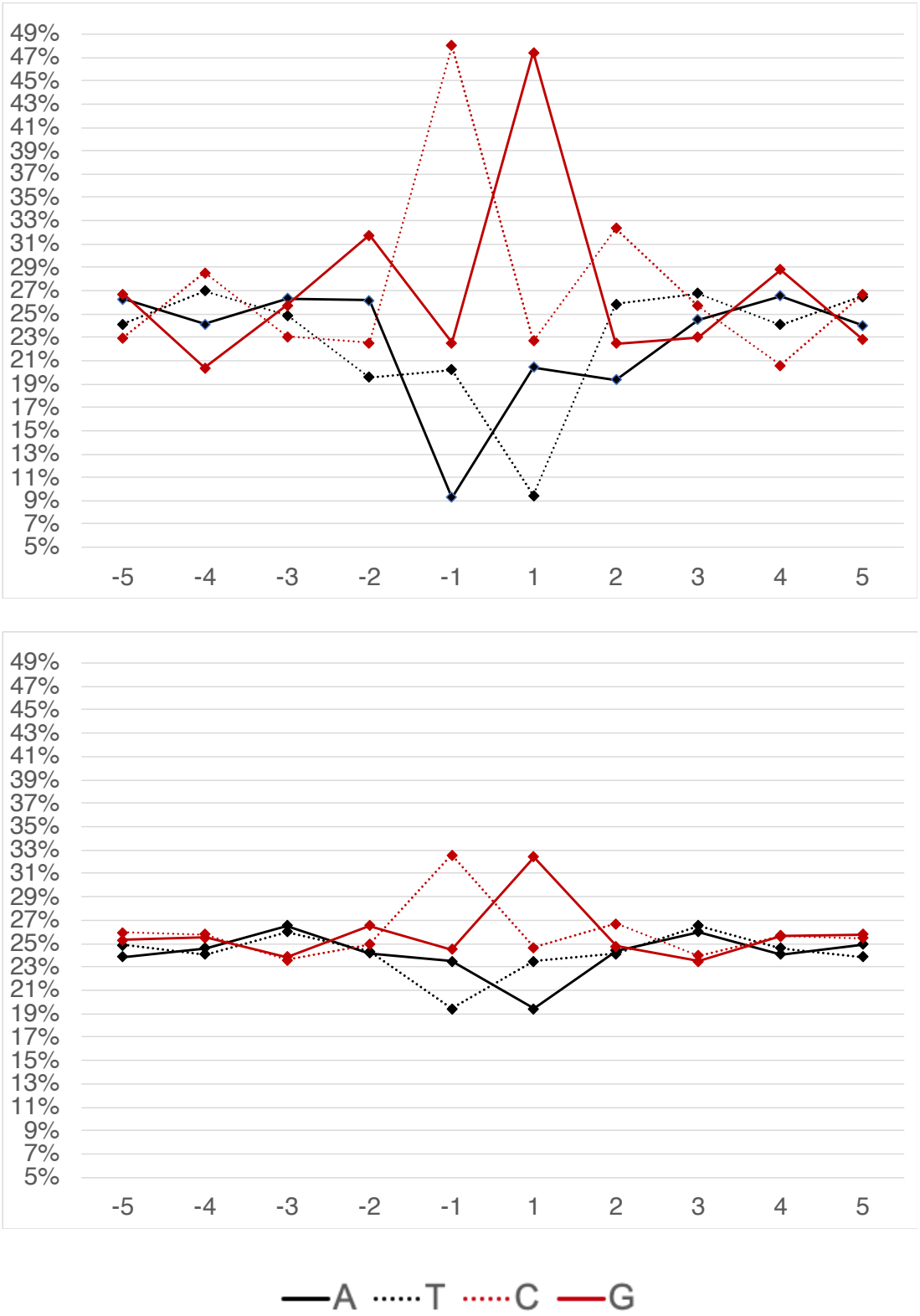


Figura 34: Ocorrência de bases nas proximidades do SNP W (A/T) e a frequência basal esperada no genoma, nas regiões sinônimas em humanos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma *"baseline frequency"* no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

análise das regiões em torno de um SNP as maiores frequências de Citosina e Guanina permanecem nas posições -1 e +1, para os cinco organismos tanto nas regiões 3' e 5'UTR quanto nas regiões intrônicas. Quanto ao conteúdo de Adenina e Timina, esses valores estão espelhados se comparados com o esperado, tendo Adenina uma frequência mínima em -1 e Timina uma frequência mínima em 1 - isso ocorre para *Sus scrofa* nessas três regiões (Figuras suplementares 78, 79, 81), *Rattus norvegicus* nas regiões 3'UTR e intron (Figura suplementares 74 e 77), e *Bos taurus* também nas regiões intrônicas (Figura suplementar 67). É notável também nas regiões 3'UTR e de íntron um padrão em torno das posições +2 e -2, pronunciadamente em bovinos, humanos e camundongos. Nestas posições há um aumento na frequência relativa de Adenina e Timina e uma diminuição na frequência relativa de Citosina e Guanina.

Quando olhamos a ocorrência de bases próximas ao SNP W, há uma divergência na região 5'UTR em comparação com as demais regiões não codificadoras, primeiramente o conteúdo GC tende a ser maior que o AT - diferentemente da frequência basal do genoma - mesmo em organismos como *Sus scrofa* (Figura suplementar 79), que essa discrepância no conteúdo é menor, esse fenômeno ainda é bem nítido nas posições próximos ao SNP. Outro ponto, é a diferença significativa entre os valores esperados e observados que ocorre nas posições -1 e 1 de *Homo sapiens* (Figura suplementar 69), único caso marcado pelo teste de qui-quadrado das regiões não codificadoras. Ainda que em bovinos (Figura suplementar 65) tenhamos novamente um forte viés na posição 5, que apresenta os maiores valores de Adenina e Timina e os menores para Citosina e Guanina, conseguimos ver o desenho das curvas similar aos das curvas de frequência basal do genoma.

4.3.5.2 Frequência de base na vizinhança de W em regiões codificadoras

Nas regiões codificadoras que geram mutações sinônimas, há uma predominância do conteúdo GC em relação ao de AT, tanto na frequência basal do genoma, quanto nos nucleotídeos ao redor do SNP. *Bos taurus, Homo sapiens, Mus musculus e Rattus norvegicus* apresentaram diferença significativa dentre essas duas frequências nas posições -1 e +1, ainda que os valores máximos e mínimos de frequência relativa dos cinco organismos correspondam ao que se é esperado aos das demais regiões.

4.3.6 Frequência de base na vizinhança de S

Na substituição S, observamos o mesmo efeito de auto complementaridade observado na substituição W. A maior frequência relativa - quando olhamos a frequência basal do genoma, nos gráficos inferiores das (Figura 35 e Figuras suplementares 82-100) - de Timina, por exemplo, se encontra na posição +1, e de forma complementar, a maior frequência de Adenina se encontra na posição -1, sendo curvas visualmente espelhadas no gráfico. Ainda em termos do genoma, observamos a menor frequência de Citosina na posição -1 e, de forma correspondente, uma menor frequência relativa de Guanina na posição +1.

4.3.6.1 Frequência de base na vizinhança de S em regiões 3'UTR

Na região de 3'UTR em torno de um SNP, *Bos taurus* é um caso a parte dos demais organismos, como podemos observar na (Figura suplementar 82), primeiramente, por apresentar um conteúdo GC superior ao AT e as maiores frequências de Citosina e Guanina nas posições +1 e -1, respectivamente, e as maiores frequências de Adenina e Timina nas posições +1 e -1, respectivamente. *H. sapiens, M. musculus, R. norvegicus* e *S. scrofa* (Figuras suplementares 86, 90, 93 e 97) apresentam um forte viés nas posições -2 e +2 para Citosina e Guanina - que são as posições em que essas bases apresentam seus valores máximos de frequência relativa. Para Adenina e Timina, os maiores valores de frequência relativa estão nas posições -1 e +1. Em comum, os cinco organismos apresentam os valores mínimos para Citosina e Guanina nas posições -1 e +1 conforme também visto na frequência basal do genoma.

4.3.6.2 Frequência de base na vizinhança de S em regiões 5'UTR

Enquanto na frequência basal do genoma para a região de 5'UTR se espera um conteúdo AT superior ao GC - conforme vistos nos gráficos inferiores das (Figuras suplementares 83, 87, 91, 94 e 98), quando analisamos a vizinhança dos SNPs, nos cinco organismos - conforme vistos nos gráficos superiores dessas mesmas figuras - há uma preponderância de maior conteúdo GC do que o de AT nesta região. Novamente, como nas substituições Y, M e K, em bovinos, temos um forte viés na posição 5, que apresenta os maiores valores para Adenina e Timina e, os menores valores para Citosina e Guanina - e com isso, apresentou diferença significativa de acordo com o teste de qui-quadrado das posições -5 a +4. Camundongos e *Sus scrofa* também apresentaram diferenças estatisticamente significativas, ambos, nas posições -1

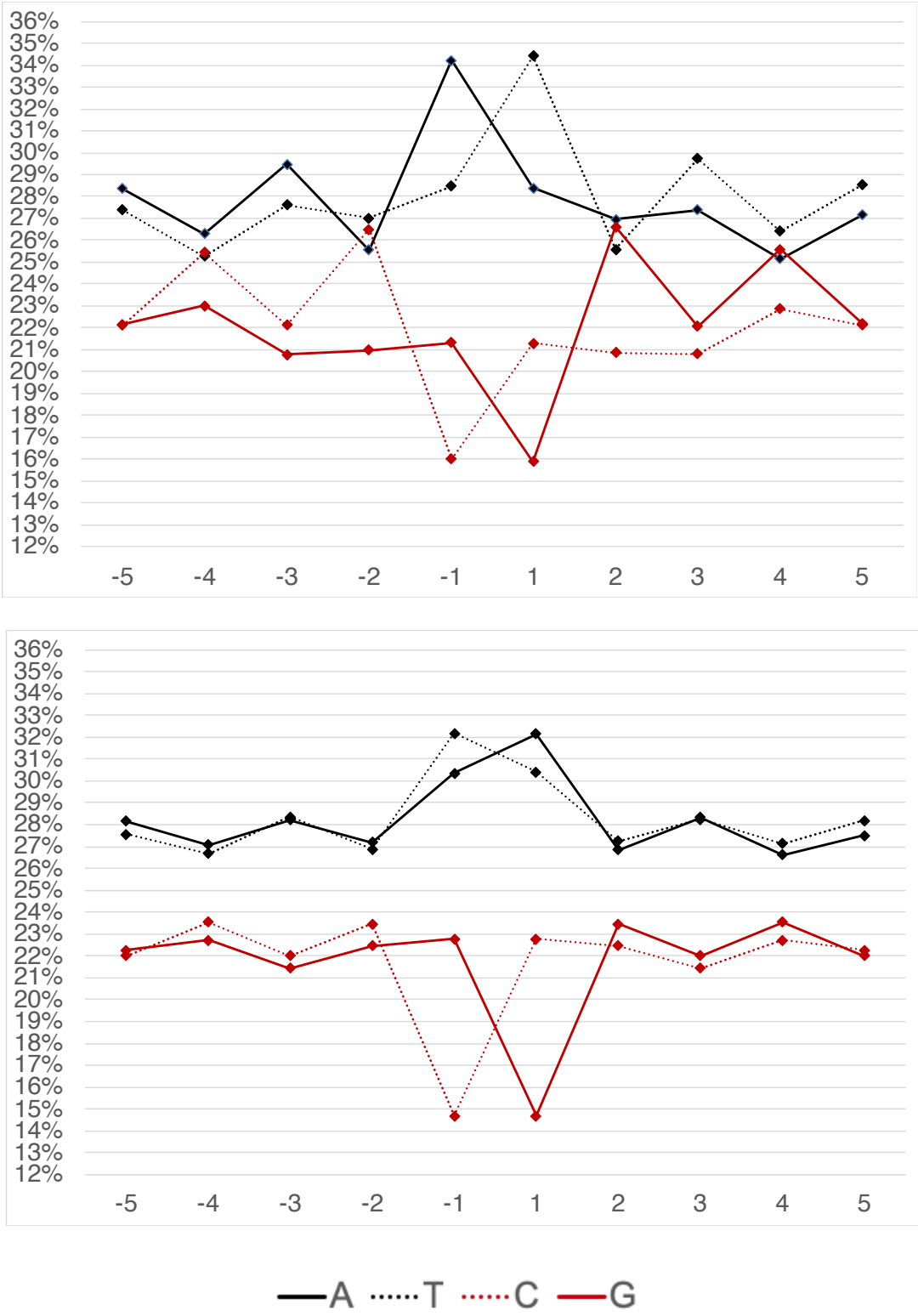


Figura 35: Ocorrência de bases nas proximidades do SNP S (C/G) e a frequência basal esperada no genoma, nas regiões intrônicas em camundongos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma *"baseline frequency"* no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

e +1. Os menores valores para Citosina e Guanina, conforme o esperado na frequência basal do genoma, foram comuns - a exceção de *H. sapiens* -, respectivamente nas posições -1 e 1. Em humanos o valor mínimo de Citosina se deu na posição -1 e, o mínimo para Guanina na posição +1. Em relação aos valores máximos para essas bases que fazem 3 pontes de hidrogênio, esse organismo, juntamente com *S. scrofa* apresentaram nas posições -2 e +2 os maiores valores para Citosina e Guanina. Os valores máximos de Adenina na posição -1 e de Timina na posição +1 foram comuns a humanos, camundongos, ratos e *S. scrofa*.

4.3.6.3 Frequência de base na vizinhança de S em regiões CDS-Sinônimas

Diferentemente das demais regiões, nas regiões codificadoras sinônimas, a frequência basal do genoma possui um conteúdo GC superior ao AT e isso também se observa na vizinhança em torno do SNP - com exceção das adjacências imediatas ao SNP, posições +1 e -1 de humanos (Figura suplementar 88). Nos cinco organismos, os valores máximos de Adenina e Timina seguem a frequência máxima esperada nas posições +1 e -1. Em *Bos taurus* (Figura suplementar 84), o teste de qui-quadrado demonstrou que existe diferença significativa entre a frequência basal do genoma e a frequência observada nas proximidades de um SNP das posições -3 a +1. Nesse organismo observamos também que os valores máximos e mínimos para Citosina e Guanina não estão nas adjacências imediatas da mutação. Para Citosina, a maior frequência relativa observada foi na posição +3 e as menores nas posições +4 e -5. De forma correspondente, a maior frequência de Guanina se deu na posição -3 e as menores nas posições -4 e +5.

Em camundongos e ratos (Figuras suplementares 92 e 95), tivemos diferenças significativas no teste qui-quadrado nas posições -1 e +1 das regiões codificadoras sinônimas. Os valores de frequências máximas e mínimas foram coincidentes com o de *Sus scrofa* (Figura suplementar 99), tendo a frequência mínima para Citosina e Guanina nas posições +1 e -1, conforme a frequência basal do genoma, e máximas em -1 e +1 respectivamente, mas este organismo não apresentou diferença significativa entre a frequência esperada e a frequência observada em torno do SNP. Em *Homo sapiens*, também não tivemos diferença significativa em nenhuma posição, e os valores máximos para Adenina e Timina e, mínimos para Citosina e Guanina, coincidiram com a frequência basal do genoma. O que se observou de distinto dos demais organismos foi um viés em torno das posições -2 e +2, que concentraram os valores máximos de frequência relativa para Citosina e Guanina nesta ordem.

4.3.6.4 Frequência de base na vizinhança de S em regiões intrônicas

Nas regiões intrônicas em torno de um SNP esse viés também pode ser visto para , *M. musculus, H. sapiens, R. norvegicus e S. scrofa* (nos gráficos superiores das Figura 35 e figuras suplementares 89, 96 e 100), juntamente com as demais frequências se dando nas posições conforme a frequência basal do genoma (nos gráficos inferiores). Em bovinos, que pode ser visualizado na (Figura suplementar 85), temos as frequências máximas de Adenina e Timina e, mínima para Citosina e Guanina conforme o esperado e, não houve um viés em torno das posições -2 e +2 para os valores máximos de Citosina e Guanina.

4.4 Análise da frequência relativa de cada dinucleotídeo

Uma vez que apenas 8% das comparações entre a frequência basal do genoma e a frequência em torno do SNP apresentaram uma diferença estatística significativa, nos propomos a investigar a preferência por determinadas bases mais adjacentes a outras no genoma. Como podemos observar nos gráficos das Figuras 36 e 37, temos uma divergência na frequência dos dinucleotídeos conforme região.

Nota-se que os cinco organismos analisados nas quatro regiões possuem frequências dos dinucleotídeos semelhantes. A tendência de maior conteúdo GC em regiões codificadoras sinônimas, observada tanto na frequência basal do genoma quanto nas proximidades de um SNP, é novamente visualizada com uma maior frequência dos dinucleotídeos CpG, CpC, GpC e GpG e, menor frequência dos dinucleotídeos ApT, ApA, TpT e TpA.

A região de 5'UTR apresenta um maior conteúdo GC nas proximidades de um SNP, mas quando analisada em sua totalidade - frequência basal -, ela apresenta frequências relativas de dinucleotídeos como das demais regiões não codificadoras analisadas (3'UTR e intrônicas).

Uma característica que chama a atenção é a baixa frequência do dinucleotídeo CpG - somente nas regiões sinônimas, o que pode ser atribuído aos efeitos de hipermutabilidade deste dinucleotídeo pela desaminação de citosinas metiladas, que contribuem significativamente para o aumento da taxa de transições nos genomas de plantas e animais (DUNCAN; MILLER, 1980 e JIANG et al., 2008), dessa forma, essa metilação resultaria na baixa frequência do dinucleotídeo CpG e em maiores frequências dos nucleotídeos TpG, CpA.

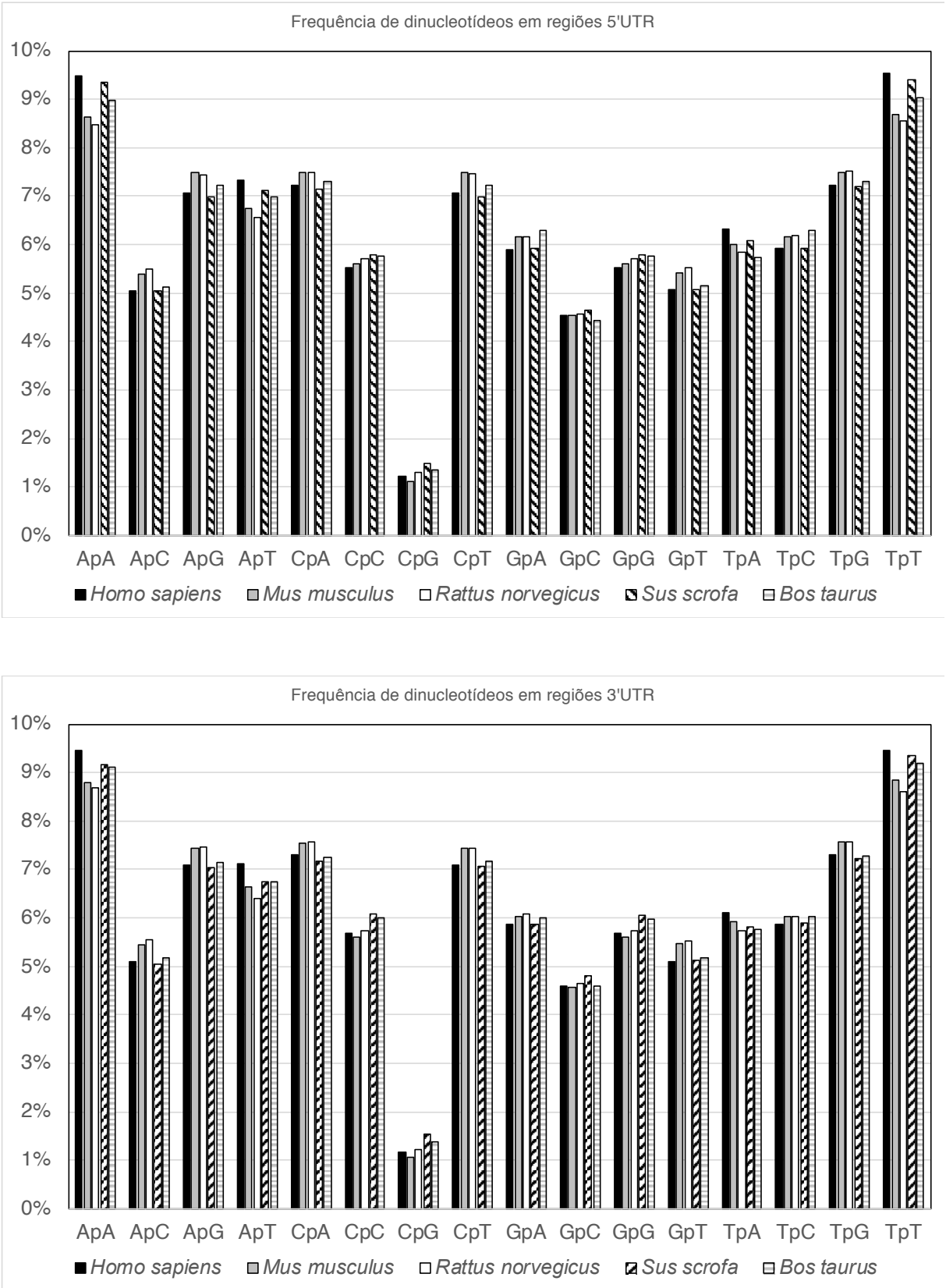


Figura 36: Frequência relativa de cada dinucleotídeo conforme organismo nas regiões 3' e 5' UTR. No eixo X temos as diferentes combinações de dinucleotídeos e no eixo Y a frequência relativa desses dinucleotídeos conforme o organismo. São mostrados resultados para o genoma de cada organismo os dados correspondentes a frequência relativa de dinucleotídeos nas regiões 3'UTR(gráfico superior) ou 5'UTR (gráfico inferior).

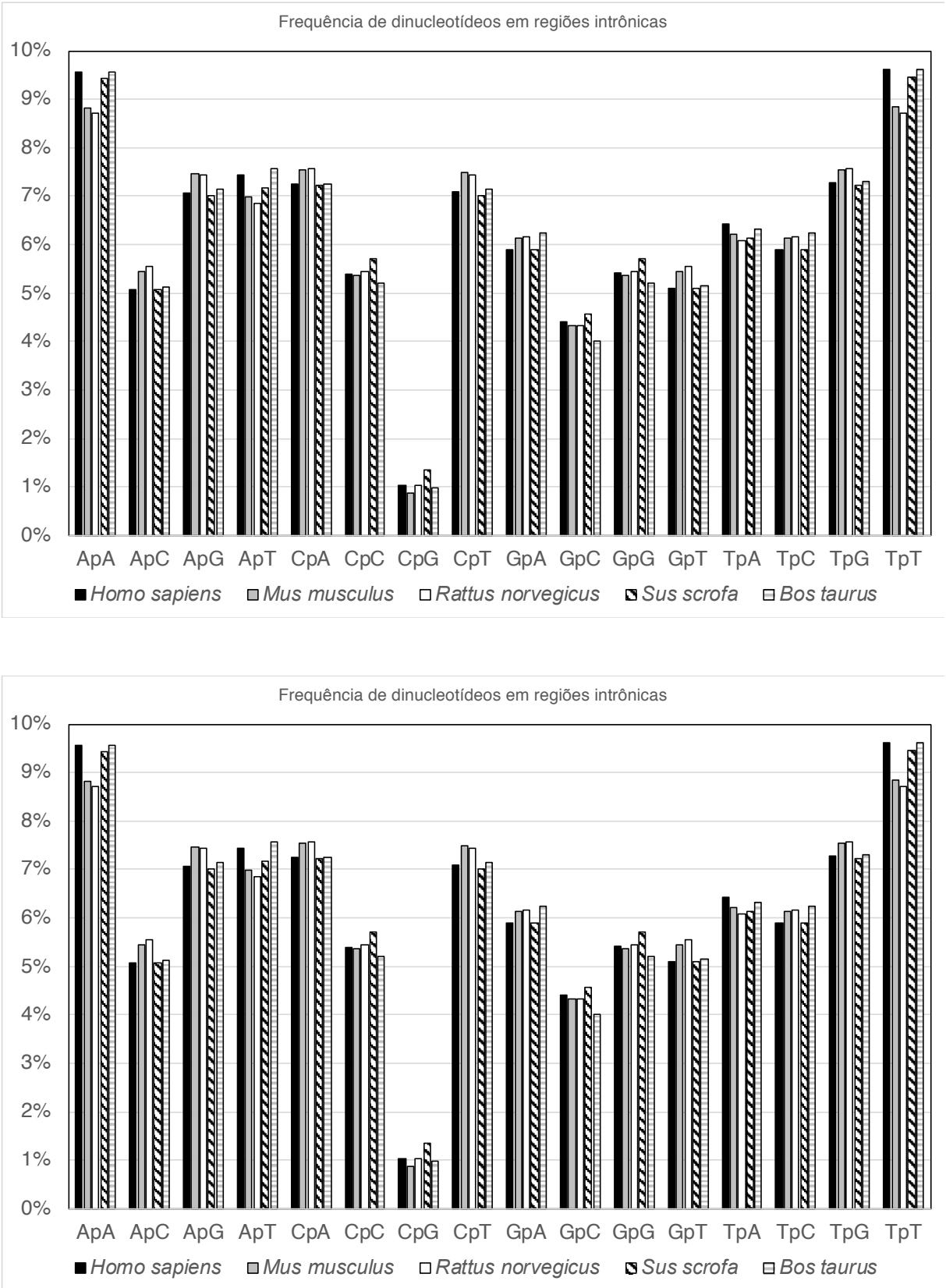


Figura 37: Frequência relativa de cada dinucleotídeo conforme organismo nas regiões intrônicas e codificadoras sinônimas. No eixo X temos as diferentes combinações de dinucleotídeos e no eixo Y a frequência relativa desses dinucleotídeos conforme o organismo. São mostrados resultados para o genoma de cada organismo os dados correspondentes a frequência relativa de dinucleotídeos nas regiões intrônicas (gráfico superior) ou codificadoras (gráfico inferior).

4.5 Análise do efeito da aminação dos dinucleotídeos TpG e CpA

Estudos prévios indicam que o fator mais determinante para que uma transição ocorra é o efeito dos dinucleotídeos CpG mutarem por metilação em C seguida de desaminação e transição para T, ocorrendo transição de G para A na replicação (ZHAO, 2002), e que essas mutações representam cerca de 25% dos SNPs (HESS'; BLAKE; BLAKE', [s.d.]; KRAWCZAK; BALL; COOPER, 1998 e OLIVEIRA, 2011) e resultam na sub-representação do dinucleotídeo CpG (ILLINGWORTH; BIRD, 2009; PLYLER et al., 2015). Com o objetivo de reverter essa sub-representação desse dinucleotídeo e verificar a influência no efeito dos nucleotídeos vizinhos, através de simulação computacional realizamos inicialmente uma suposta metilação em todas as citosinas de CpG e desaminação de C para T, e obtivemos efeito reduzido. Então de forma reversa revertemos computacionalmente T para C, antes de G. Assim, realizamos a aminação em diferentes proporções do número total dos dinucleotídeos CA e TG no genoma. Neste procedimento, os dinucleotídeos CA e TG foram revertidos para CG. Um exemplo da distribuição desses dinucleotídeos onde 25% e 35% deles foram aminados pode ser visualizado na Figura 38.

Tendo diferentes regiões genômicas (3'UTR, 5'UTR, regiões codificadoras e intrônicas) nos cinco mamíferos em análise (*Bos taurus, Homo sapiens, Mus musculus, Sus scrofa e Rattus Norvegicus*), calculamos novamente as frequências das bases vizinhas para cada uma das seis substituições (R,Y, M, K,W, S) de acordo com os percentuais de aminação de 5% a 95%. Para cada caso, calculamos a distância euclidiana entre a média das frequências relativas dos nucleotídeos das posições -5 e +5 para com a média das frequências relativas dos nucleotídeos das posições -1 e +1, de tal forma que, quanto menor for essa distância, mais o gráfico tenderia a uma linearidade e dessa forma, não teríamos mais o viés da vizinhança.

Quando plotamos essas distâncias conforme o organismo e a região genômica visto nas Figura 39 e Figuras suplementares 101-104, é possível notar algumas características em comum. Uma delas é a sobreposição das curvas dos tipos de SNPs que são complementares: K com M e R com Y. Nota-se também que, a curva das transições possui pouca variação na distância euclidiana conforme há um aumento na proporção de aminação, de tal forma que o efeito dos nucleotídeos vizinhos é pouco alterado. Isto tem que ser melhor estudado, mas note-se que as transições ocorrem entre as bases envolvidas na aminação e desaminação. As transições variam de duas formas distintas, entre as regiões codificadoras e as regiões não codificadoras (que são as regiões 3'UTR, 5'UTR e intrônicas), e isso pode estar relacionado com a diferença do conteúdo de nucleotídeos dessas duas regiões - regiões codificadoras

apresentam um maior conteúdo GC enquanto as regiões não codificadoras tendem a apresentar um maior conteúdo AT. Nas regiões codificadoras observa-se que as curvas das substituições W e S estão menos sobrepostas, tendo a curva da substituição S menores valores de distância euclidiana em relação a curva W para cada taxa de aminação. A partir desta observação sugere-se que a aminação dos dinucleotídeos TpG e CpA em regiões codificadoras em uma substituição sinônima de C/G, tem uma maior influência na redução da diferença nas frequências relativas de cada dinucleotídeo entre as posições mais distantes para com as mais próximas do que em uma substituição A/T. Apesar disso, o percentual de aminação em que essa distância euclidiana tem o seu valor mínimo é coincidente para essas duas substituições, ficando

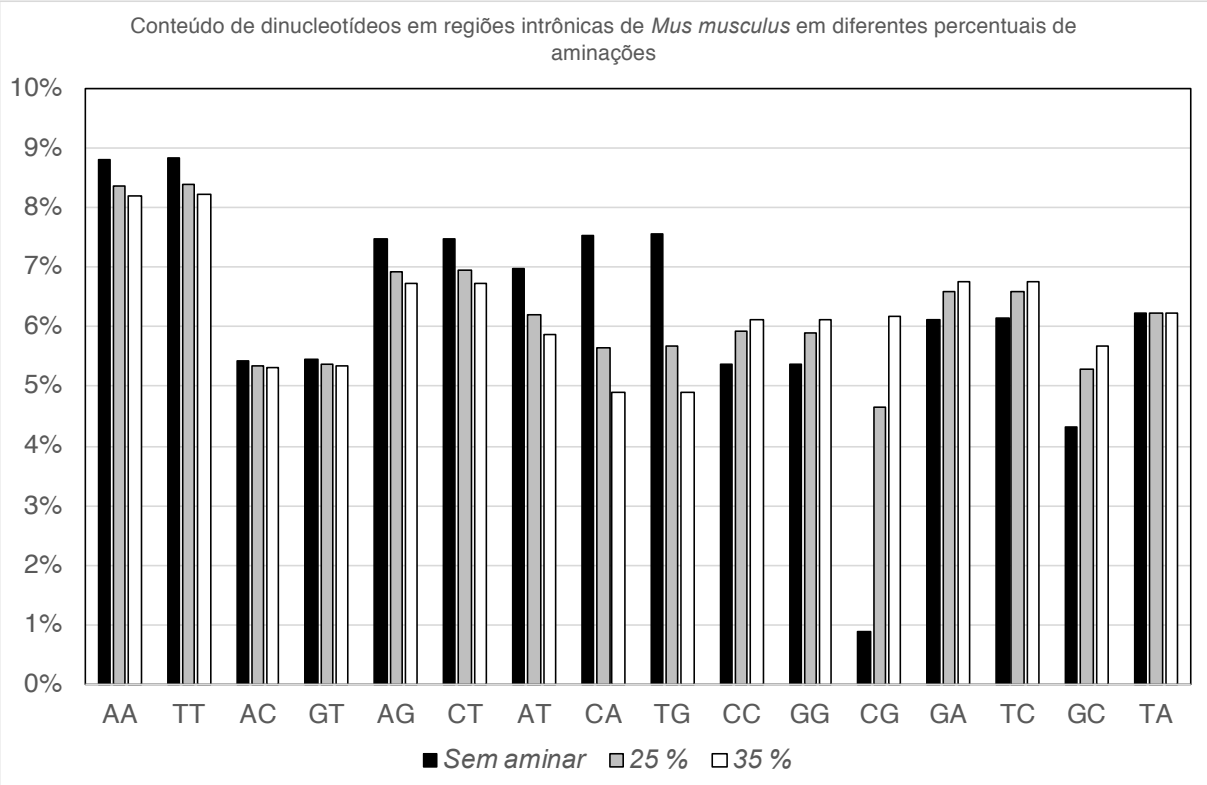


Figura 38: Frequência relativa de cada dinucleotídeo conforme em diferentes simulações de aminação nas regiões intrônicas de *Mus musculus*. No eixo X temos as diferentes combinações de dinucleotídeos e no eixo Y a frequência relativa desses dinucleotídeos conforme o percentual de aminação.

entre cerca de 35% a 40% nas regiões codificadoras que sofreram mutações sinônimas. Por sua vez, para as transições K e M nessas regiões essa porcentagem de aminação que torna a distância euclidiana mínima foi em torno de 30%.

Para as regiões não codificadoras, observamos que as curvas W e S tendem a se coincidirem e as porcentagens de aminação que minimizam a distância euclidiana estão entre 30 e 35%. Em *Sus scrofa* e *Homo sapiens* apresentaram menores valores de distância euclidiana

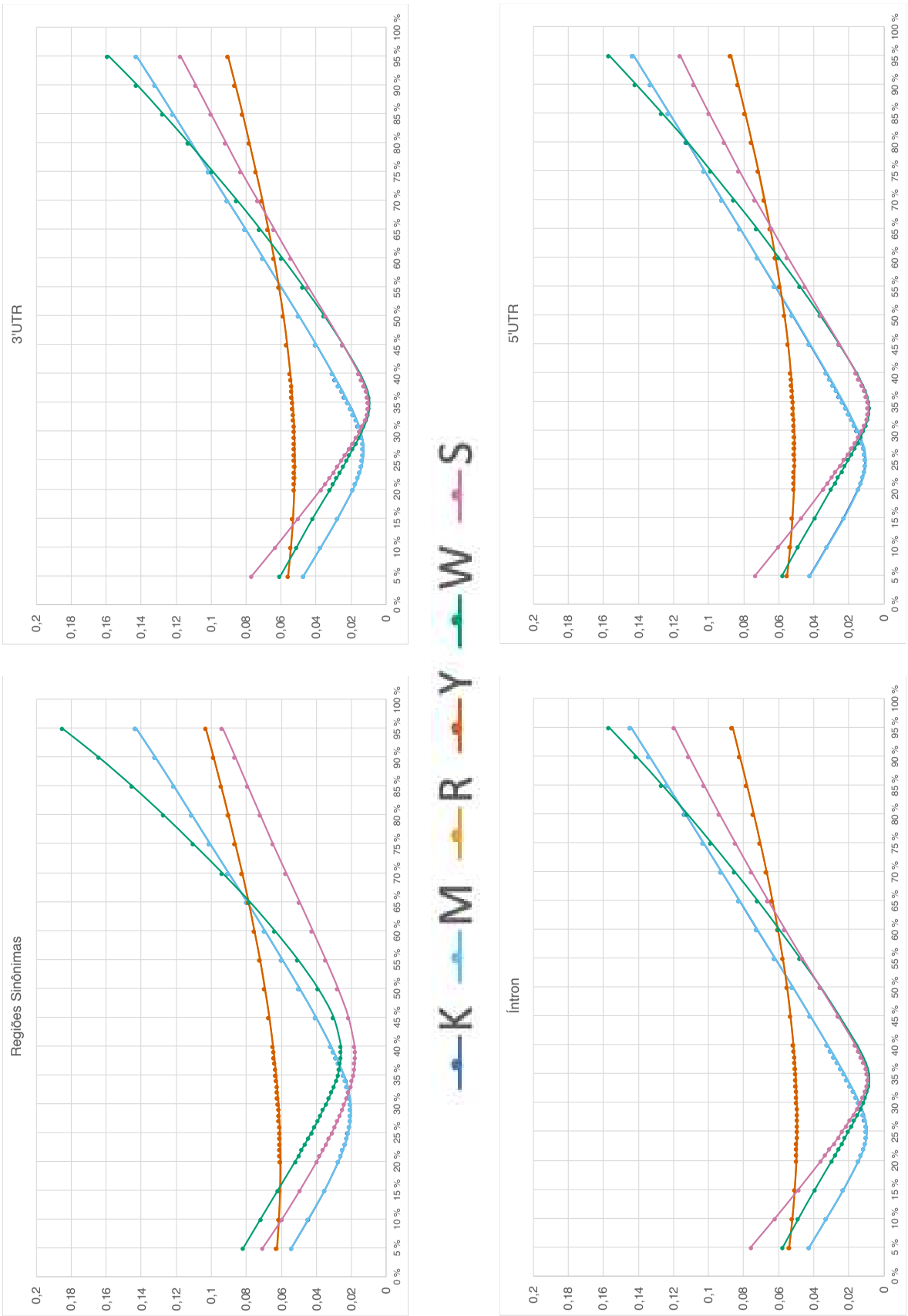


Figura 39: Distância Euclidiana entre as médias das posições 5 e 1 em diferentes regiões em *Mus musculus*. No eixo X temos os diferentes percentuais de aminação que da respectiva região do genoma e no eixo Y os valores da distância euclidiana entre a média das posições +1 e -1 e a média das posições +5 e -5.

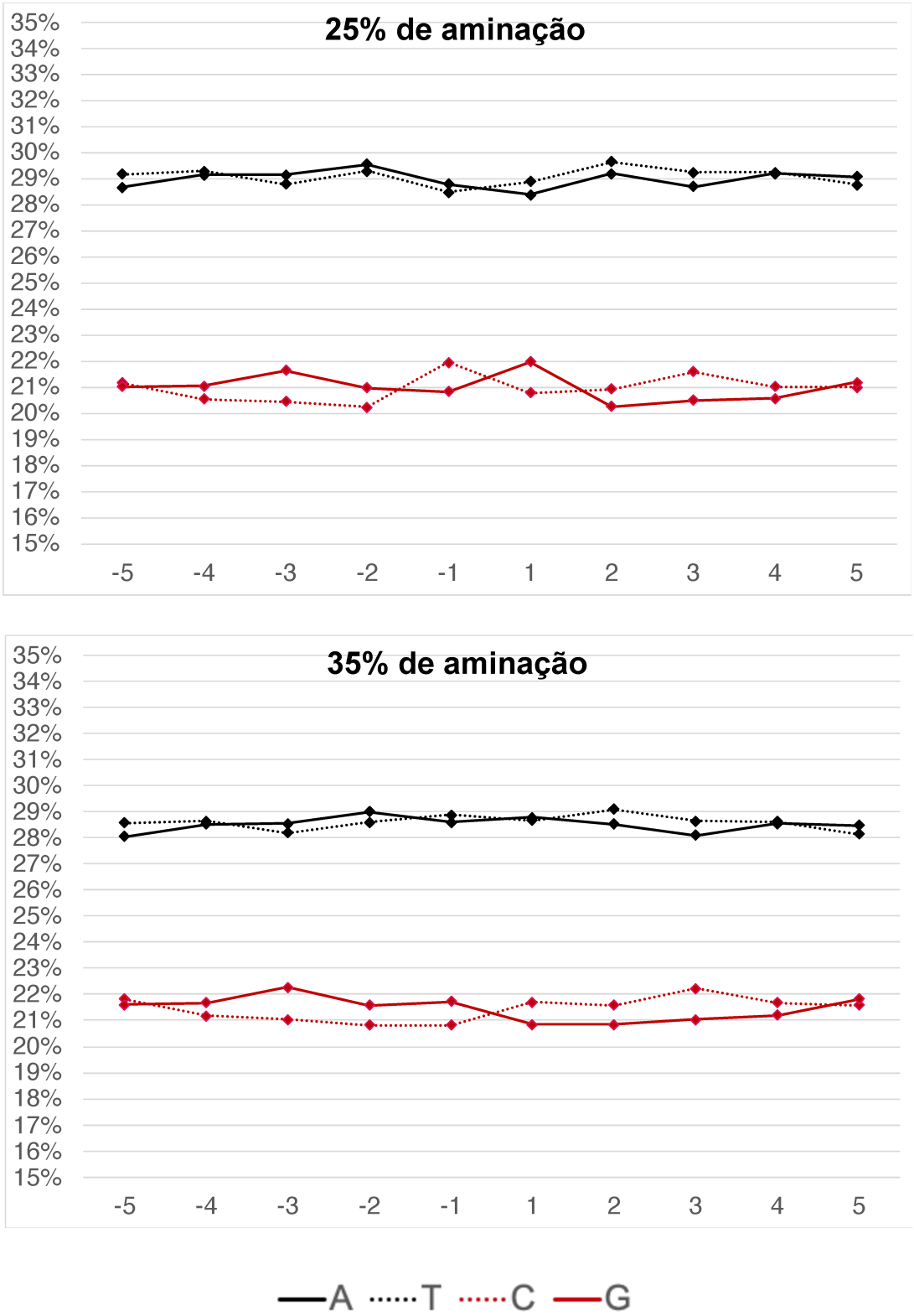


Figura 40: Ocorrência de bases nas proximidades do SNP W (A/T) e a frequência basal esperada no genoma, nas regiões intrônicas de *Homo sapiens* em duas situações de aminação distintas. São mostrados resultados para a simulação do genoma aminado em 25% no gráfico superior e a simulação do genoma aminado em 35% no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

do que nos demais organismos. Nas substituições K e M esses valores não chegaram a ser tão baixos quanto em W e S, que acontecem num percentual de 25% a 30%. Isto sugere que a aminação tem um maior efeito, no sentido de diminuir a variação da frequência de nucleotídeos que ocorre entre as posições mais próximos e mais distantes a um SNP, nas trocas A/T e C/G do que nas trocas G/T e A/C.

Como exemplo, temos na Figura 40 o efeito de diferentes percentuais de aminação para nas frequências dos nucleotídeos próximos aos SNPs do tipo W nas regiões intrônicas de *Homo* conteúdo AT. Em 25% de aminação, há uma maior frequência relativa de Citosina e Guanina nas posições 1 e -1; e a menor frequência de Adenina e Timina nas posições 1 e -1 que podem ser observadas de forma pouco pronunciadas. Com 35% de aminação observamos que as curvas do gráfico tendem a uma linearidade - sem sobressaltos de maior frequência relativa inclusive nas adjacências imediatas à simulação de uma substituição.

Para cada uma das 120 categorias conduzidas nesse estudo - quatro diferentes regiões genômicas em seis substituições distintas em cada um dos cinco organismos analisados - as frequências de nucleotídeos após a aminação em 25% e 35% de TpG e CpA podem ser visualizadas na base de dados SNP LANE. Para isso, após a seleção da categoria, basta localizar o menu *dropdown*, que fica abaixo dos gráficos de frequência basal do genoma, e escolher o percentual de aminação. São mostradas representações gráficas como as apresentadas na Figura 40. De forma marcante, o aparente efeito de vizinhança que SNP LANE retrata para todas as 120 categorias, é apagado quando se retorna a condições supostamente ancestrais de metilação em CpG que com o tempo foram sofrendo desaminação em C, gerando T. A Figura 40 mostra os resultados que foram mais conspícuos, envolvendo a vizinhança de W e S. E sugerem que o efeito de vizinhança se apaga com a aminação computacional, sugerindo que não há efeito de vizinhança. Estudos posteriores podem retratar melhor o porquê das nuances quando as diversas mutações são estudadas.

5. DISCUSSÃO

5.1 Frequência de transições e transversões em diferentes mamíferos

Para inferir nas amostras contidas em nosso banco de dados se havia um efeito de vizinhança sobre um dado tipo de SNP - e qual seria, neste trabalho analisamos um total de 392.591.193 SNPs, de cinco organismos e posicionados em quatro regiões genômicas. Essa separação por regiões foi em função da conhecida diferença de conteúdo GC entre éxons e íntrons e também porquê regiões codificadoras de proteínas ou sequências nas quais a mutação pode afetar a função, e sofrem uma maior pressão seletiva do que regiões não funcionais do genoma - (WATSON, CRICK, 1953; Oliveira, 2011; (LING et al., 2019). De fato observamos que quando comparamos os resultados de SNPs intrônicos com aqueles SNPs que causam uma alteração sinônima em regiões codificadoras, as frequências são distintas. SNPs sinônimos foram escolhidos por escaparem, se não quase que totalmente, da pressão seletiva que recai sobre os não sinônimos. A princípio poderiam apresentar resultados similares a de regiões não codificadoras, todavia a é composição diferente. E *a priori* não seria possível estabelecer que as regiões não traduzidas, tanto 5' como 3' apresentariam comportamento similar a íntrons - assim se justificam as inclusões dessas categorias. Os organismos e as regiões foram escolhidos também pela alta quantidade de dados já classificados no dbSNP, fonte dos dados.

Em comparações com trabalhos prévios de *H. sapiens*, analisamos um volume de dados que é 37 vezes maior do que estudos prévios de 2006 (ZHAO; ZHANG, 2006), o que se deve ao acelerado ritmo dos avanços na tecnologia de sequenciamento do genoma que ultrapassou a Lei de Moore e possibilitou um barateamento no custo de sequenciamento e, um aumento na velocidade de sequenciamento do genoma. Apesar da grande diferença no volume de dados de SNPs analisados, as proporções de transições e transversões encontradas no estudo de 2006 por Zhao; Zhang (ZHAO; ZHANG, 2006) foram bem próximas da que foi obtido neste estudo (Tabela 5).

A forma de comparação com o que se esperaria de um genoma sem as mutações, nesse estudo, se baseou em contagens das proporções de cada nucleotídeo no genoma, considerando a diferença dessas composições de nucleotídeos em regiões distintas do genoma. A esta medida chamamos de frequência basal de vizinhança conjuntamente a uma dupla de bases que estariam compondo um SNP.

Tabela 5: Comparativo da frequência de cada substituição em *H. sapiens* em diferentes estudos.

Tipo de substituição:	Nossos dados:	*Zhao; Zhang:*
R	32,89 %	32,77 %
Y	32,81 %	32,81 %
K	8,86 %	9,06 %
M	8,80 %	8,98 %
S	9,10 %	8,92 %
W	7,55 %	7,55 %

Em *Mus musculus*, a comparação foi bem próxima de *Homo sapiens*, houve um aumento do volume de dados para com a análise prévia de *Zhang; Zhao* (ZHANG; ZHAO, 2004) em 30 vezes, mas a proporção dos dados encontrada de transições e transversões foram bastante similar.

No entanto, quando comparamos com estudos prévios de *Jiang et al*. (JIANG et al., 2008) em bovinos, notamos grande distinção em relação as proporções de transições e transversões, diferença esta que pode ser atribuído ao crescimento de volume de dados disponíveis na análise. Enquanto o primeiro estudo analisou 15.110 SNPs bovinos disponíveis no dbSNP em 2008, em nossa análise, baseada no último release do dbSNP de 2017, analisamos um volume de dados 2521 vezes maior. As cinco espécies analisadas, em geral, compartilham os mesmos cenários adjacentes de vizinhança que flanqueavam os SNPs e *baseline frequency*, exceto os bovinos que apresentaram maior divergência com os demais organismos e análises prévias em relação aos padrões de transições e transversões (Tabela 6).

Tabela 6: Comparativo da frequência de cada substituição em *B. taurus* em diferentes estudos.

Tipo de substituição:	Nossos dados:	*Jiang et al.:*
R	25,11 %	34,63 %
Y	25,35 %	35,13 %
K	14,45 %	8,27 %
M	14,53 %	8,15 %
S	10,96 %	7,68 %
W	9,59 %	6,14 %

A inclusão de *Sus scrofa* em nossa análise foi importante pois, assim como *Bos taurus*, ele é classificado na superordem Laurasiatheria, distinta da superordem Euarchontoglires que reúne os primatas, e os roedores (como camundongo e, rato), além de lagomorfos (coelhos), escadêncios (ordem Scadentia) e os colugos (ordem Dermoptera). A literatura continha dados

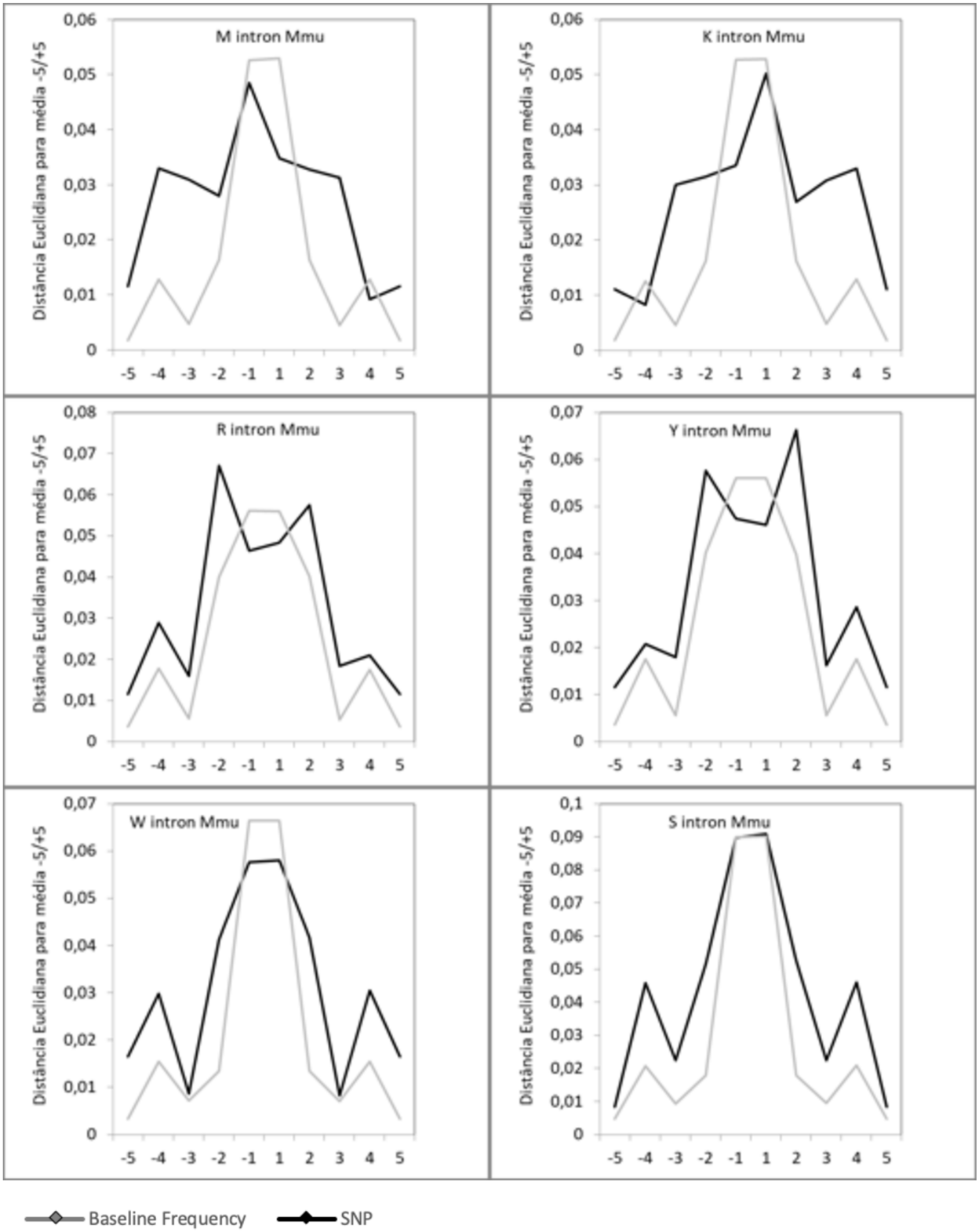


Figura 41: Distância Euclidiana entre cada posição e as médias das posições +5 e -5 em diferentes regiões em *Mus musculus*. No eixo X temos os diferentes posições da respectiva região do genoma e no eixo Y os valores da distância euclidiana entre àquela posição e a média das posições +5 e -5. As linhas mais escuras representam as distâncias euclidianas calculadas com baseadas nas frequências relativas em torno de um SNP e as linhas em cinza claro representam as distâncias euclidianas baseadas na *baseline frequency*.

de bovinos e eles eram discrepantes na quantidade de transições e transversões em relação aos humanos e roedores, como mostrado na Tabela 4. Desta forma, como o porco gerou resultados semelhantes a primata e roedores, podemos inferir que os dados de bovinos devem estar enviesados e que, possivelmente, esta discrepância não é característico de um clado mais distante, neste caso uma outra superordem.

Como apontado acima, a quantidade de dados é sensivelmente maior em nossos estudos, e demandou intenso processamento. Para evitar que outros tenham que reprocessar os dados para obter todas as combinações e observar os dados, diferentemente dos estudos anteriores, organizamos o material em uma base de dados, SNP LANE. A base possui quatro tipos de representações gráficas para cada combinação, logo 480 gráficos. Pretendemos ampliar as posições de vizinhança para 10 a 20 bases flanqueando as substituições - ainda que a complexidade do modelo aumenta exponencialmente à medida que o tamanho do contexto aumenta, segundo a literatura (LING et al., 2019). A razão de utilizarmos cinco foi o grande tempo de processamento, pois trata-se de um projeto que pode ser qualificado como de "big data".

Como mencionado acima, outro ponto de importância desse trabalho foi disponibilizar os dados na base SNP LANE: SNP LocAl NEighborhood, que mostra a pequena vizinhança de todas os tipos de substituição, em cinco animais e quatro regiões distintas, de fácil acesso, de forma que outros investigadores podem inspecioná-los para gerar novas abordagens com facilidade.

5.2 Efeito dos nucleotídeos vizinhos para a ocorrência de uma mutação

A partir das posições dos SNPs que ocorrem nos genomas dos organismos analisados obtidos no banco de dados dbSNP, foi possível verificar para cada tipo de variação a frequência de cada nucleotídeo nas posições vizinhas a ela. Para verificarmos se de fato alguma posição possui valores de frequências nucleotídicas fora do esperado, adotamos métodos de simulação para estimar a frequência esperada de um nucleotídeo a k posições de uma variação. O cálculo desta frequência poderia ser obtida sistematicamente verificando para cada posição do genoma a ocorrência de cada nucleotídeo na sua vizinhança. No entanto, adotamos o procedimento baseado em simulação computacional por permitir a obtenção destes resultados em um tempo hábil. A hipótese de que que a não aleatoriedade do genoma causaria o efeito aparente de vizinhança, pois é preciso um T depois de C para que o SNP T/A possa ser observado depois

de C, poderia ser analisada matematicamente, mas preferimos adotar uma simulação computacional. Visamos com a simulação inferir a linha de base da vizinhança local. O algoritmo percorria as sequências de uma dada região, intron, por exemplo, e sorteava posições. Na análise de W, se a base sorteada fosse A ou T, guardavam-se as bases nas posições vizinhas. Ao final determinava-se o padrão de bases na vizinhança do par A/T. As curvas observadas e a linha de base são mostradas para todas as combinações de organismo, região e tipo de substituição na SNP LANE, para inspeção. Verificamos primeiramente com análise de W que os perfis das plotagens eram muito semelhantes. Depois de testar várias métricas de distância para a comparação entre os gráficos, concluímos por utilizar o teste de qui-quadrado (*chi-square*) comparando as curvas como um todo - teste também classificado como o melhor para esse tipo de análise no estudo de *Plyler et al* (PLYLER et al., 2015). Verificamos que a grande maioria não apresenta diferença significativa no teste, os 92 casos em que houveram diferença significativa podem ser observados na Tabela suplementar 1 do apêndice dessa dissertação. Na SNP LANE são mostrados, sob posicionamento do *mouse* ("*mouse over*"), as proporções de bases em cada posição. Estas proporções foram comparadas entre os dados observados e a linha de base com o teste de qui-quadrado. As posições, poucas, que apresentam diferença significativa entre os dois grupos de valores, são notadas no gráfico com asterisco, (ex: -1*). Assim o usuário pode prontamente verificar o resultado do teste ao inspecionar o site. Uma lista de situações onde ocorrem diferenças significativas será adicionada em breve ao site.

A análise preliminar dos gráficos obtidos indicava que as curvas se estabilizavam muitas vezes quando se chegava às posições mais distantes da variação, neste trabalho -5 e +5. Depois de vários testes concluímos que seria interessante medir a distância Euclidiana entre (i) a média das proporções de A, C, G, T nas posições -5 e 5, e (ii) as proporções das mesmas bases em uma dada posição. Com isso geramos gráficos como os da Figura 41. Futuramente pretendemos adicionar este formato de figura ao site.

Quando analisamos a frequência de substituições em vários organismos e regiões, já nos deparamos com a influência que a ocorrência basal de bases no genoma teria sobre as médias. É de ampla documentação na literatura que a ocorrência de bases não é aleatória, sendo a frequência de dinucleotídeos e a riqueza de GC dois fatores que poderiam influir na medida de um possível efeito de vizinhança (ROMIGUIER et al., 2013; ROMIGUIER; ROUX, 2017; ZHAO; ZHANG, 2006). Exemplificando melhor, como seria possível medir níveis globais da transversão G/C tratada aqui como S, em regiões virtualmente constituídas somente por AT? Aplicando o exemplo às medidas de efeito de vizinhança sobre a geração de substituições, se é obtido que a base C é encontrada em maior frequência a -1 da substituição W, isso poderia ser

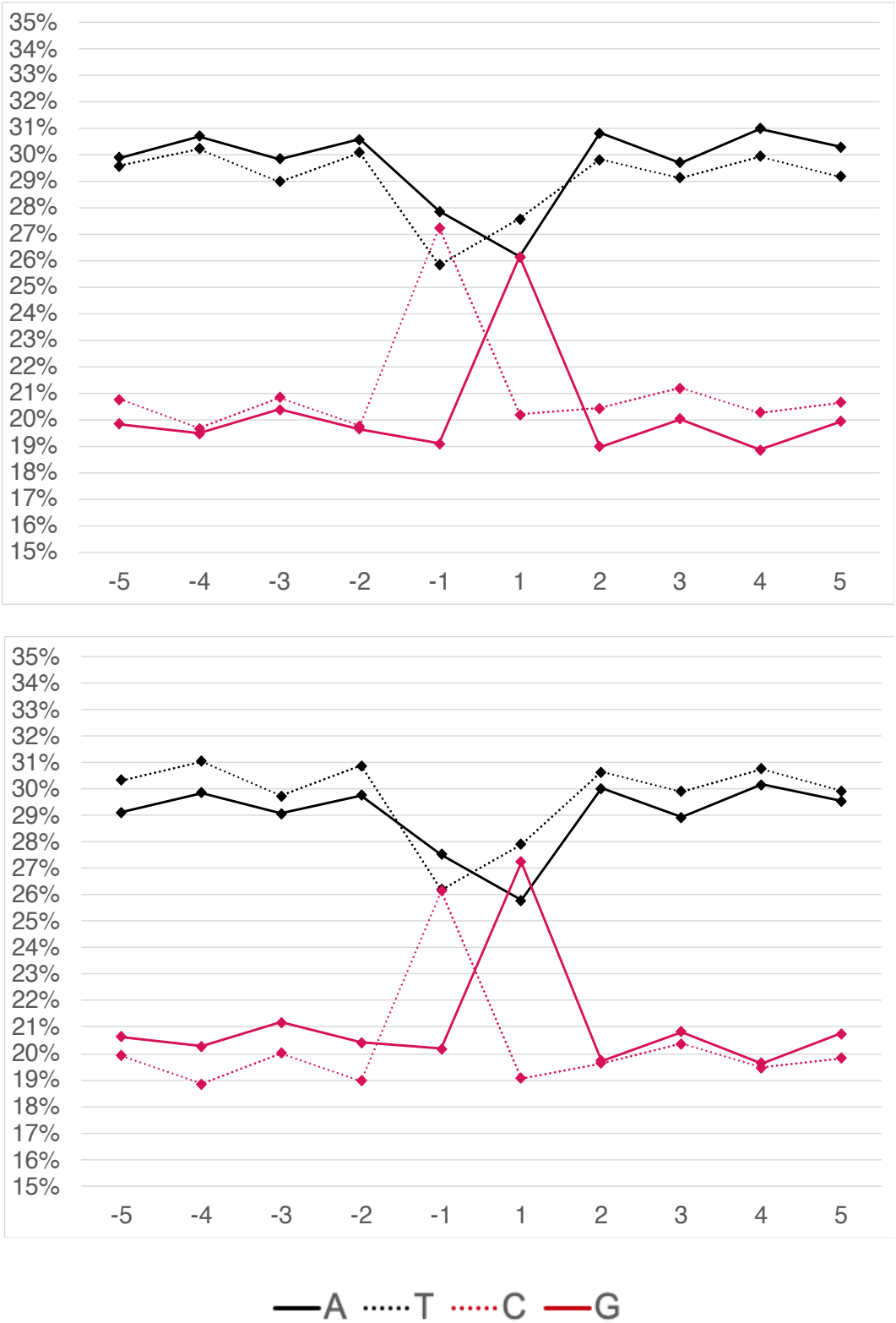


Figura 42: Ocorrência de bases nas proximidades do SNP W (A/T) em simulações de desaminação do dinucleotídeo CpG em *Mus musculus* em regiões intrônicas. São mostrados resultados para a desaminação do dinucleotídeo CpG em CpA no gráfico superior e desaminação do dinucleotídeo CpG em TpG no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

explicado pela maior ocorrência de uma base T para se detectar a substituição T/A, hipótese que foi nomeada "It takes Two to T/Ango" informalmente. Isso apoia a visão anterior de que uma sequência rica em AT pode ter mais probabilidade de mudanças de nucleotídeos entre A e T, do que uma sequência rica em dinucleotídeos CA e/ou CT, resultando em um aparente efeito de vizinhança que poderia, no entanto, não existir. Durante nosso trabalho trabalhamos com a hipótese de que é preciso haver GC pode ter mais probabilidade de mudanças de nucleotídeos entre C e G (Li, 1997; (ZHAO; ZHANG, 2006).

5.3 Desaminação *in silico* das citosinas em CpG

Alguns trabalhos (LING et al., 2019; PLYLER et al., 2015; JIANG et al., 2008; ZHAO; ZHANG, 2006; ZHANG, 2005; ZHANG; ZHAO, 2004; ZHAO, 2002) sugeriram que havia o efeito de vizinhança e que dele faria parte o fenômeno de desmetilação de Citosinas no dinucleotídeo CpG. Quando a citosina sofre desaminação, o que pode ocorrer espontaneamente, ela se transforma na base U, todavia a citosina metilada na posição 5 se transforma em T. Como CpG é complementar a CpG, quanto uma citosina metilada é desaminada em uma fita, gera TpG, mas se é a complementar ao G da fita considerada, a base complementar a este G muta de C a T, logo este G se propaga como A, e o dinucleotídeo resultante é CpA. Para testar a contribuição desse fenômeno nos perfis em SNP LANE, nós inicialmente desaminamos computacionalmente todas as citosinas de dinucleotídeos CpG, esperando obter um efeito enorme nos perfis. No entanto, a diferença encontrada foi pequena (Figura 42). De fato, quando dois sítios adjacentes imediatos fornecem estruturas para CpG, aumentou significativamente transições em comparação com as estruturas sem o CpG. No entanto essa comparação em nossos dados foi tanto verdadeira para regiões onde ocorre o polimorfismo quanto em simulações da frequência basal do genoma.

O que nos indicou que estávamos no caminho correto foi o fato de que quando mutávamos CpG da fita em análise para TpG, aumentava um pouco a frequência na posição -1 (Figura 42), e quando o script alterava os dinucleotídeos CpG desta mesma fita para CpA, a alteração residia na posição +1, embora pequena. Então mudamos de abordagem e decidimos computacionalmente aminar as timinas em TpG, revertendo-as a citosinas metiladas, gerando supostos CpG que teriam sido desaminados a partir de um genoma original. Como o script trabalha em uma fita, também convertemos CpA em CpG, pois a aminação (em contraposição a grupo cetona que ocorre na timina) da base T complementar ao A de CpA geraria um C e

consequentemente na fita em análise o dinucleotídeo reverteria para CpG. E foi gerado um script que executava essa operação em diferentes porcentagens de TpG e CpA (Figura 40). Nossa simulação de aminação de bases T gerou resultados bastante claros para as substituições W e S. As curvas perdem totalmente o viés (Figura 42). A distância Euclidiana foi medido entre as médias das posições -5 e +5 contra as demais posições (Figura 40) e consideramos a média das distâncias com as posições -1 e +1 para inferir qual porcentagem de aminação diminuiria ao máximo o efeito de vizinhança (Figura 39 e Figuras suplementares 101-104). Para W e S, esse valor foi de 35%. O efeito não é tão claro para as substituições K e M, mas observa-se efeito similar quando a porcentagem de aminação é de 25%. Não sabemos explicar totalmente a razão dos dados não serem claros para as substituições R e Y, mas uma possível explicação é que a mudança de CpG para TpG e de CpG para CpA já envolvem em si as transições, respectivamente, Y e R. A curva de redução da distância Euclidiana aponta uma tendência mas muito suave (Figura 39). Futuros estudos com modelagem matemática podem clarear essa questão.

A sub-representação do dinucleotídeo CpG e consequente previsão de um genoma ancestral contendo 25% a 35% de citosinas metiladas - é discutida desde o sequenciamento do genoma humano, que estimou em 21% esse valor para o genoma de *Homo sapiens* como um todo (ILLINGWORTH; BIRD, 2009; VENTER et al., 2001). Não havia essa previsão de diferentes sub-representações do dinucleotídeo CpG conforme região genômica e em organismos não-humanos.

Encontramos um maior volume de SNPs na vizinhança imediata de um CpG, ou dentro de um próprio dinucleotídeo CpG, e essa sub-representação de dinucleotídeos CpG é também visualizada no uso de códons em *Mus musculus* e *Homo sapiens*. Por exemplo, dos tripletos múltiplos de nucleotídeos disponíveis para serina, prolina, treonina e alanina (seis para serina; quatro para prolina, treonina e alanina), os códons contendo CG são marcadamente subutilizados). A sub-representação de dinucleotídeos CG dentro de éxons de camundongos foi relatada anteriormente (INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM, 2001) diminuiria a formação exônica de SNP e ajudaria a preservar o DNA codificador de proteínas ao longo da escala evolutiva de tempo (LING et al., 2019). No entanto é a primeira vez que a não aleatoriedade do genoma aparentando o efeito de vizinhança na probabilidade de SNPs é amplamente documentada e que simulações de retorno de CpG a sua frequência anterior eliminam o viés.

6. CONCLUSÕES

Após um estudo aprofundado em sequencias contendo SNPs chegamos a várias conclusões. A taxa de transições e transversões - à exceção de *Bos taurus* - aproximou-se da proporção 3:1:1. Estendendo a análise da sequência para a vizinhança dos SNPs notamos vieses fortes e evidentes da proporção de nucleotídeos flanqueando os SNPs analisados em sequências de 1-5 pb em ambos os lados de um SNP. No entanto, uma diferença entre os resultados relatados aqui e aqueles relatados por outros pesquisadores é que em uma comparação com a frequência basal do genoma, observamos esses mesmos vieses da proporção de nucleotídeos com apenas cerca de 7,6% dos casos apresentando diferença significativa entre essas frequências relativas de cada nucleotídeo por posição.

Os achados não são compatíveis com um modelo evolutivo de DNA tipo "neutro", pois o genoma é fortemente não aleatório, como evidenciado pelo contexto local. Evidenciamos que as taxas de mutação apresentam uma especificidade conforme a região genômica e é similar em diferentes mamíferos. O exemplo mais notável é o processo de desaminação de citosinas metiladas - geralmente presentes em dinucleotídeos CpG - a dependência de contexto para a composição do genoma pode apresentar vários fatores para a sua causalidade, incluindo a fidelidade e a correção da polimerase (SIEPEL, 2003), valores de energia livre e flexibilidade de uma microrregião (Oliveira, 2011).

Outro ponto de importância desse trabalho foi disponibilizar os dados na base SNP LANE: SNP LocAl NEighborhood que, mostra a pequena vizinhança de todas os tipos de substituição, em cinco animais e quatro regiões distintas, de fácil acesso, de forma que outros investigadores possam inspecioná-los para gerar novas abordagens com facilidade.

7. PERSPECTIVAS

- Atualizar o pipeline de trabalho para os novos formatos do dbSNP e EVA;

- Expandir o SNP LANE de forma a incluir outros organismos-modelo como *Drosophila melanogaster*, *Gallus gallus*, *Danio rerio*, *Pan troglodytes,* para verificar se há efeito de vizinhança nesses organismos e visualizar em um contexto evolutivo;

- Expandir o SNP LANE para incluir outras regiões genômicas como regiões de SNPs *missense* e *non-sense* - de forma a explorar a influência da pressão seletiva;

- Usar o pipeline de trabalho para sequências de SNPs tumorais do TCGA, objetivando ver se há um padrão específico de efeito de vizinhança para SNPs tumorais ou se as frequências relativas em torno desses SNPs é equivalente aos demais SNPs ou com a frequência basal do genoma;

- Analisar os SNPs em uma perspectiva de duplas, observando as frequências de pares em torno do SNP, por exemplo, analisar concomitantemente as posições +1 e -1 e ver quais as combinações de nucleotídeos apresentam maior frequência relativa entre os diferentes organismos.

8. REFERÊNCIAS BIBLIOGRÁFICAS

BILLINGS, T. et al. Patterns of Recombination Activity on Mouse Chromosome 11 Revealed by High Resolution Mapping. **PLoS ONE**, v. 5, n. 12, p. e15340, 8 dez. 2010.

BODMER, W. Genetic Characterization of Human Populations: From ABO to a Genetic Map of the British People. **Genetics**, v. 199, n. 2, p. 267–279, fev. 2015.

dbSNP architecture redesign supports future human variation data expansion; changes to be introduced over the next year. **NCBI INSIGHTS,** 09 de maio de 2017. Disponível em: https://ncbiinsights.ncbi.nlm.nih.gov/2017/07/07/dbsnp-redesign-supports-future-data-expansion/ Acesso em: 01 de junho de 2019

ELLEGREN, H.; GALTIER, N. Determinants of genetic diversity. **Nature Reviews Genetics**, v. 17, n. 7, p. 422–433, jul. 2016.

ENSEMBL. Disponível em: https://www.ensembl.org/Sus_scrofa/Info/Annotation Acesso em: 19 de set de 2019

ENSEMBL. Disponível em: https://www.ensembl.org/Mus_musculus/Info/Annotation Acesso em: 19 de set de 2019

European Variation Archive (EVA): A database of all types of genetic variation data from all species. **EUROPEAN MOLECULAR BIOLOGY LABORATORY.** Disponível em: https://www.ebi.ac.uk/training/events/2018/european-variation-archive-eva-database-all-types-genetic-variation-data-all-species Acesso em: 01 de junho de 2019

FORCADA, J.; HOFFMAN, J. I. Climate change selects for heterozygosity in a declining fur seal population. **Nature**, v. 511, n. 7510, p. 462–465, jul. 2014.

GOJOBORI, T.; LI, W.-H.; GRAUR, D. Patterns of nucleotide substitution in pseudogenes and functional genes. **Journal of Molecular Evolution**, v. 18, n. 5, p. 360–369, set. 1982.

HAKE, S.; ROSS-IBARRA, J. Genetic, evolutionary and plant breeding insights from the domestication of maize. **eLife**, v. 4, p. e05861, 25 mar. 2015.

HESS', S. T.; BLAKE, J. D.; BLAKE', R. D. Wide Variations in Neighbor-dependent Substitution. p. 12, [s.d.].

HODGKINSON, A.; EYRE-WALKER, A. Variation in the mutation rate across mammalian genomes. **Nature Reviews Genetics**, v. 12, n. 11, p. 756–766, nov. 2011.

ILLINGWORTH, R. S.; BIRD, A. P. CpG islands - 'A rough guide'. **FEBS Letters**, v. 583, n. 11, p. 1713–1720, 5 jun. 2009.

INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM. Initial sequencing and analysis of the human genome. **Nature**, v. 409, n. 6822, p. 860–921, fev. 2001.

JIANG, Z. et al. A Novel Type of Sequence Variation: Multiple-Nucleotide Length Polymorphisms Discovered in the Bovine Genome. **Genetics**, v. 176, n. 1, p. 403–407, maio 2007.

JIANG, Z. et al. The Complementary Neighborhood Patterns and Methylation-to-Mutation Likelihood Structures of 15,110 Single-Nucleotide Polymorphisms in the Bovine Genome. **Genetics**, v. 180, n. 1, p. 639–647, set. 2008.

KIM, S.; MISRA, A. SNP Genotyping: Technologies and Biomedical Applications. **Annual Review of Biomedical Engineering**, v. 9, n. 1, p. 289–320, 15 ago. 2007.

KITTS, A. et al. The Database of Short Genetic Variation (dbSNP). p. 41, [s.d.].

KRAWCZAK, M.; BALL, E. V.; COOPER, D. N. Neighboring-Nucleotide Effects on the Rates of Germ-Line Single-Base-Pair Substitution in Human Genes. **The American Journal of Human Genetics**, v. 63, n. 2, p. 474–488, ago. 1998.

LI, Wen-hsiung. **Molecular Evolution.** Portland: Sinauer Associates Inc, 1997. 487 p.

LING, G. et al. **A Bayesian framework for inferring the influence of sequence context on single base modifications**. [s.l.] Bioinformatics, 9 mar. 2019. Disponível em: <http://biorxiv.org/lookup/doi/10.1101/571646>. Acesso em: 5 jun. 2019.

LYNCH, M. Evolution of the mutation rate. **Trends in Genetics**, v. 26, n. 8, p. 345–352, ago. 2010.

MAJEWSKI, J. Dependence of Mutational Asymmetry on Gene-Expression Levels in the Human Genome. **The American Journal of Human Genetics**, v. 73, n. 3, p. 688–692, set. 2003.

MILLER, D. E. et al. A Whole-Chromosome Analysis of Meiotic Recombination in *Drosophila melanogaster*. **G3: Genes|Genomes|Genetics**, v. 2, n. 2, p. 249–260, fev. 2012.

Non-human dbSNP Import Status. **EUROPEAN MOLECULAR BIOLOGY LABORATORY.** Disponível em: https://www.ebi.ac.uk/eva/?dbSNP-Import-Progress Acesso em: 01 de junho de 2019

NUSSBAUM, Robert L.; MCINNES, Roderick R.; WILLARD, Huntington F.. **Thompson & Thompson Genética Médica.** 8. ed. Rio de Janeiro: Elsevier, 2016. 560 p.

OLIVEIRA, Luciana Márcia de. **Estudos Genômicos de flexibilidade e energia associadas a distribuição de SNPs.** 2011. 105 f. Dissertação (Mestrado) - Curso de Programa de Pós-graduação em Biotecnologia, Núcleo de Pesquisas em Ciências Biológicas, Universidade Federal de Ouro Preto, Ouro Preto, 2011.

PEARSON, W. R.; LIPMAN, D. J. Improved tools for biological sequence comparison. **Proceedings of the National Academy of Sciences**, v. 85, n. 8, p. 2444–2448, 1 abr. 1988.

PELLEY, John W.. Organization, Synthesis, and Repair of DNA. **Elsevier's Integrated Review Biochemistry**, [s.l.], p.125-135, 2012. Elsevier. http://dx.doi.org/10.1016/b978-0-323-07446-9.00015-5.

Phasing out support for non-human genome organism data in dbSNP and dbVar. **NCBI INSIGHTS,** 09 de maio de 2017. Disponível em:

https://ncbiinsights.ncbi.nlm.nih.gov/2017/05/09/phasing-out-support-for-non-human-genome-organism-data-in-dbsnp-and-dbvar/ Acesso em: 01 de junho de 2019

PLYLER, Z. E. et al. SNP Formation Bias in the Murine Genome Provides Evidence for Parallel Evolution. **Genome Biology and Evolution**, p. evv150, 6 ago. 2015.

PRAY, L. (2008) Discovery of DNA structure and function: Watson and Crick. **Nature Education** 1(1):100. Disponível em: <https://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397>. Acesso em 19/03/2019

QUINTANA-MURCI, L.; CLARK, A. G. Population genetic tools for dissecting innate immunity in humans. **Nature Reviews Immunology**, v. 13, n. 4, p. 280–293, abr. 2013.

RILEY, J. H. et al. The use of single nucleotide polymorphisms in the isolation of common disease genes. **Pharmacogenomics**, v. 1, n. 1, p. 39–47, fev. 2000.

ROMIGUIER, J.; ROUX, C. Analytical Biases Associated with GC-Content in Molecular Evolution. **Frontiers in Genetics**, v. 8, 15 fev. 2017.

ROMIGUIER, J. et al. Less Is More in Mammalian Phylogenomics: AT-Rich Genes Minimize Tree Conflicts and Unravel the Root of Placental Mammals. Molecular Biology and Evolution, v. 30, n. 9, p. 2134–2144, set. 2013.

ROSNER, Bernard. **Fundamentals of biostatistics.** 7. ed. Boston: Brooks/cole, Cengage Learning, 2011.

SANTOS, Fabrício R.; ORTEGA, José Miguel. Bioinformática aplicada à Genômica. In: SANTOS, Fabrício R.; ORTEGA, José Miguel. **Biowork IV.** Belo Horizonte: Usp, 2007. p. 1-21. Disponível em : https://edisciplinas.usp.br/pluginfile.php/4472700/mod_resource/content/1/Bioinform%C3%A1tica%20e%20gen%C3%B4mica.pdf>. Acesso em: 25 set. 2019.

SAUNA, Z. E. et al. Silent Polymorphisms Speak: How They Affect Pharmacogenomics and the Treatment of Cancer. **Cancer Research**, v. 67, n. 20, p. 9609–9612, 15 out. 2007.

SHERRY, S. T. dbSNP: the NCBI database of genetic variation. **Nucleic Acids Research**, v. 29, n. 1, p. 308–311, 1 jan. 2001.

SIEPEL, A. Phylogenetic Estimation of Context-Dependent Substitution Rates by Maximum Likelihood. **Molecular Biology and Evolution**, v. 21, n. 3, p. 468–488, 5 dez. 2003.

SOARES, M. P.; WEISS, G. The Iron age of host-microbe interactions. **EMBO reports**, v. 16, n. 11, p. 1482–1500, 1 nov. 2015.

VANDER WAL, E. et al. Evolutionary rescue in vertebrates: evidence, applications and uncertainty. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 368, n. 1610, p. 20120090–20120090, 3 dez. 2012.

UniProtKB - P26358 (DNMT1_HUMAN). UNIPROT CONSORTIUM**.** Disponível em: https://www.uniprot.org/uniprot/P26358 Acesso em: 01 de junho de 2019

VENTER, J. C. et al. The Sequence of the Human Genome. **THE HUMAN GENOME**, v. 291, p. 51, 2001.

WANG, D. et al. Multidrug resistance polypeptide 1 (MDR1, ABCB1) variant 3435C>T affects mRNA stability: **Pharmacogenetics and Genomics**, v. 15, n. 10, p. 693–704, out. 2005.

WATSON, J.; CRICK, F. Genetic implications of the structure of DNA. *Nature*, v. 171, p. 964–967, 1953.

What are single nucleotide polymorphisms (SNPs)? **NIH.** 11 de Jun. de 2019. Disponível em: https://ghr.nlm.nih.gov/primer/genomicresearch/snp Acesso em: 16 de junho de 2019

XUE, Y. et al. Human Y Chromosome Base-Substitution Mutation Rate Measured by Direct Sequencing in a Deep-Rooting Pedigree. **Current Biology**, v. 19, n. 17, p. 1453–1457, set. 2009.

ZHANG, F.; ZHAO, Z. The influence of neighboring-nucleotide composition on single nucleotide polymorphisms (SNPs) in the mouse genome and its comparison with human SNPs. **Genomics**, v. 84, n. 5, p. 785–795, nov. 2004.

ZHANG, J. A high-resolution multistrain haplotype analysis of laboratory mouse genome reveals three distinctive genetic variation patterns. **Genome Research**, v. 15, n. 2, p. 241–249, 14 jan. 2005.

ZHAO, H. et al. Neighboring-Nucleotide Effects on the Mutation Patterns of the Rice Genome. **Genomics, Proteomics & Bioinformatics**, v. 3, n. 3, p. 158–168, 2005.

ZHAO, Z. Neighboring-Nucleotide Effects on Single Nucleotide Polymorphisms: A Study of 2.6 Million Polymorphisms Across the Human Genome. **Genome Research**, v. 12, n. 11, p. 1679–1686, 1 nov. 2002.

ZHAO, Z.; ZHANG, F. Sequence context analysis of 8.2 million single nucleotide polymorphisms in the human genome. **Gene**, v. 366, n. 2, p. 316–324, fev. 2006.

APÊNDICE A – TABELAS E FIGURAS SUPLEMENTARES

Tabela Suplementar 1: Categorias com diferenças significativas no teste qui-quadrado

Organismo:	Substituição:	Região:	Posição	p-value:
Bos taurus	AC	3utr	1	0,014301819626139
Bos taurus	AC	5utr	-5	0,008725192134267
Bos taurus	AC	5utr	-4	0,008593513140872
Bos taurus	AC	5utr	-3	0,011099253191684
Bos taurus	AC	5utr	-2	0,020849623512987
Bos taurus	AC	5utr	-1	0,001159762373932
Bos taurus	AC	5utr	1	0,000003491018775
Bos taurus	AC	5utr	2	0,002853722390027
Bos taurus	AC	5utr	3	0,011903401485624
Bos taurus	AC	5utr	4	0,006547480599961
Bos taurus	AC	cds	-3	0,025698652784768
Bos taurus	AC	cds	-1	0,000012230373485
Bos taurus	AC	cds	1	0,000332997008178
Bos taurus	AC	cds	3	0,046571095981199
Bos taurus	AG	cds	-1	0,008469625387036
Bos taurus	AG	intron	-1	0,001970210785803
Bos taurus	AT	cds	-1	0,000552182114891
Bos taurus	AT	cds	1	0,002062211723393
Bos taurus	CG	5utr	-5	0,003917163022821
Bos taurus	CG	5utr	-4	0,002283461781166
Bos taurus	CG	5utr	-3	0,001452871743482
Bos taurus	CG	5utr	-2	0,002678053464672
Bos taurus	CG	5utr	-1	0,000003352411125
Bos taurus	CG	5utr	1	0,000007153334412
Bos taurus	CG	5utr	2	0,003770455127711
Bos taurus	CG	5utr	3	0,001452871743482
Bos taurus	CG	5utr	4	0,002526391670275
Bos taurus	CG	cds	-3	0,007511656974946
Bos taurus	CG	cds	-2	0,018692735980030
Bos taurus	CG	cds	-1	0,000101632384431
Bos taurus	CG	cds	1	0,000114536846912
Bos taurus	CG	cds	3	0,005059413481620
Bos taurus	CT	5utr	-5	0,028298609892171
Bos taurus	CT	5utr	-4	0,035878771487571
Bos taurus	CT	5utr	-3	0,041798203060217
Bos taurus	CT	5utr	-2	0,043827230529726
Bos taurus	CT	5utr	-1	0,006293701148216
Bos taurus	CT	5utr	1	0,000386625954944
Bos taurus	CT	5utr	2	0,018438447537734
Bos taurus	CT	5utr	3	0,022890354777037
Bos taurus	CT	5utr	4	0,031754544394175
Bos taurus	CT	cds	1	0,010511830440889
Bos taurus	GT	5utr	-5	0,005338777529854
Bos taurus	GT	5utr	-4	0,003024745423769
Bos taurus	GT	5utr	-3	0,007655338395275

Tabela Suplementar 1 (Continuação): Categorias com diferenças significativas no teste qui-quadrado

Organismo:	Substituição:	Região:	Posição	p-value:
Bos taurus	GT	5utr	-2	0,001811825515736
Bos taurus	GT	5utr	-1	0,000000828113171
Bos taurus	GT	5utr	1	0,001182794977601
Bos taurus	GT	5utr	2	0,020849623512987
Bos taurus	GT	5utr	3	0,008336365347058
Bos taurus	GT	5utr	4	0,013050414734549
Bos taurus	GT	cds	-2	0,028981572897462
Bos taurus	GT	cds	-1	0,000475166740297
Bos taurus	GT	cds	1	0,000012230373485
Bos taurus	GT	cds	3	0,031578432194558
Bos taurus	GT	intron	-1	0,022043080174032
Homo sapiens	AG	3utr	-1	0,045262106977130
Homo sapiens	AT	5utr	-1	0,016621920766705
Homo sapiens	AT	5utr	1	0,016621920766705
Mus Musculus	AC	cds	-1	0,002765340995399
Mus Musculus	AC	cds	1	0,000000002497297
Mus Musculus	AG	3utr	-1	0,015515481898049
Mus Musculus	AG	5utr	-1	0,000360477454270
Mus Musculus	AT	cds	-1	0,000018446374716
Mus Musculus	AT	cds	1	0,000112725367045
Mus Musculus	CG	5utr	-1	0,000001297644314
Mus Musculus	CG	5utr	1	0,000000367704910
Mus Musculus	CG	cds	-1	0,000000000061714
Mus Musculus	CG	cds	1	0,000000000437821
Mus Musculus	CT	3utr	1	0,011046088301348
Mus Musculus	GT	5utr	-1	0,002398460948145
Mus Musculus	GT	cds	-1	0,000000000046966
Mus Musculus	GT	cds	1	0,006187046517847
Rattus norvegicus	AC	cds	-1	0,000611144362349
Rattus norvegicus	AC	cds	1	0,000003225951177
Rattus norvegicus	AG	3utr	-1	0,001028445718744
Rattus norvegicus	AG	5utr	-1	0,000400367047142
Rattus norvegicus	AT	cds	-1	0,000080424224086
Rattus norvegicus	AT	cds	1	0,000162860715184
Rattus norvegicus	CG	cds	-1	0,000000000628267
Rattus norvegicus	CG	cds	1	0,000000000602942
Rattus norvegicus	CT	intron	1	0,001979293946460
Rattus norvegicus	GT	5utr	-1	0,046203032765315
Rattus norvegicus	GT	cds	-1	0,000004840540246
Rattus norvegicus	GT	cds	1	0,002274081856008
Sus scrofa	AG	cds	-1	0,000001171206675
Sus scrofa	CG	5utr	-1	0,002648874224529
Sus scrofa	CG	5utr	1	0,001855199509146
Sus scrofa	CT	3utr	1	0,000012174239245
Sus scrofa	CT	5utr	1	0,000022232385890
Sus scrofa	CT	cds	1	0,000001027955510
Sus scrofa	CT	intron	1	0,000905571892090

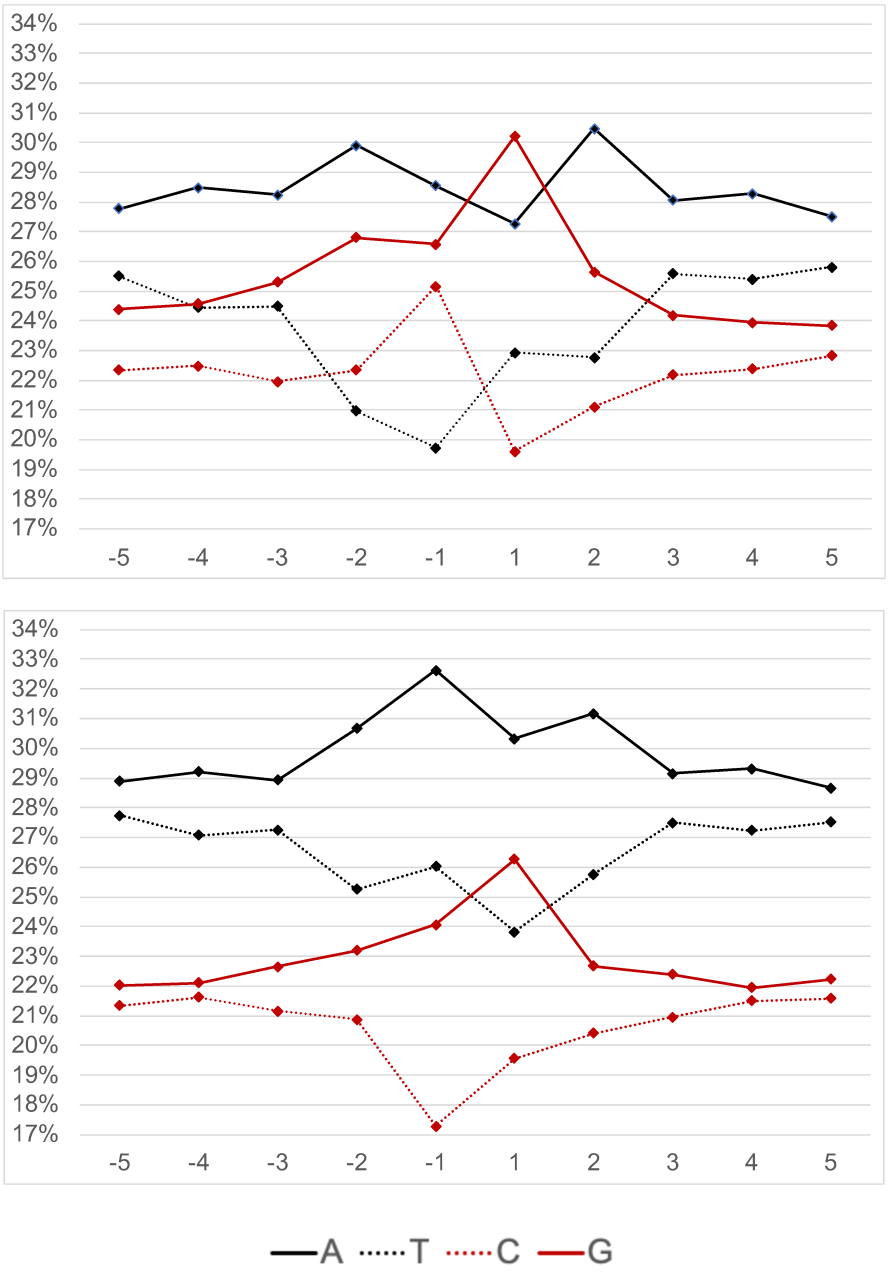


Figura Suplementar 1: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões 3'UTR em bovinos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency^.

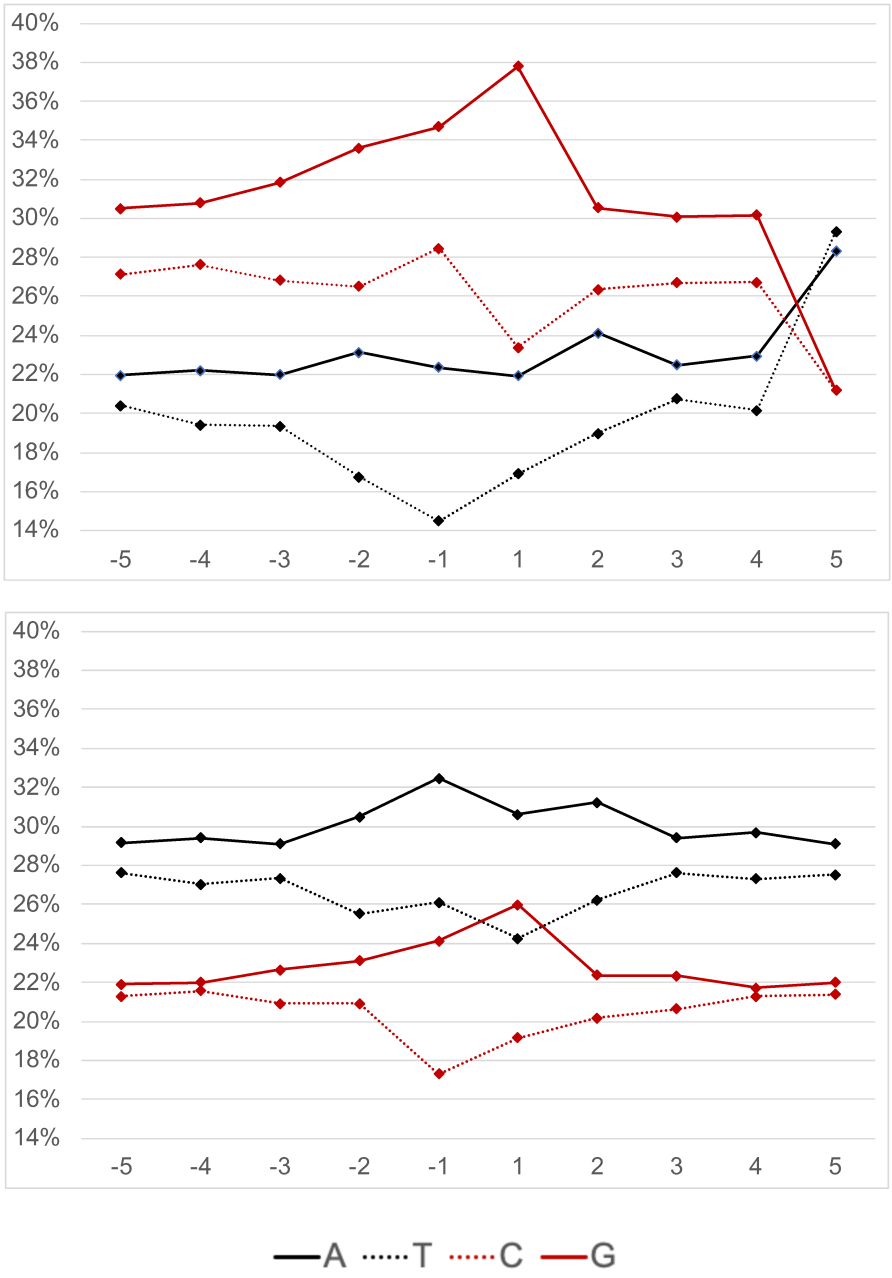


Figura Suplementar 2: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões 5'UTR em bovinos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

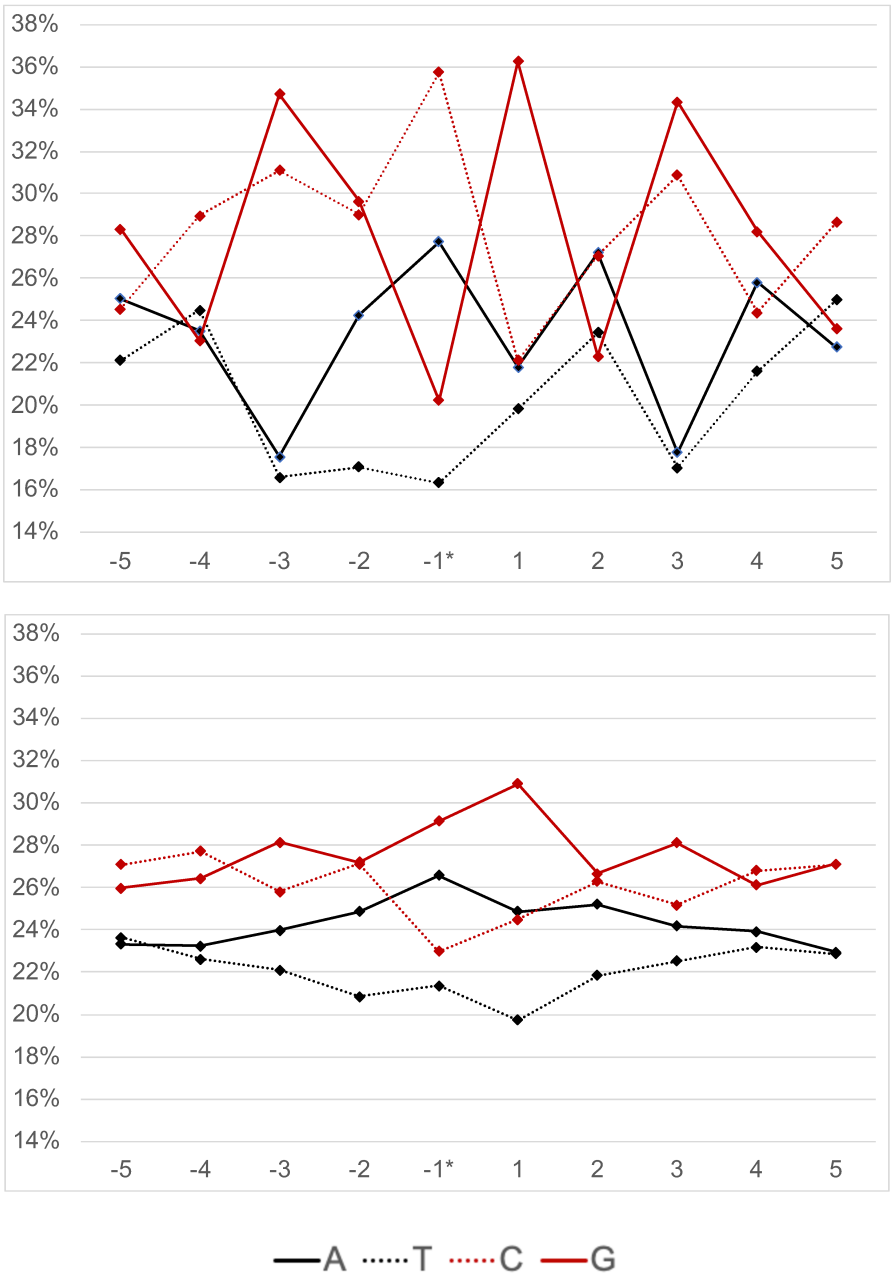


Figura Suplementar 3: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões sinônimas em bovinos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency^.

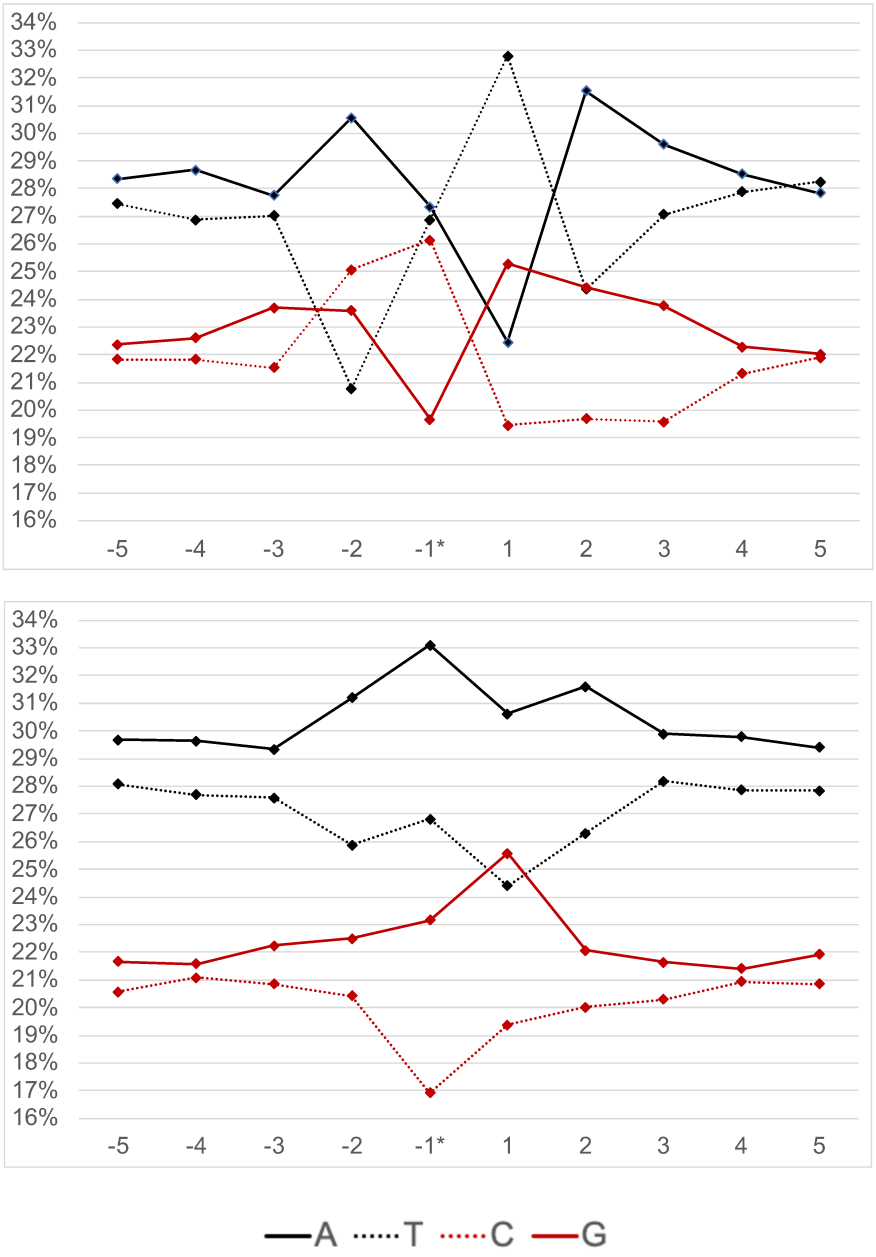


Figura Suplementar 4: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões 3'UTR em humanos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency.

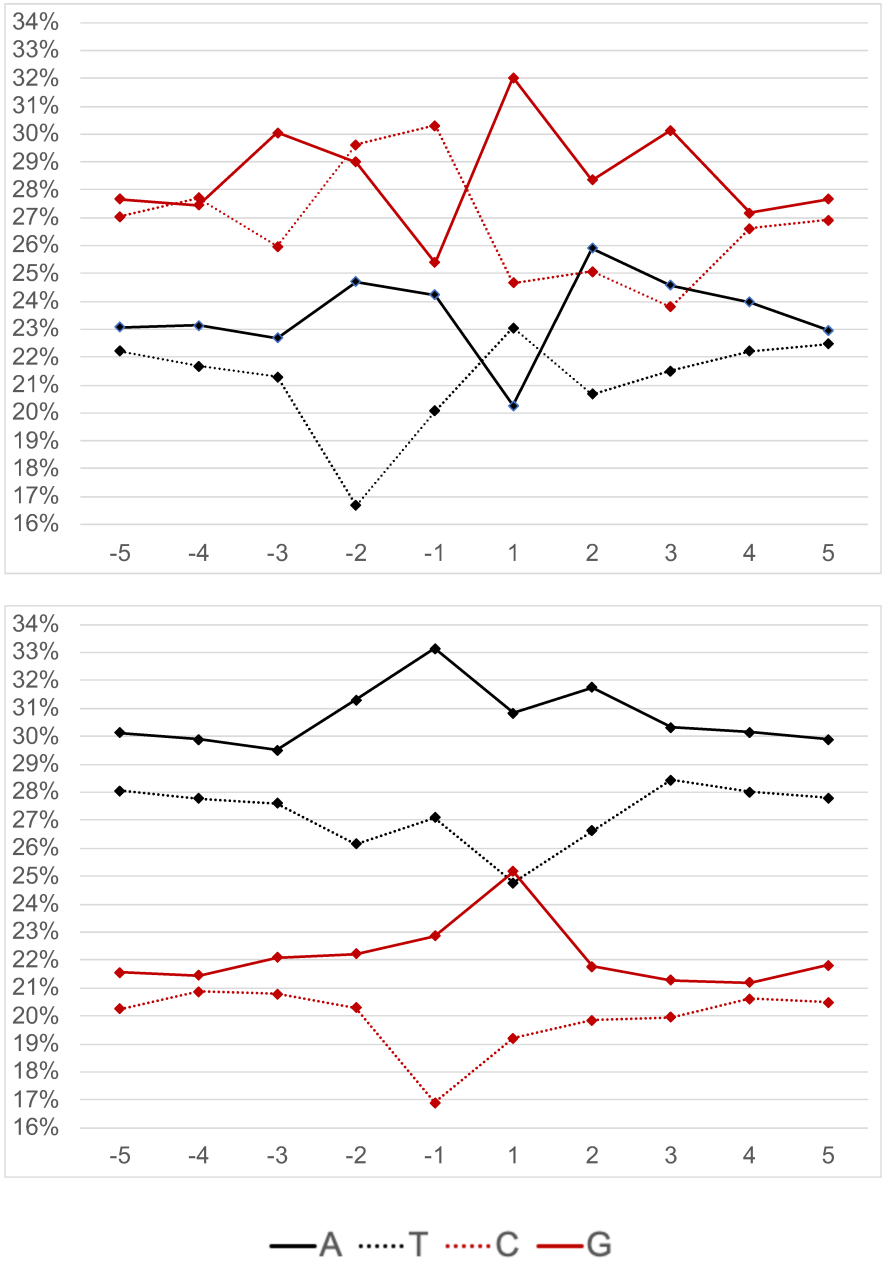


Figura Suplementar 5: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões 5'UTR em humanos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

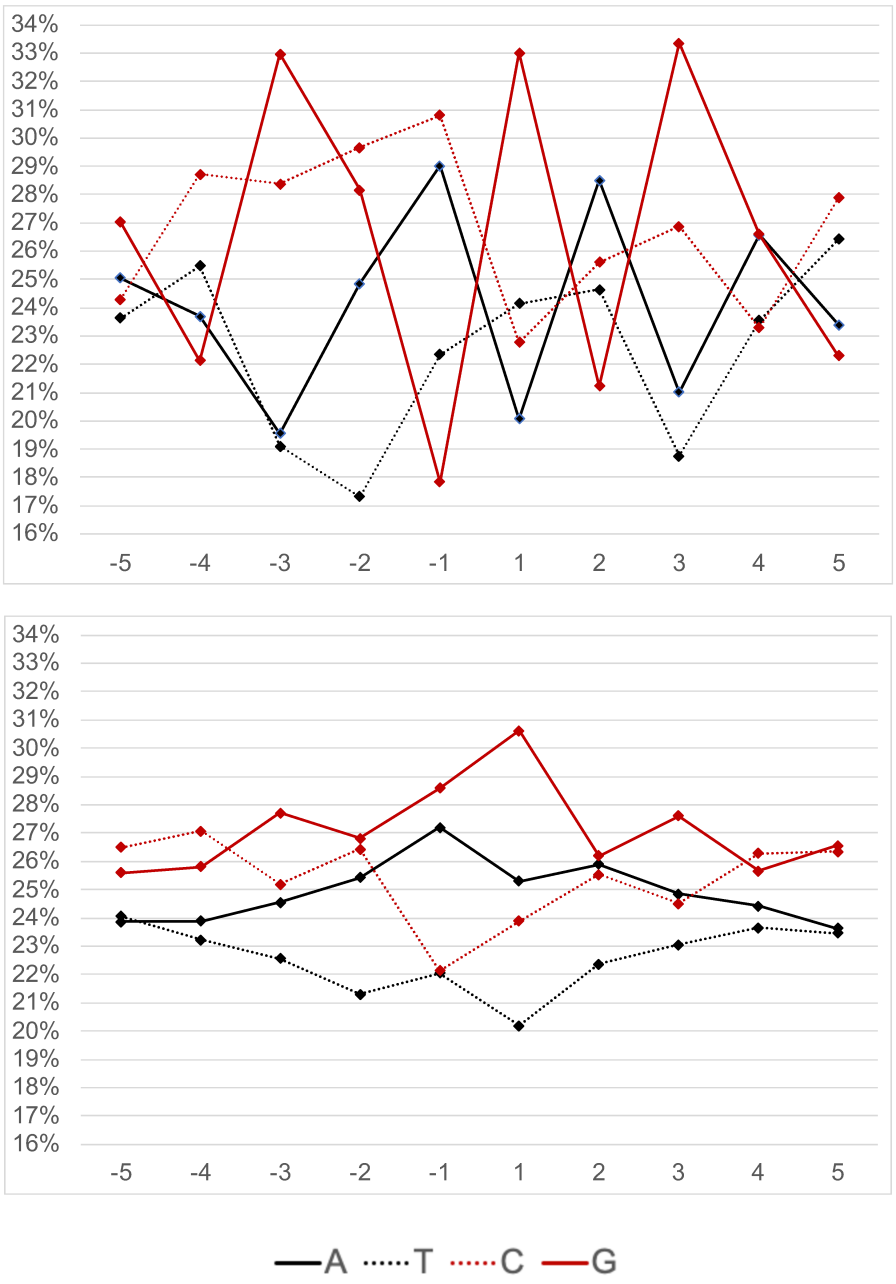


Figura Suplementar 6: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões sinônimas em humanos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

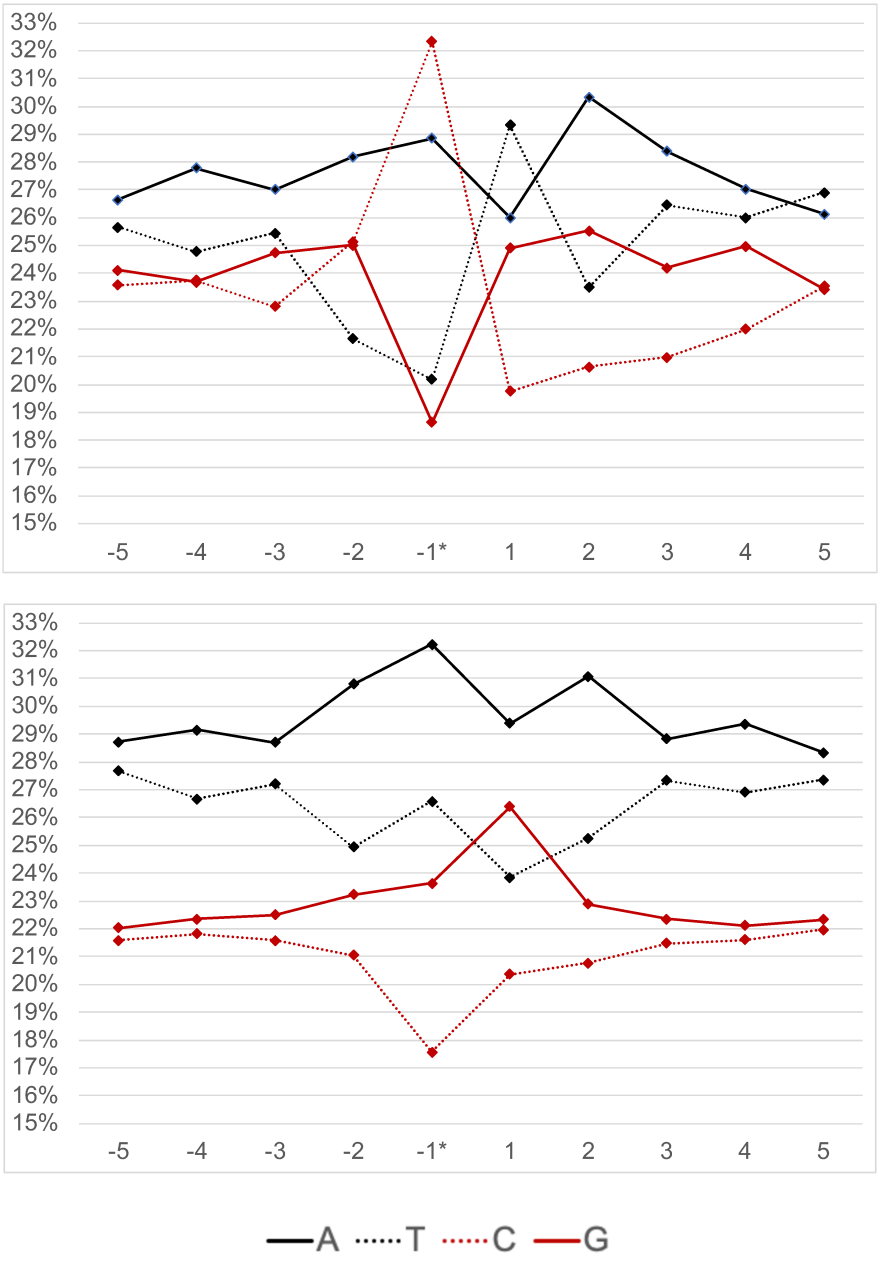


Figura Suplementar 7: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões 3'UTR em *Rattus norvegicus*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency.

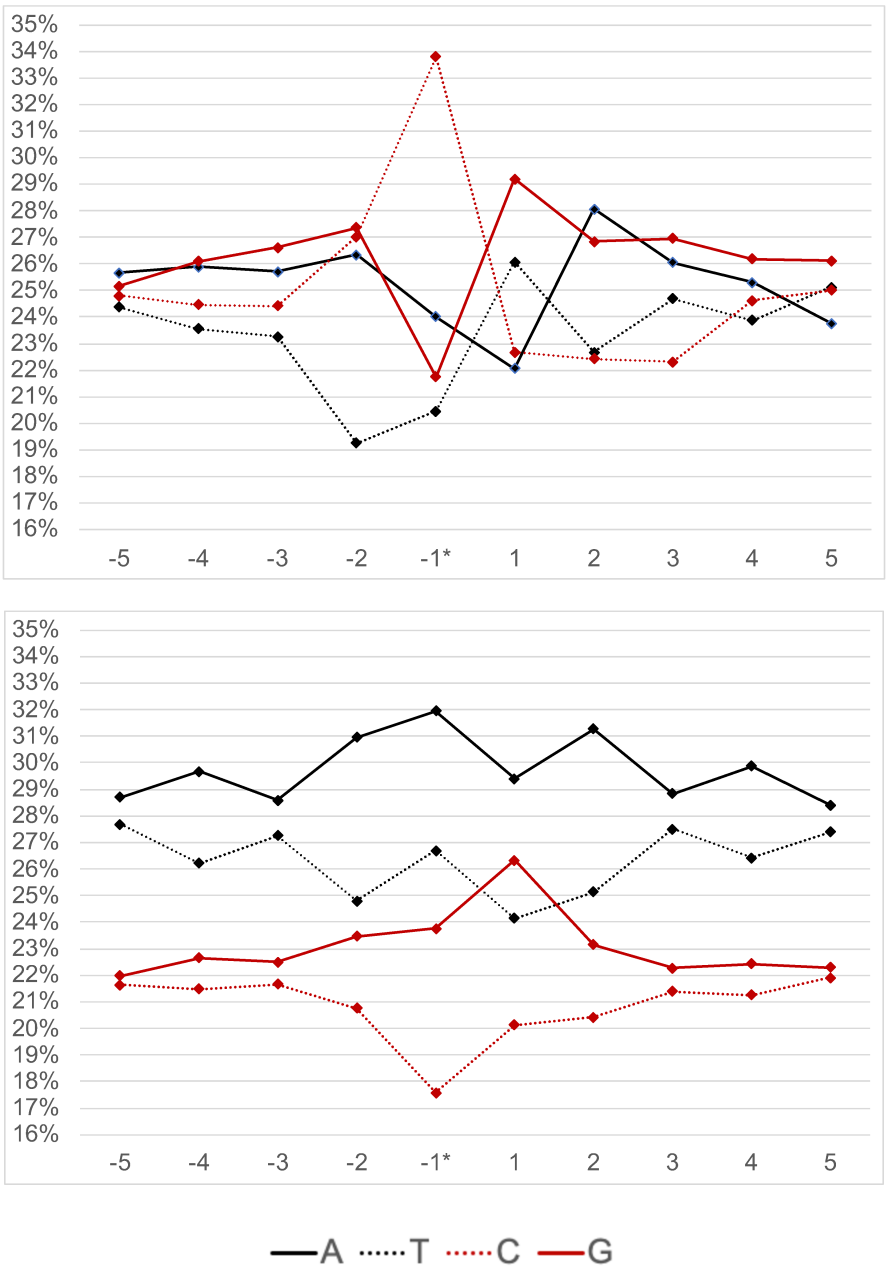


Figura Suplementar 8: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões 5'UTR em *Rattus norvegicus*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency.

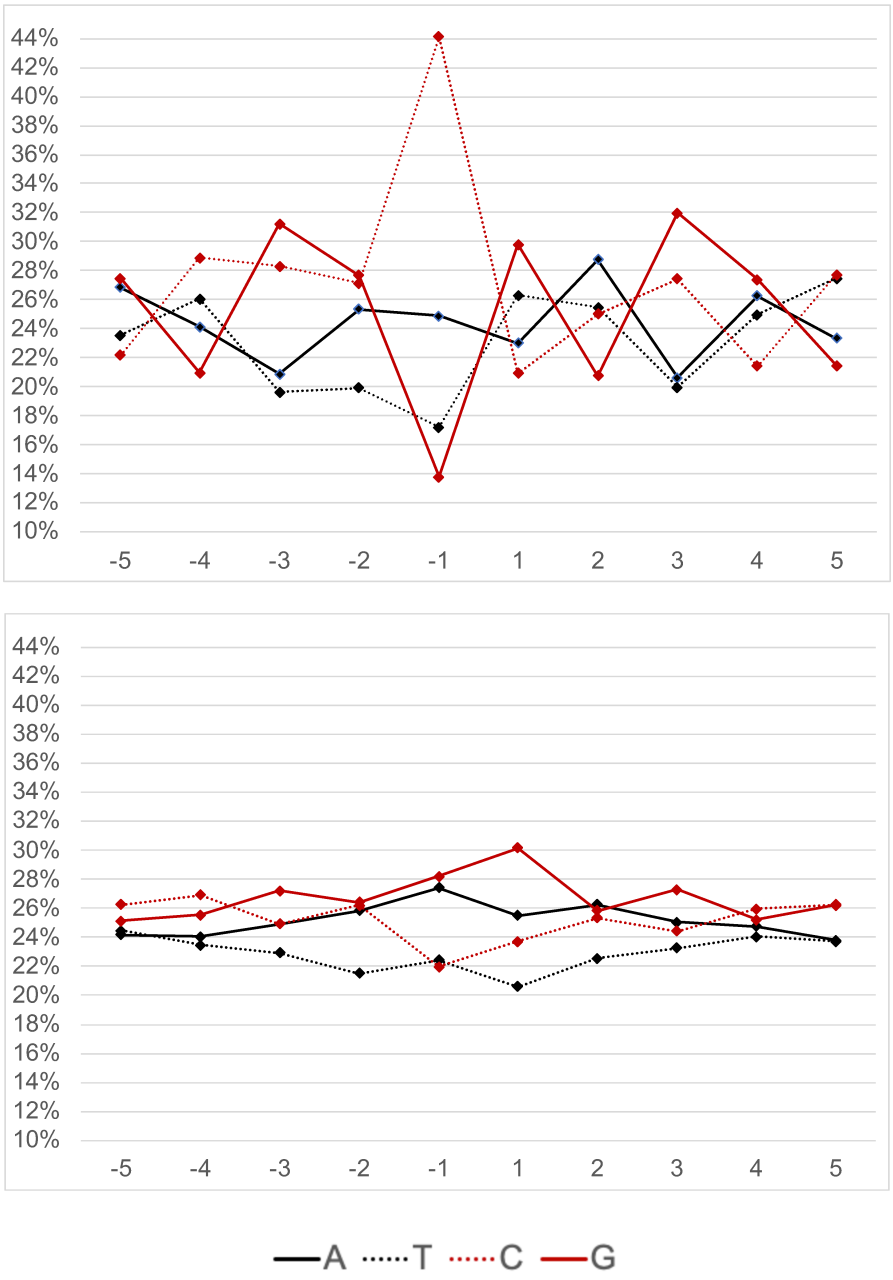


Figura Suplementar 9: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões sinônimas em *Rattus norvegicus*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

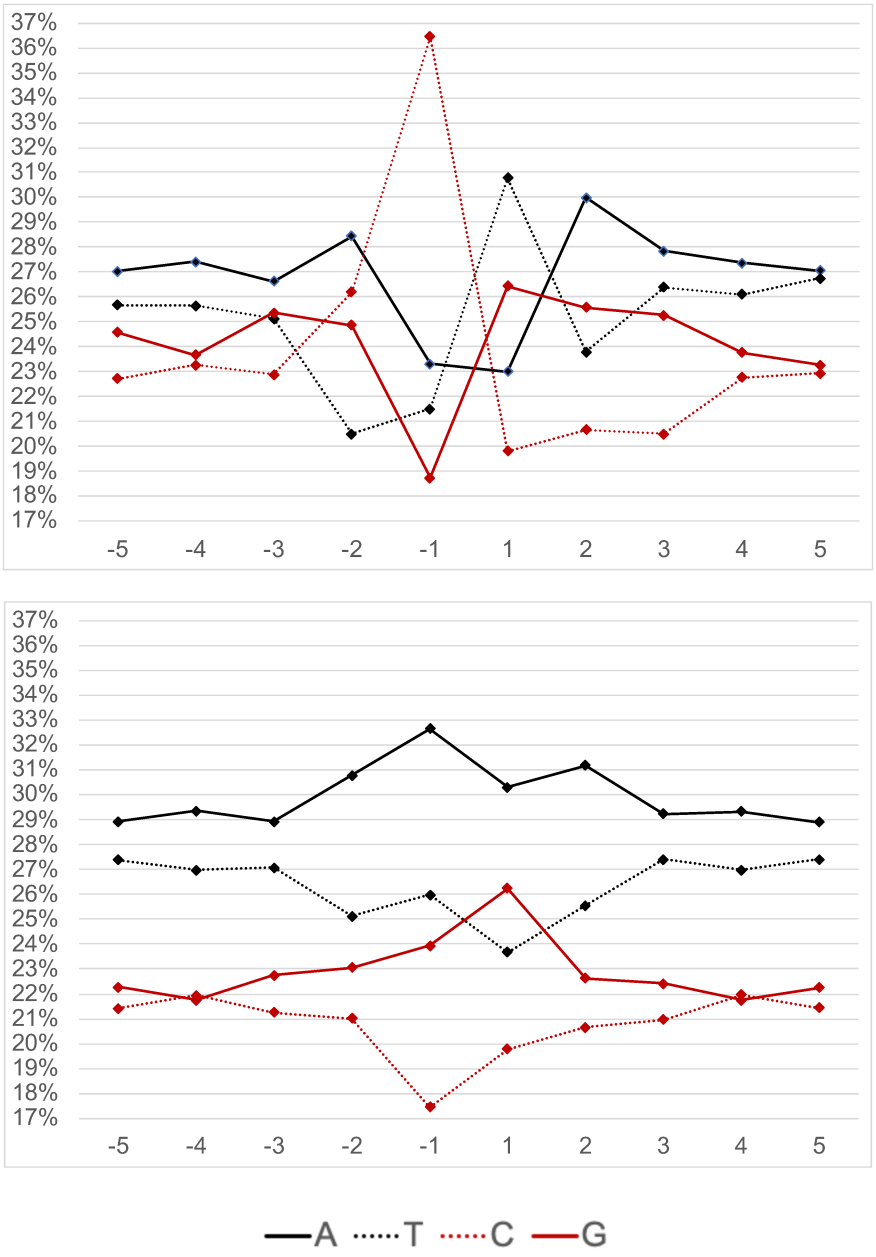


Figura Suplementar 10: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões 3'UTR em *Sus scrofa*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

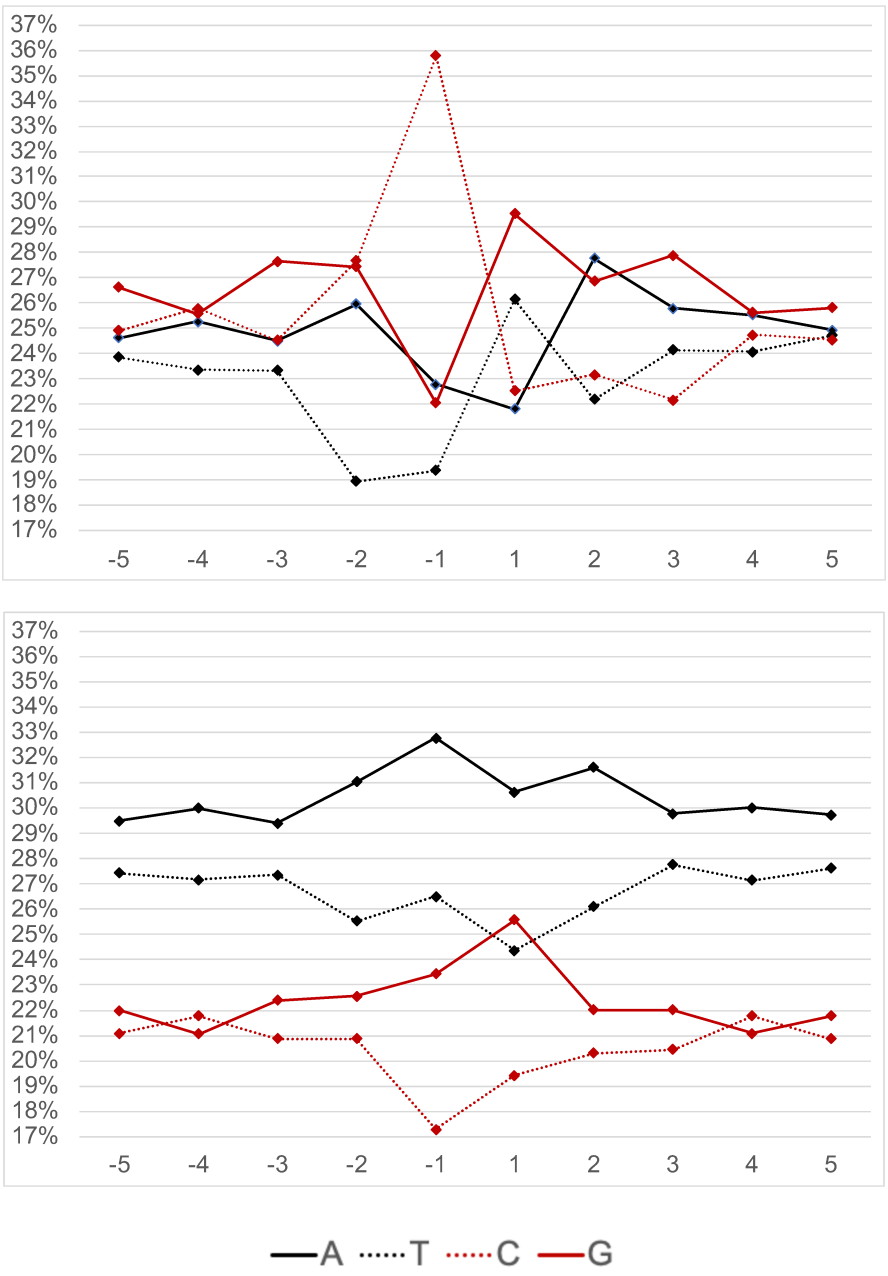


Figura Suplementar 11: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões 5'UTR em *Sus scrofa*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

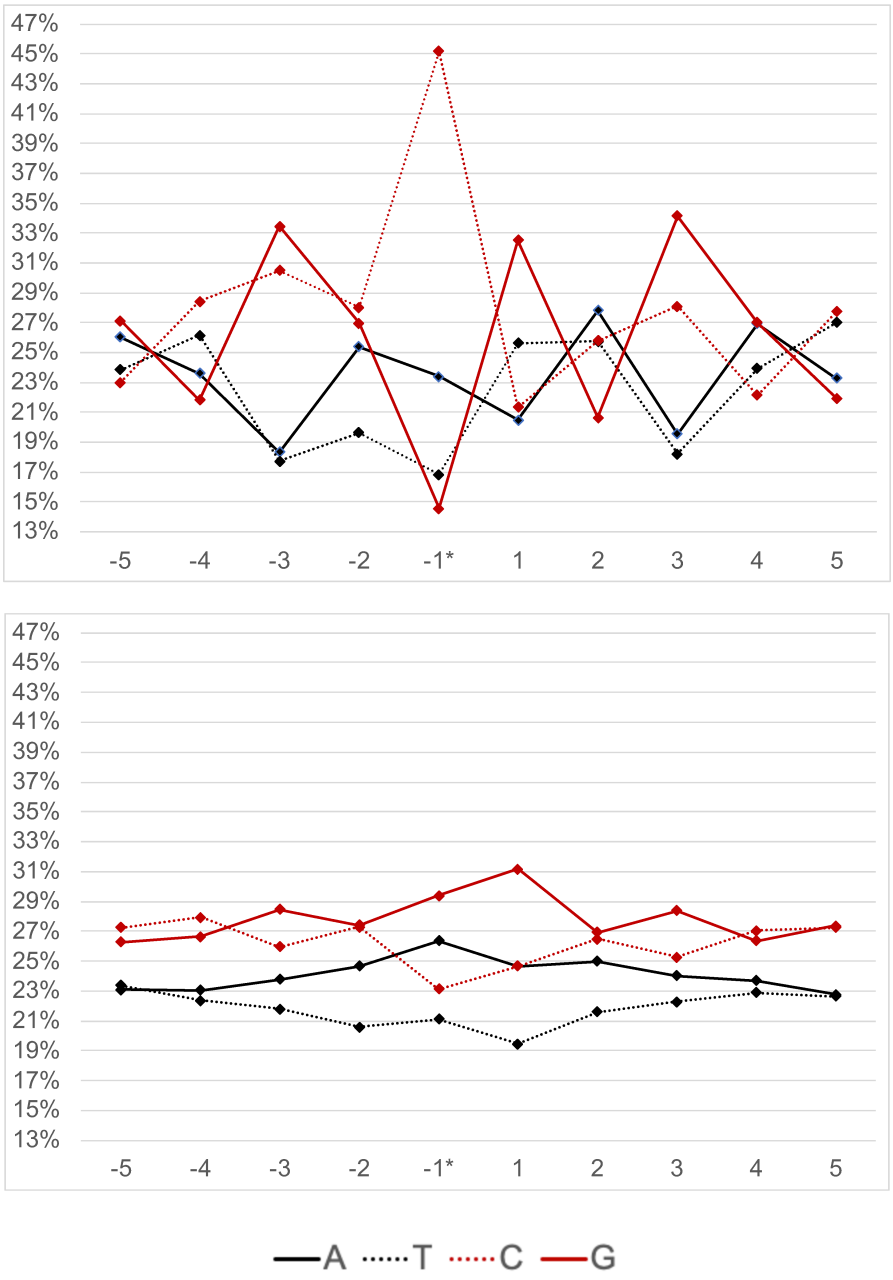


Figura Suplementar 12: Ocorrência de bases nas proximidades do SNP R (A/G) e a frequência basal esperada no genoma, nas regiões sinônimas em *Sus scrofa*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency.

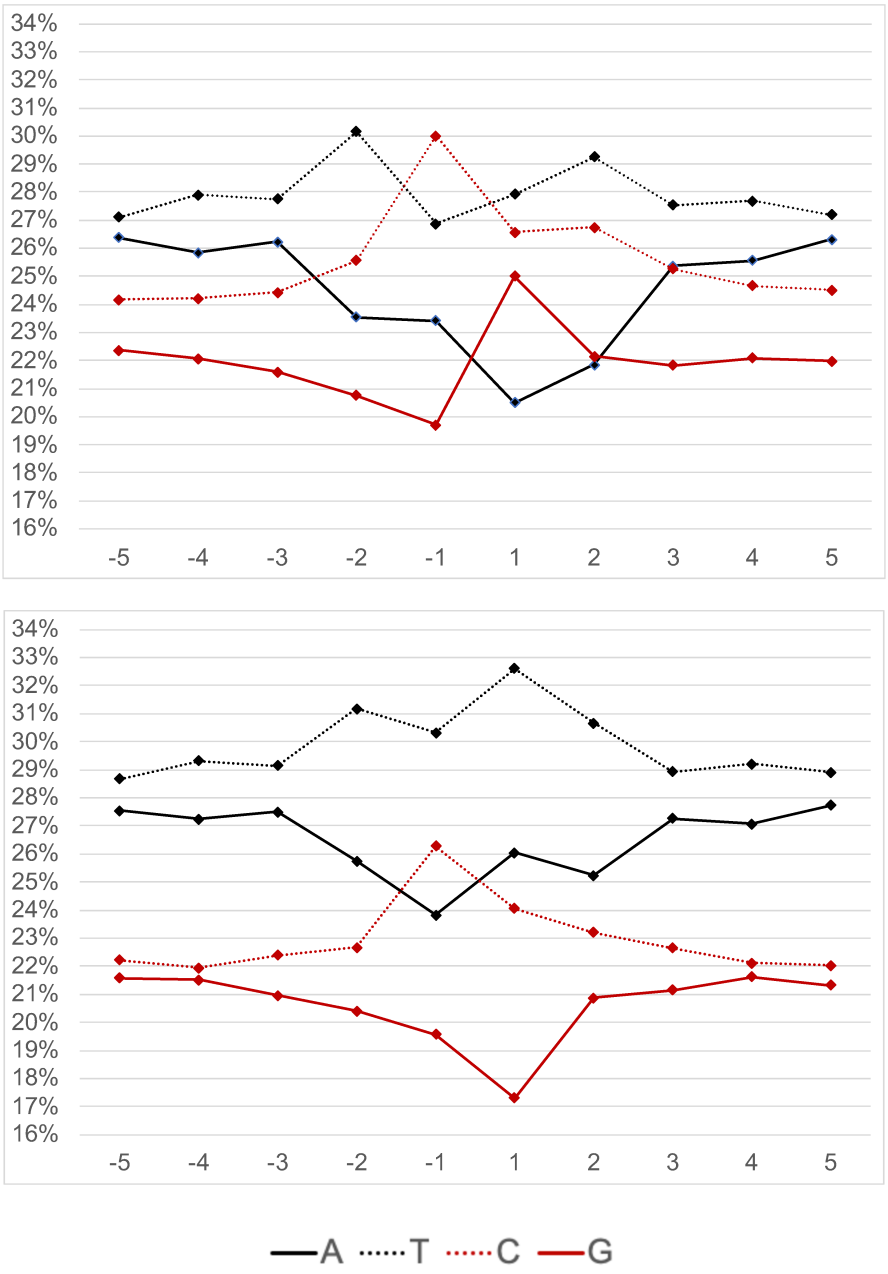


Figura Suplementar 13: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões 3'UTR em bovinos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

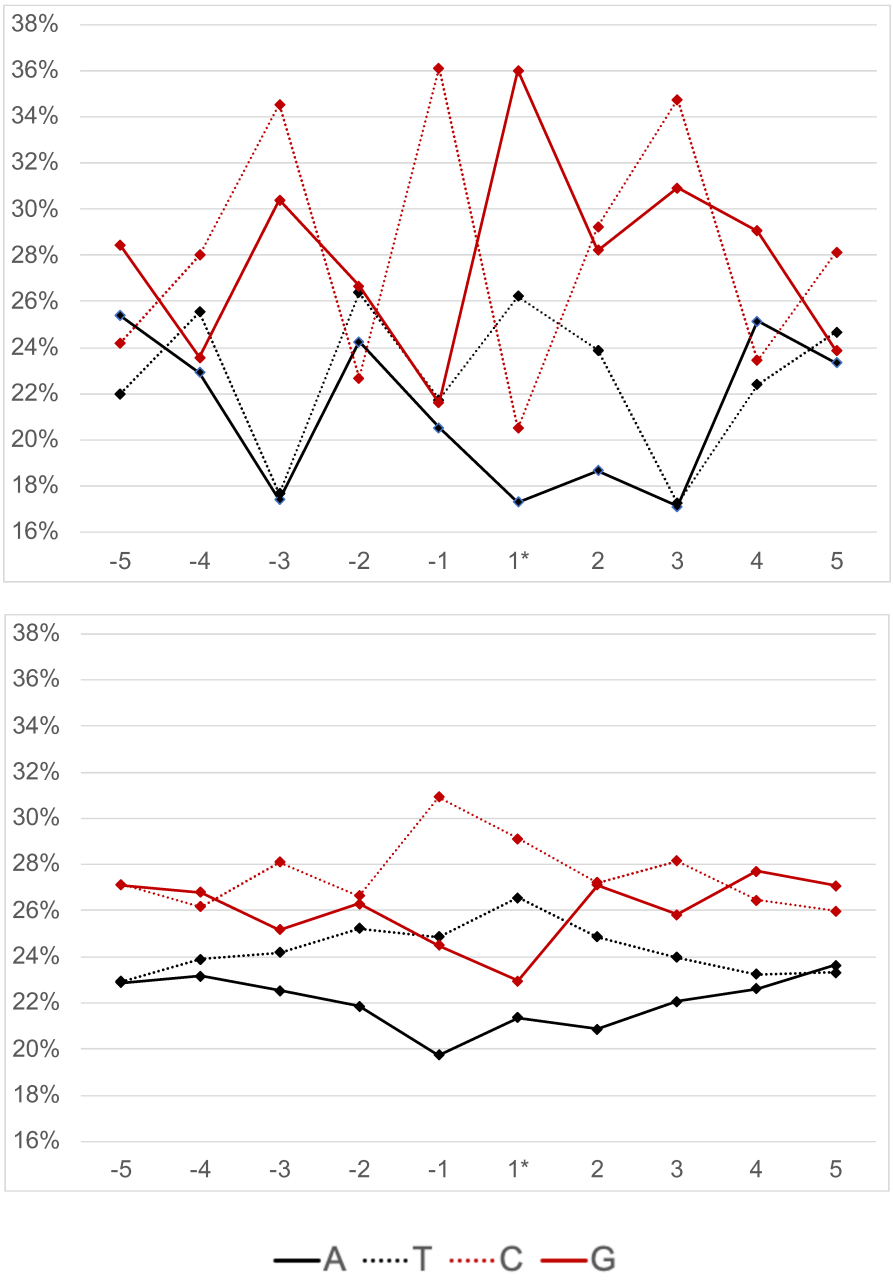


Figura Suplementar 14: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões sinônimas em bovinos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency^.

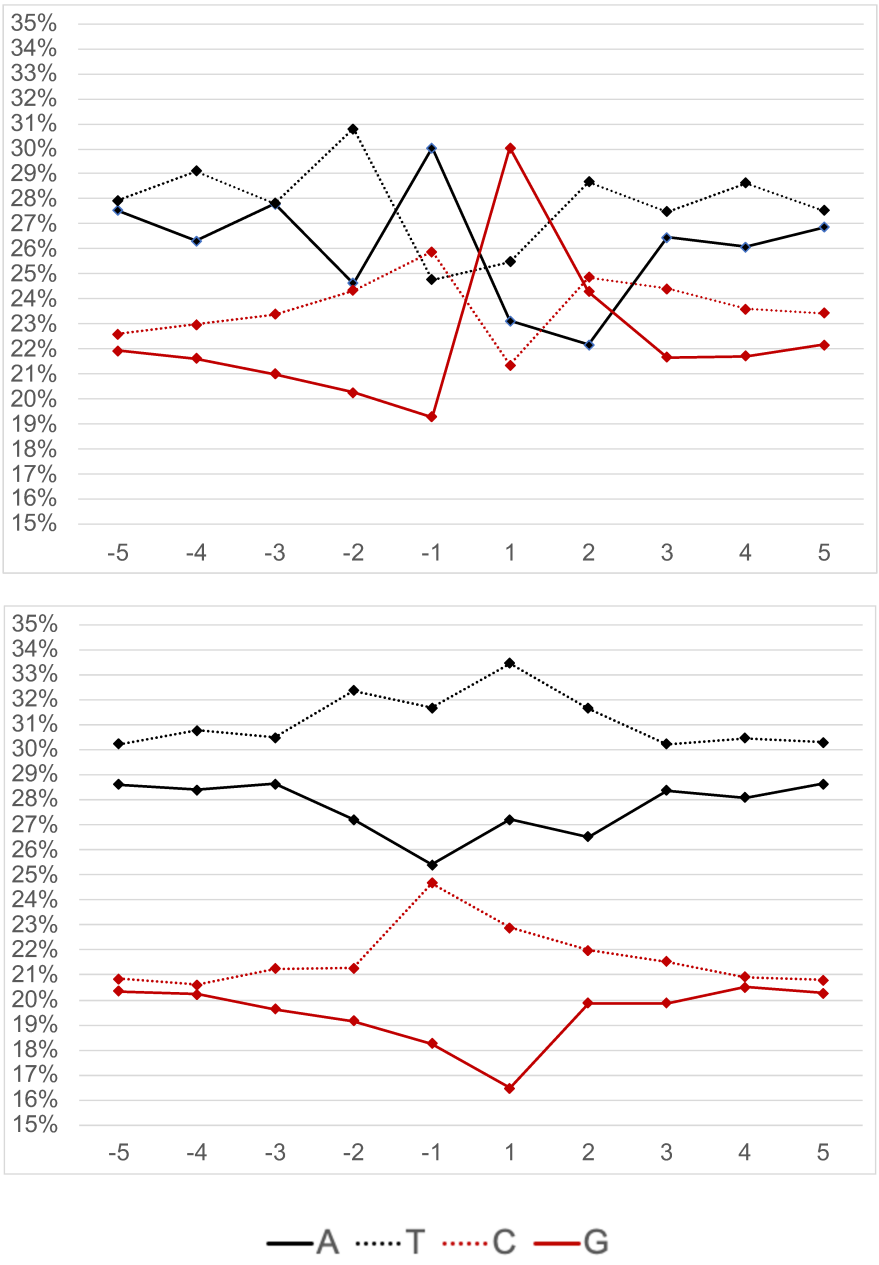


Figura Suplementar 15: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões intrônicas de bovinos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

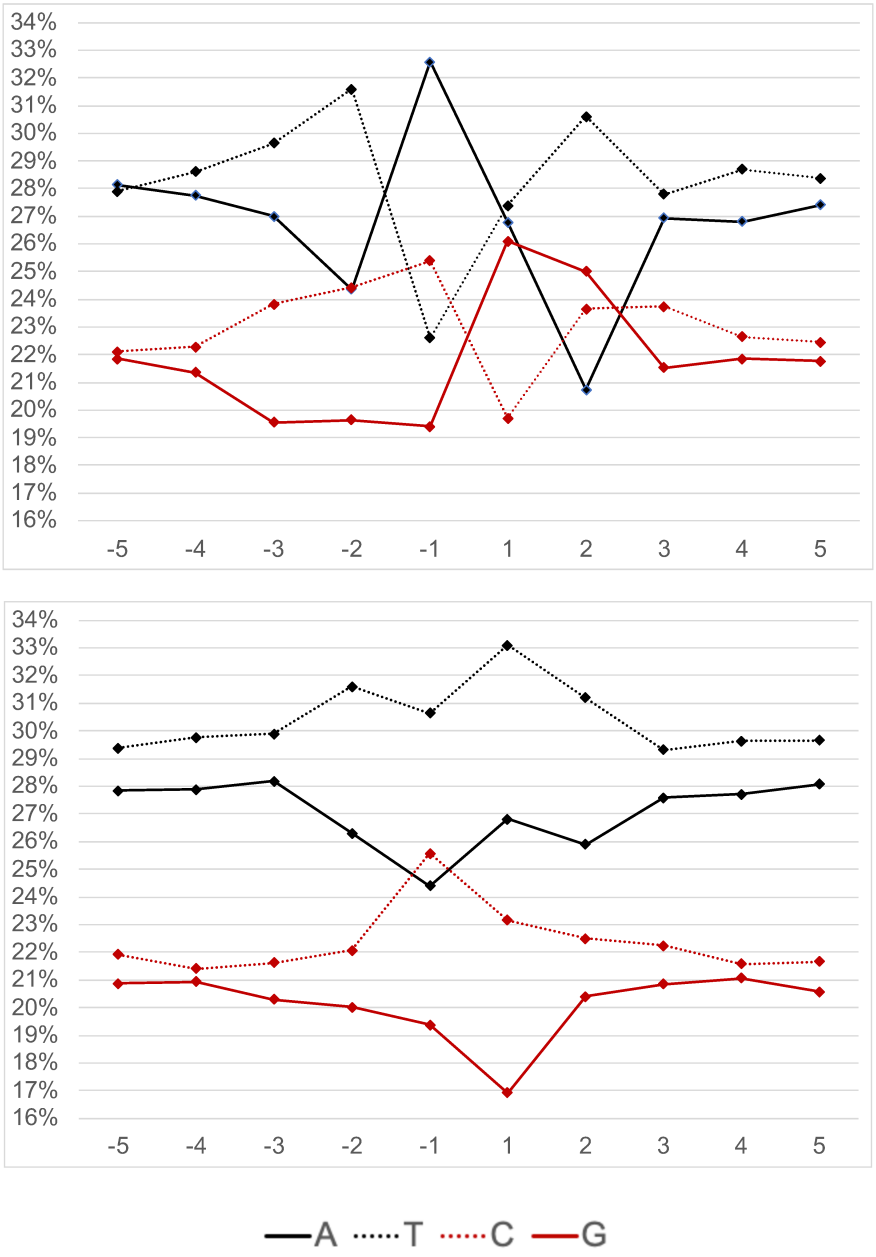


Figura Suplementar 16: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões 3'UTR em humanos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

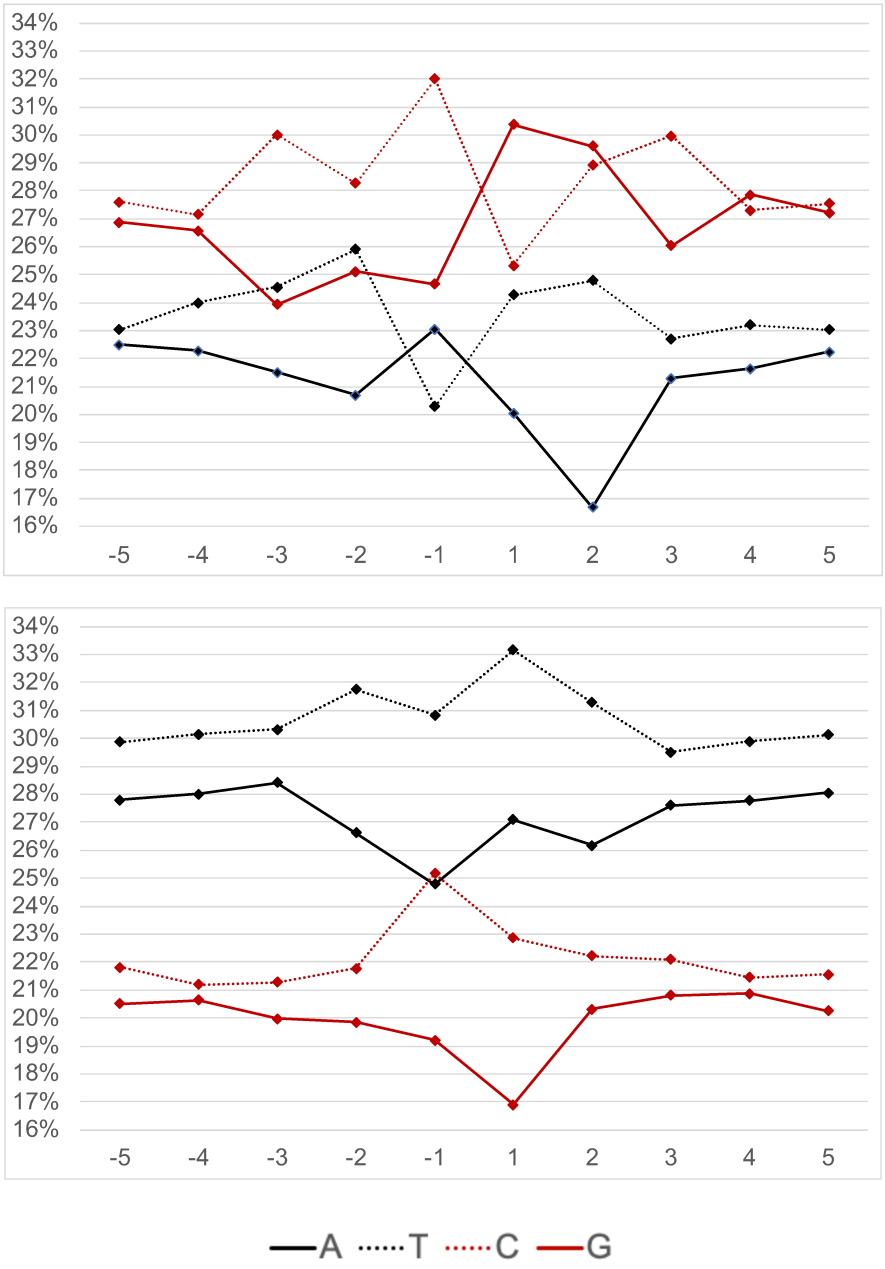


Figura Suplementar 17: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões 5'UTR em humanos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

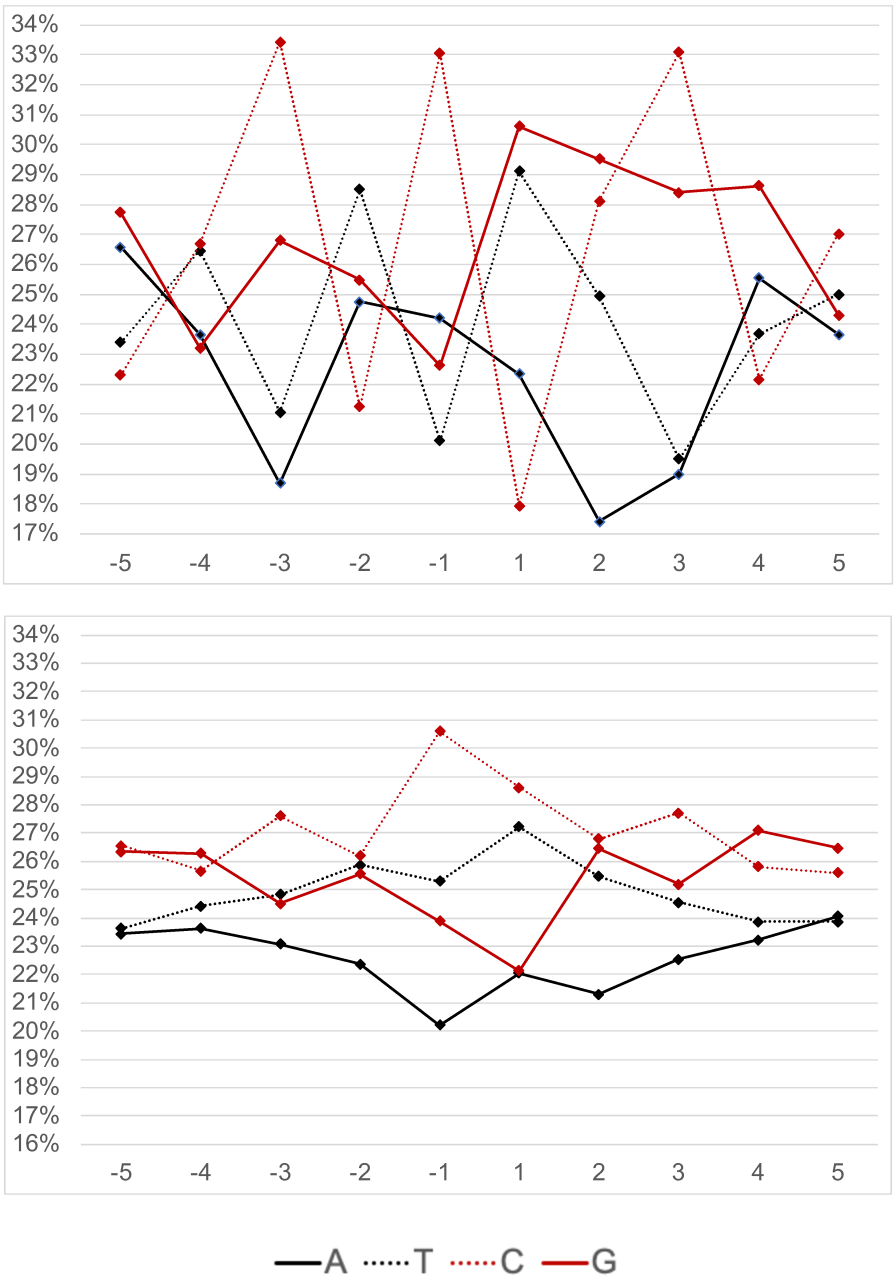


Figura Suplementar 18: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões sinônimas em humanos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

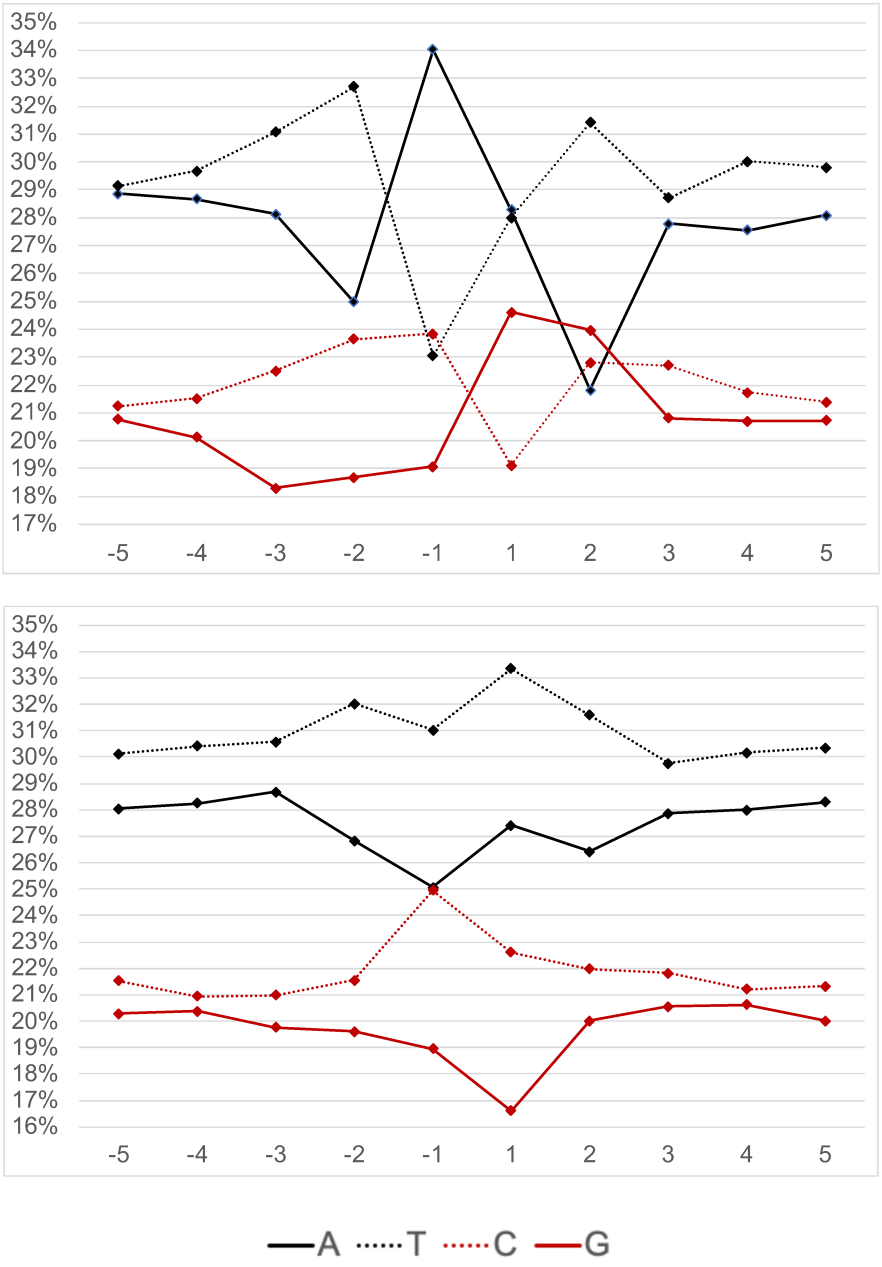


Figura Suplementar 19: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões intrônicas em humanos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

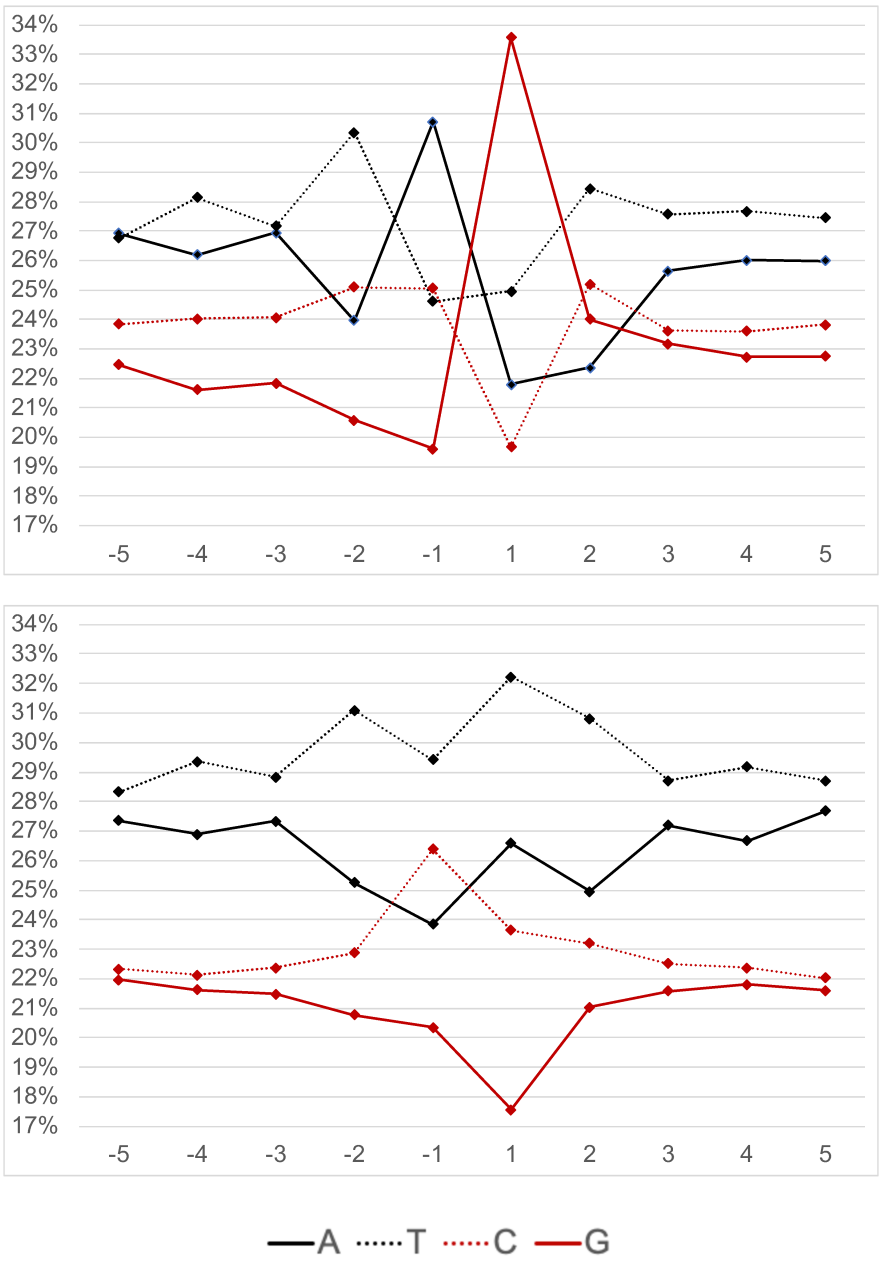


Figura Suplementar 20: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões 3'UTR em *Rattus norvegicus*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

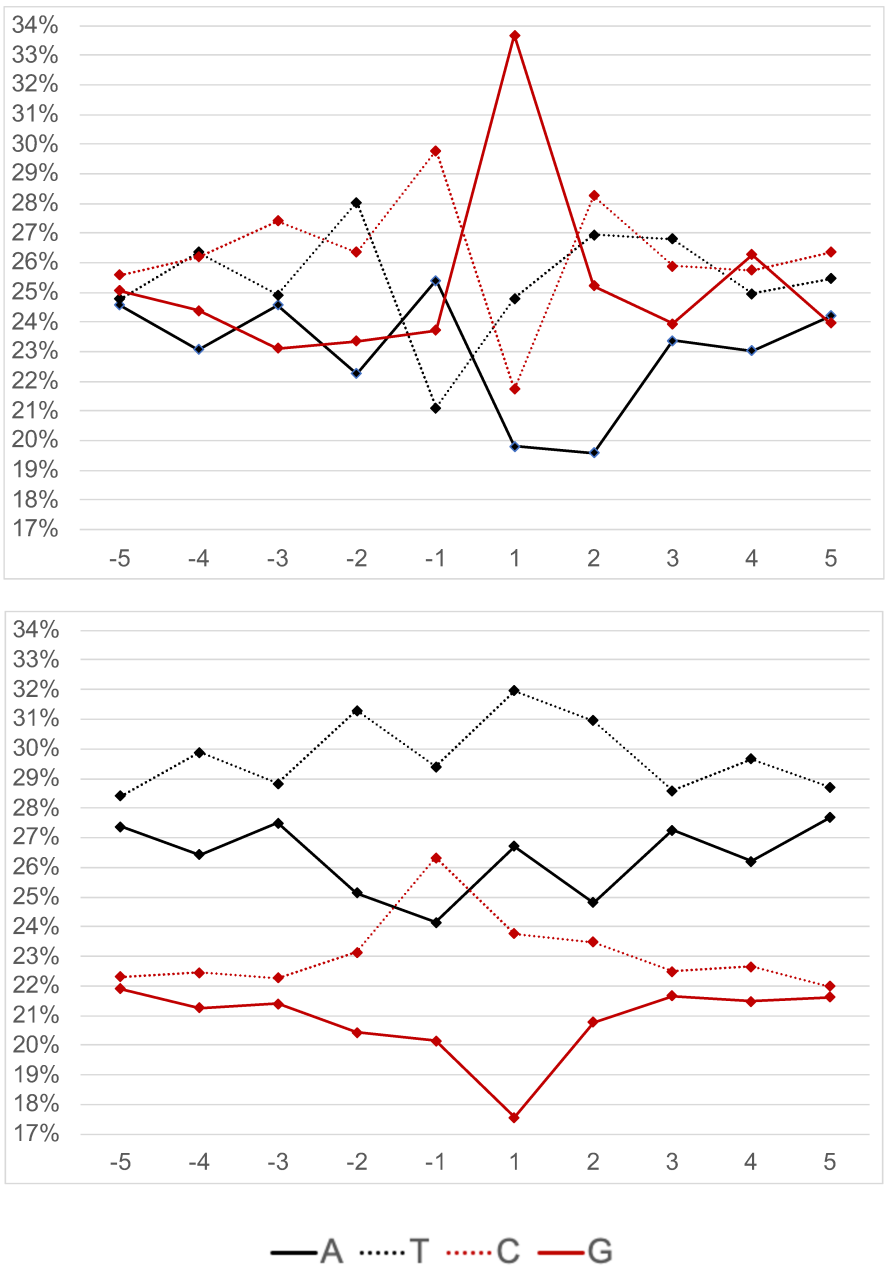


Figura Suplementar 21: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões 5'UTR em *Rattus norvegicus*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

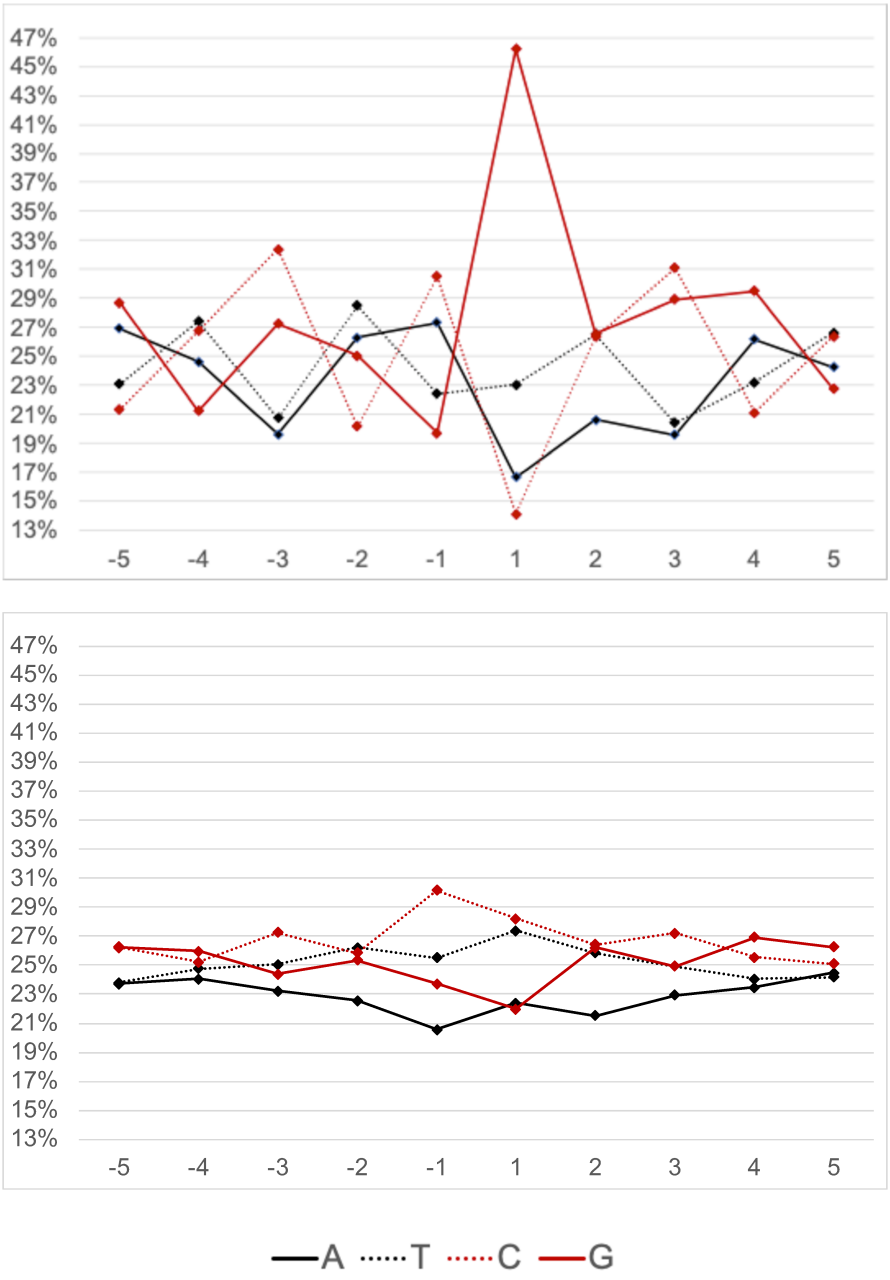


Figura Suplementar 22: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões sinônimas em *Rattus norvegicus*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.



Figura Suplementar 23: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões intrônicas em *Rattus norvegicus*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ˆbaseline frequency".

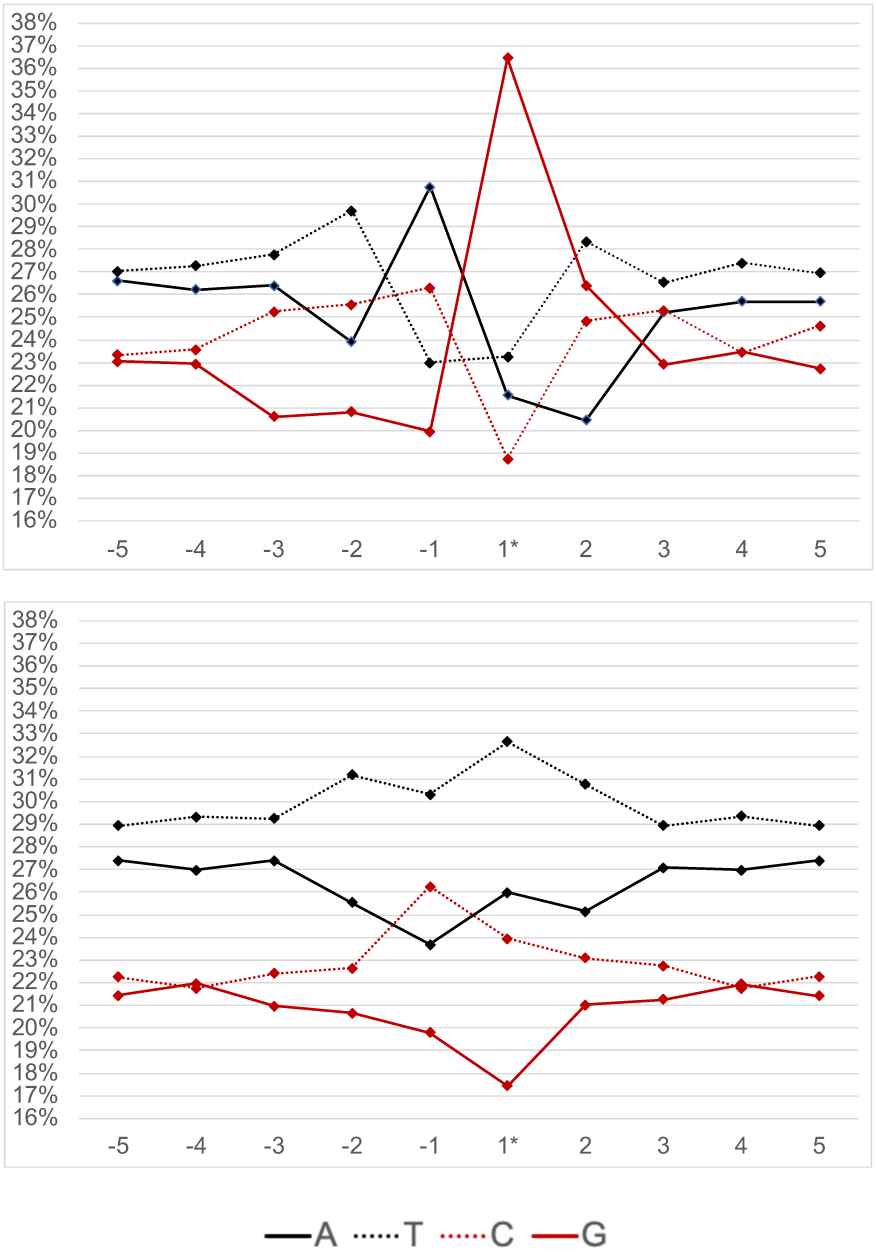


Figura Suplementar 24: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões 3'UTR em *Sus scrofa*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency".

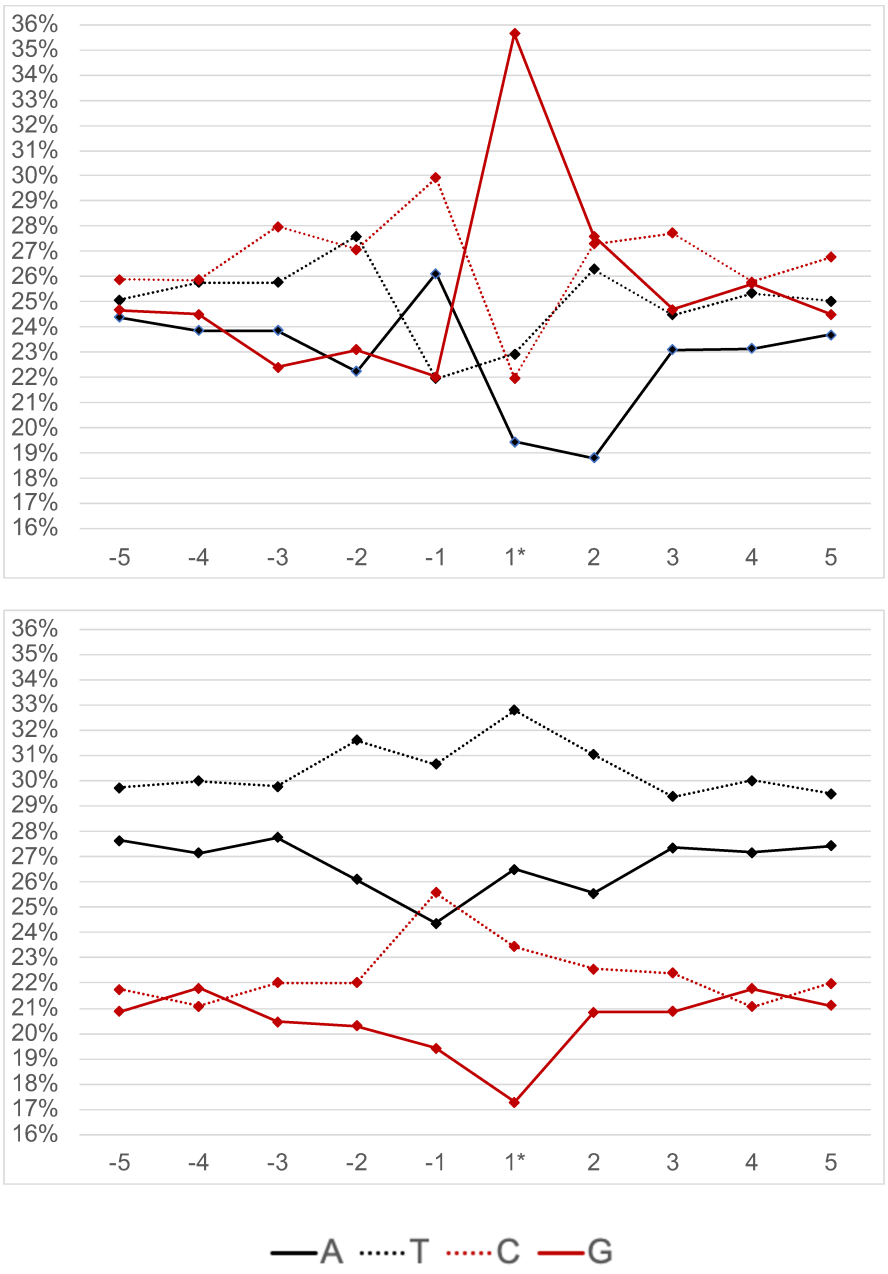


Figura Suplementar 25: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões 5'UTR em *Sus scrofa*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency".

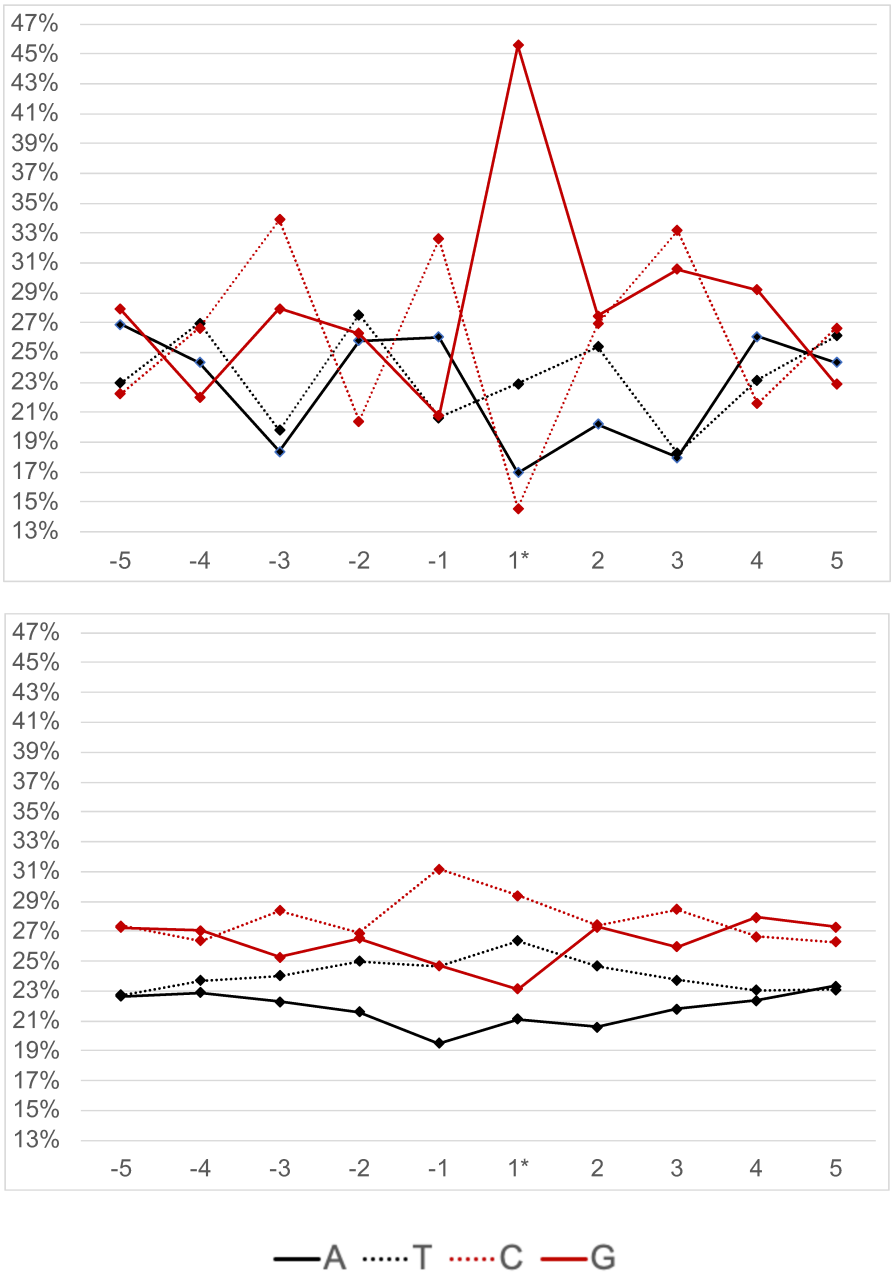


Figura Suplementar 26: Ocorrência de bases nas proximidades do SNP Y (C/T) e a frequência basal esperada no genoma, nas regiões sinônimas em *Sus scrofa*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency.

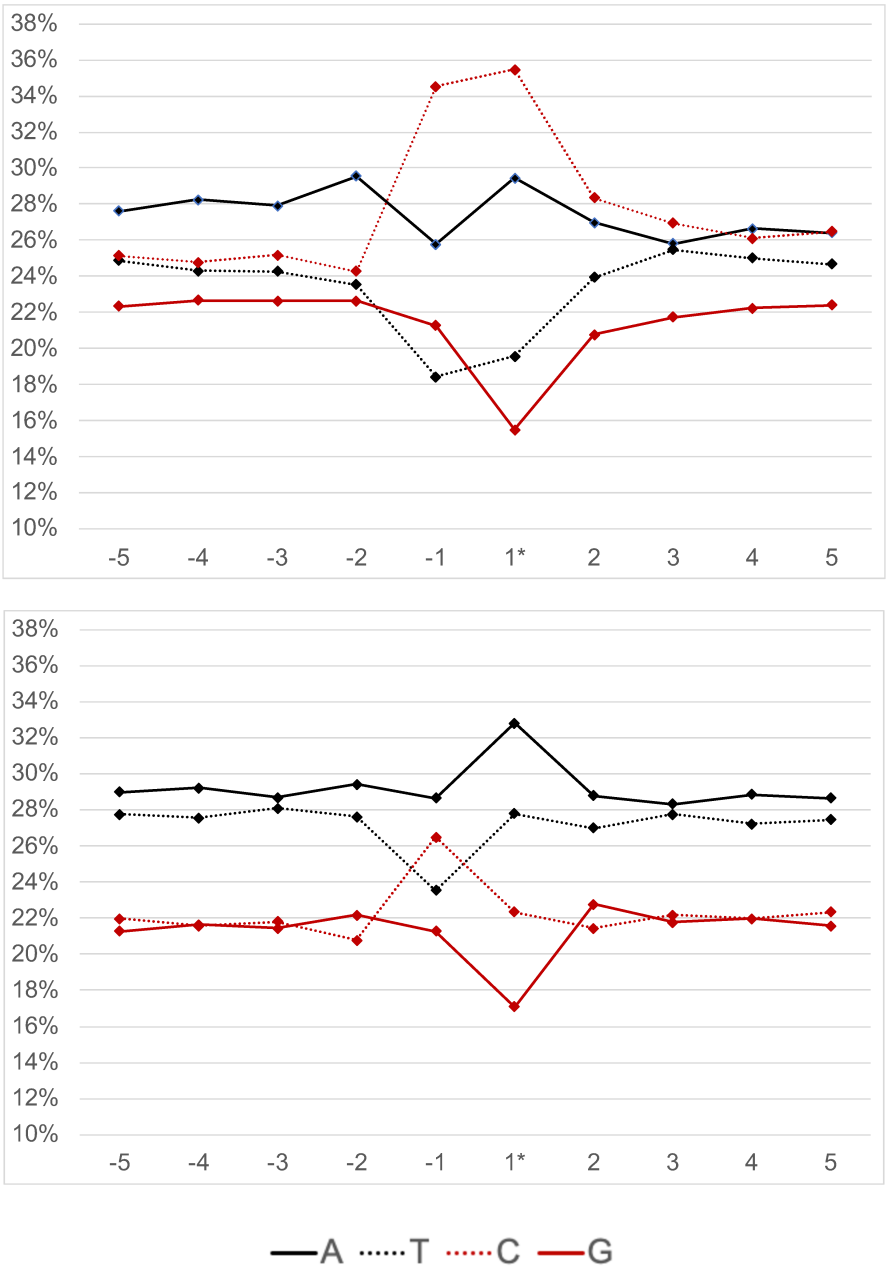


Figura Suplementar 27: Ocorrência de bases nas proximidades do SNP M (A/C) e a frequência basal esperada no genoma, nas regiões 3'UTR em bovinos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency^.

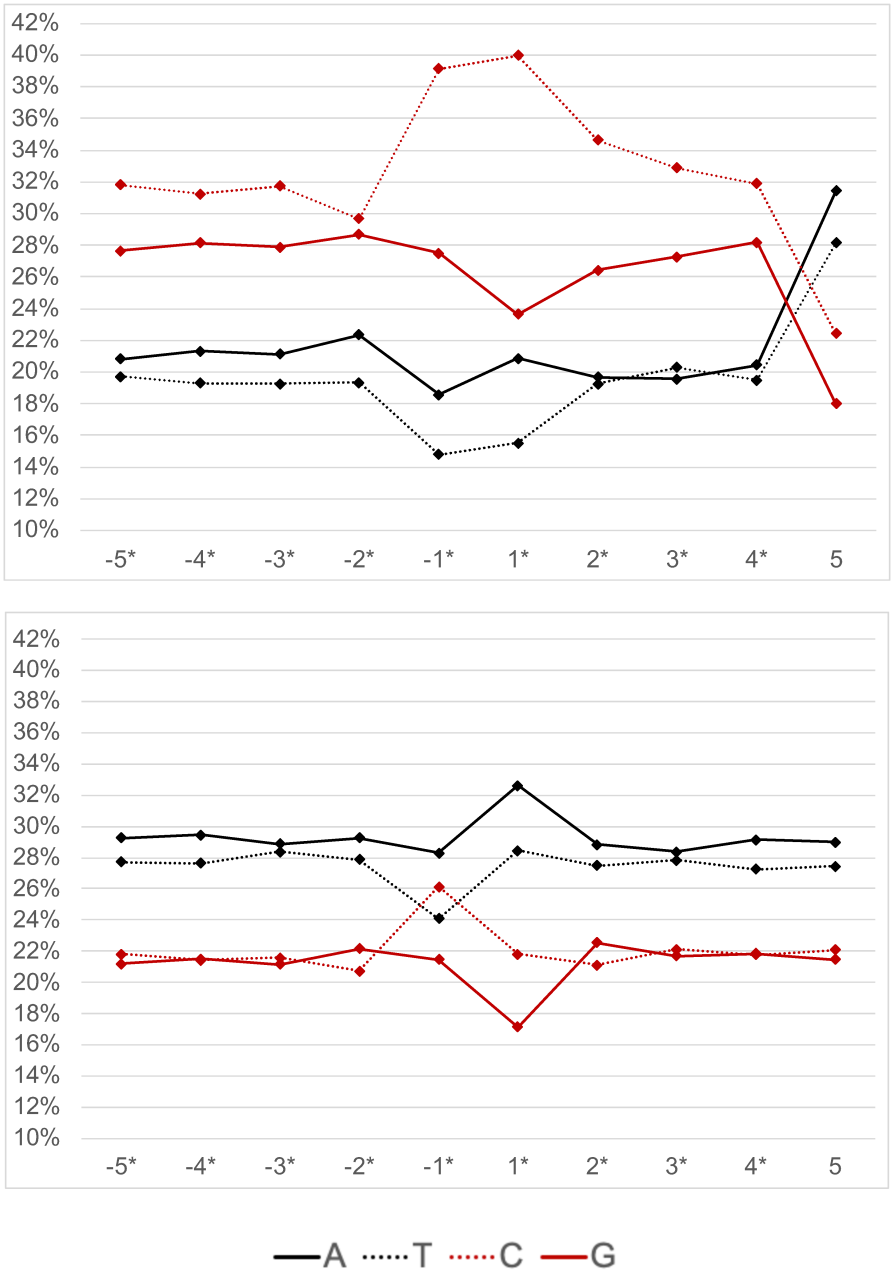


Figura Suplementar 28: Ocorrência de bases nas proximidades do SNP M (A/C) e a frequência basal esperada no genoma, nas regiões 5'UTR em bovinos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região contém o SNP e a ˆbaseline frequencyˆ.

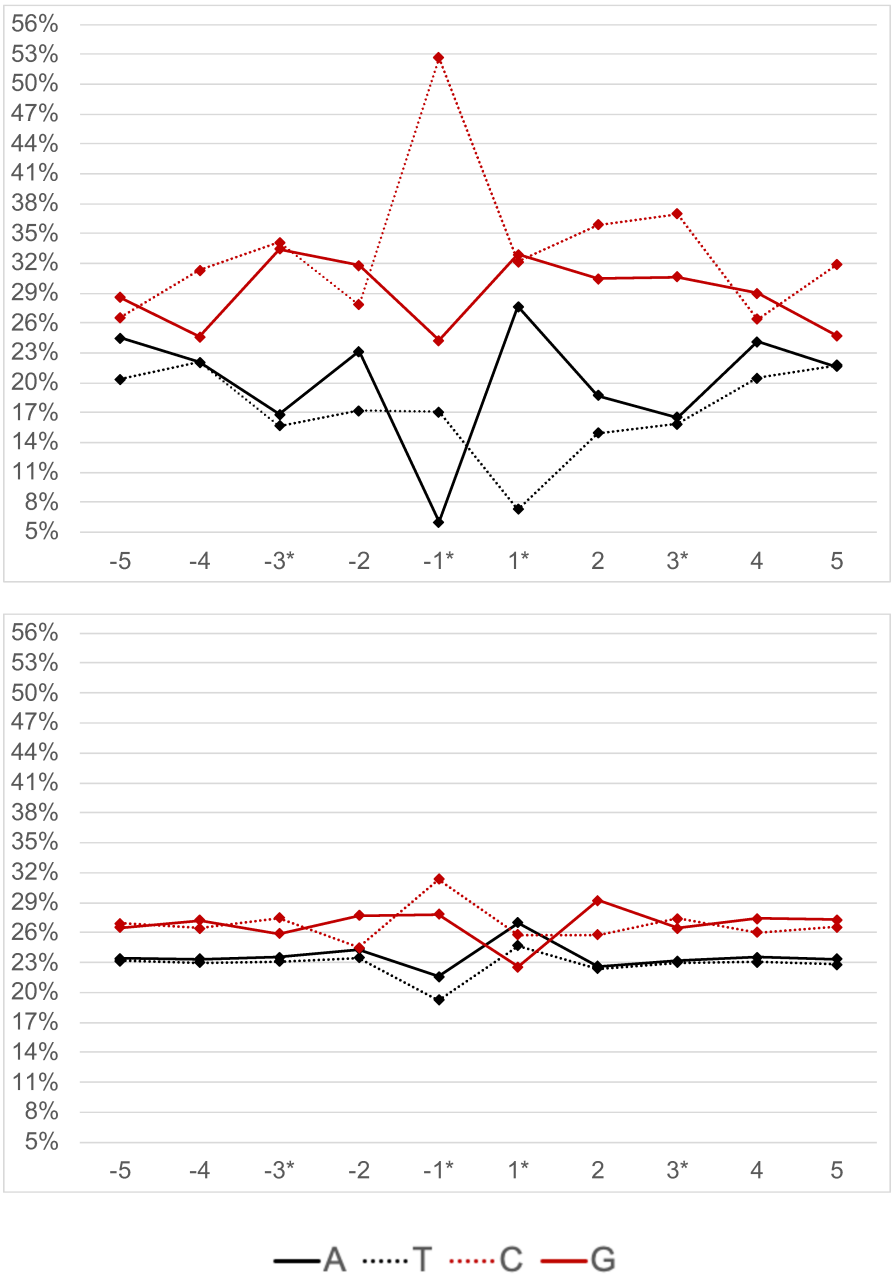


Figura Suplementar 29: Ocorrência de bases nas proximidades do SNP M (A/C) e a frequência basal esperada no genoma, nas regiões sinônimas em bovinos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency^.

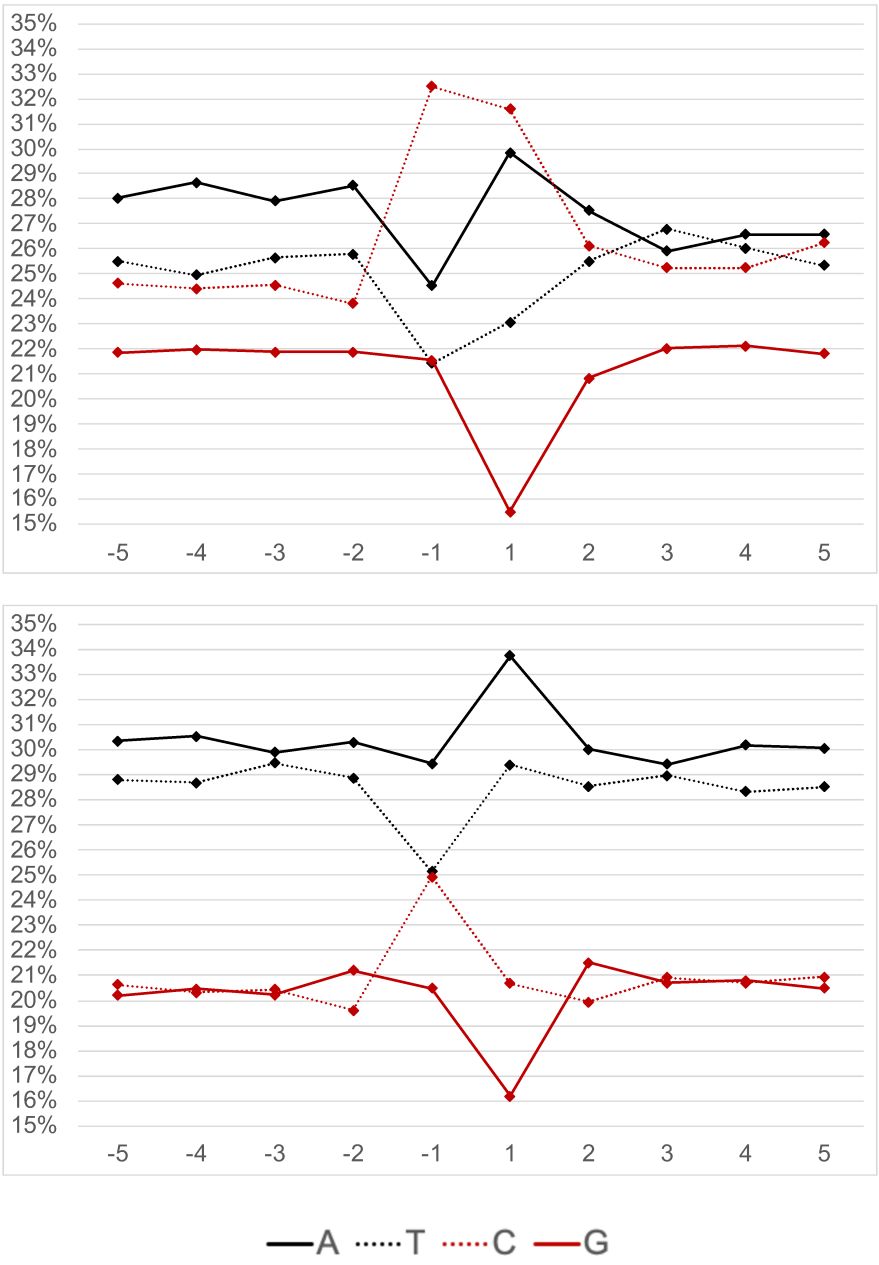


Figura Suplementar 30: Ocorrência de bases nas proximidades do SNP M (A/C) e a frequência basal esperada no genoma, nas regiões intrônicas de bovinos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

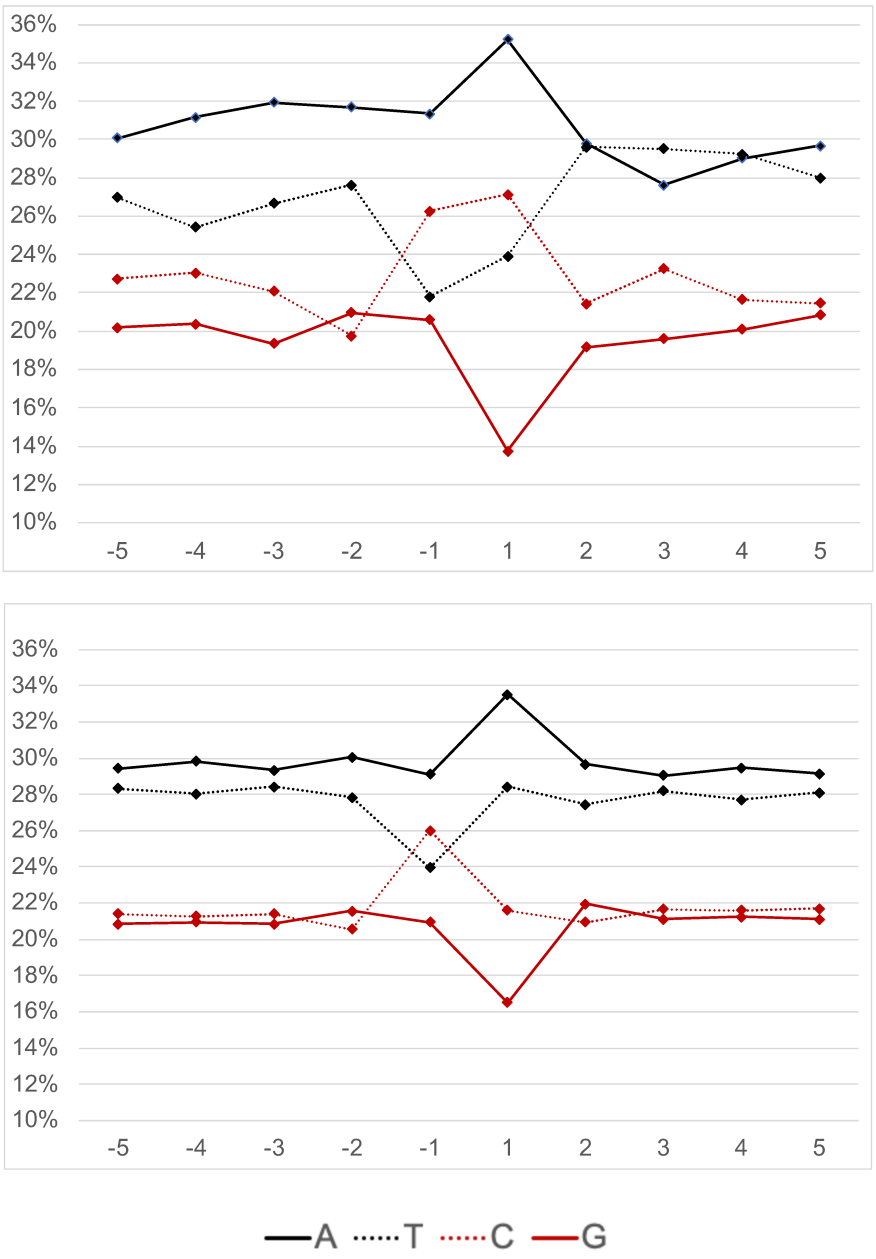


Figura Suplementar 31: Ocorrência de bases nas proximidades do SNP M (A/C) e a frequência basal esperada no genoma, nas regiões 3'UTR em humanos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

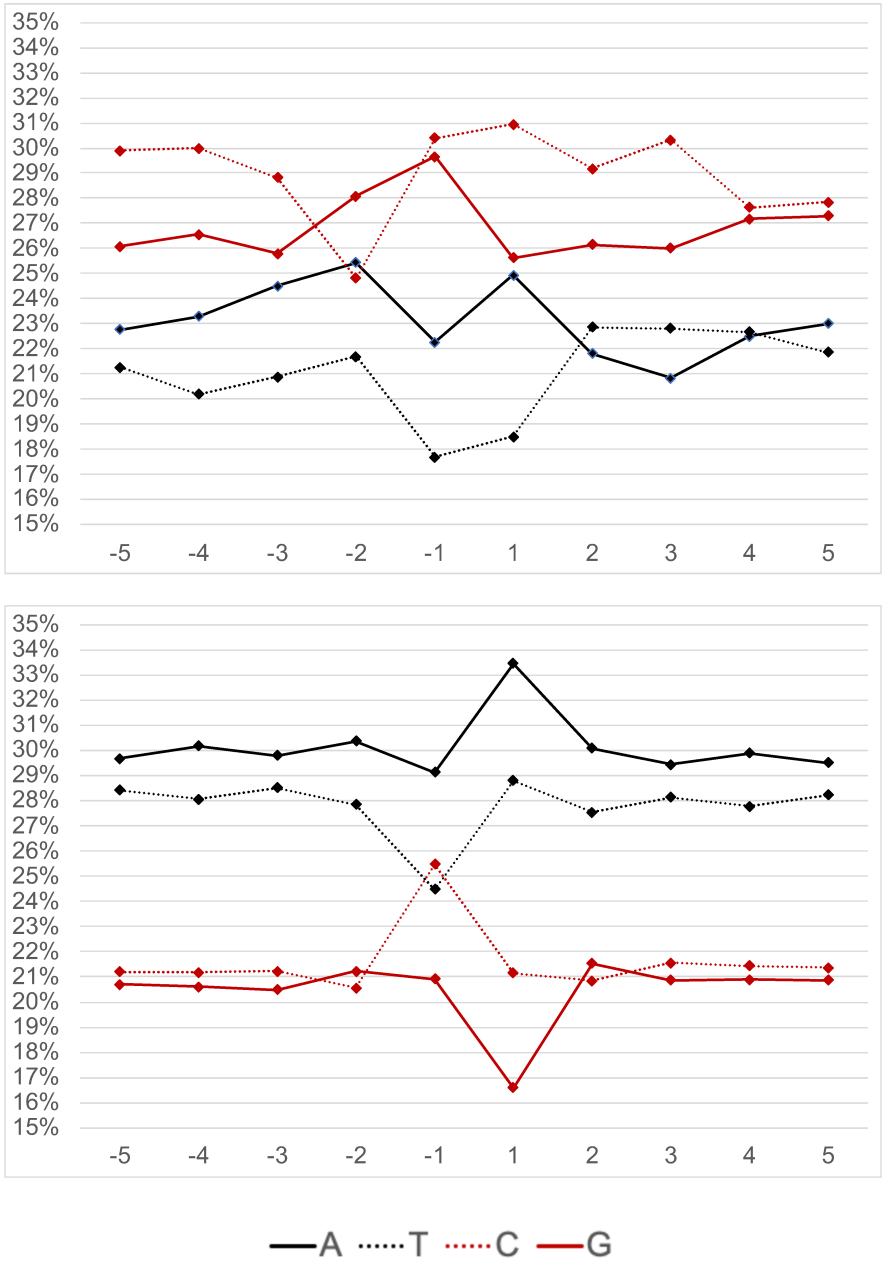


Figura Suplementar 32: Ocorrência de bases nas proximidades do SNP M (A/C) e a frequência basal esperada no genoma, nas regiões 5'UTR em humanos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

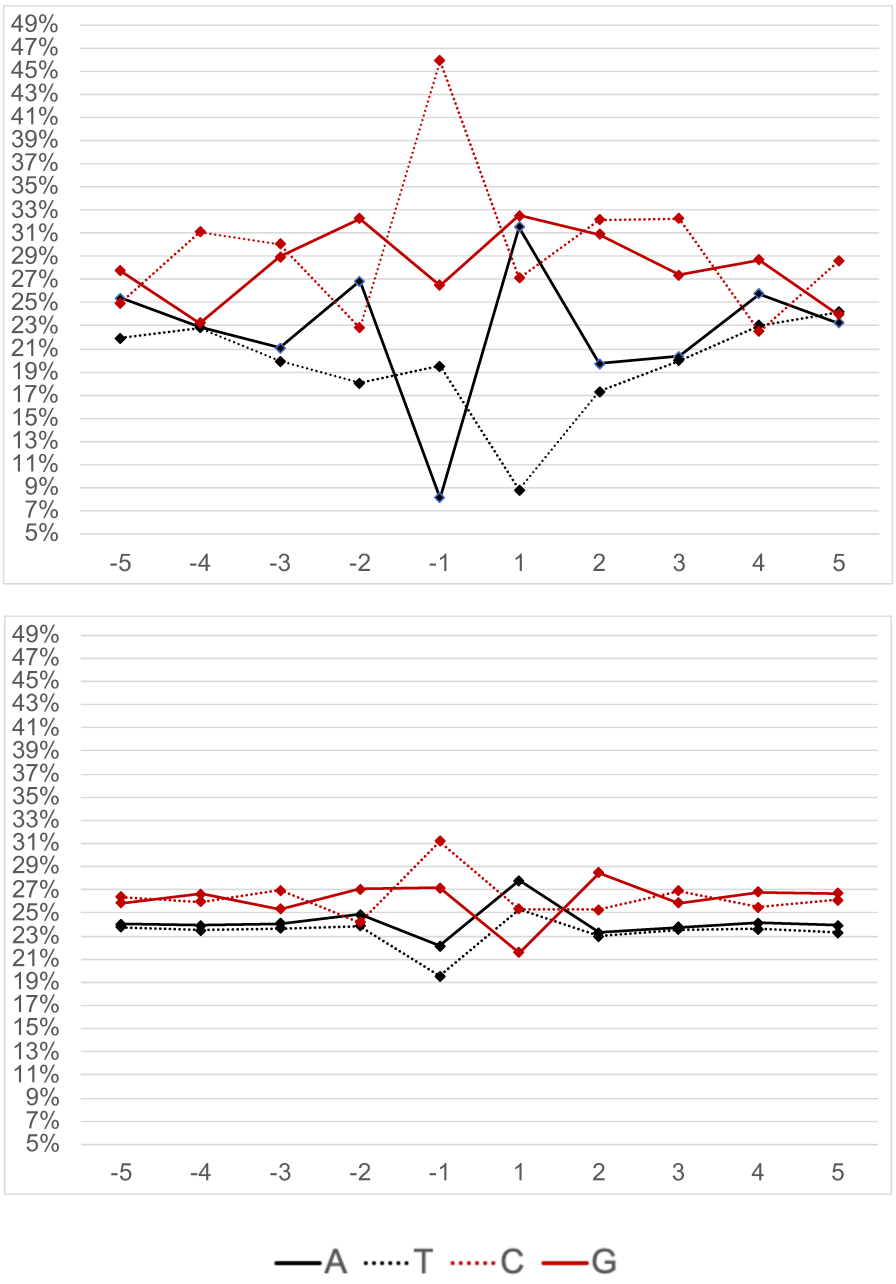


Figura Suplementar 33: Ocorrência de bases nas proximidades do SNP M (A/C) e a frequência basal esperada no genoma, nas regiões sinônimas em humanos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

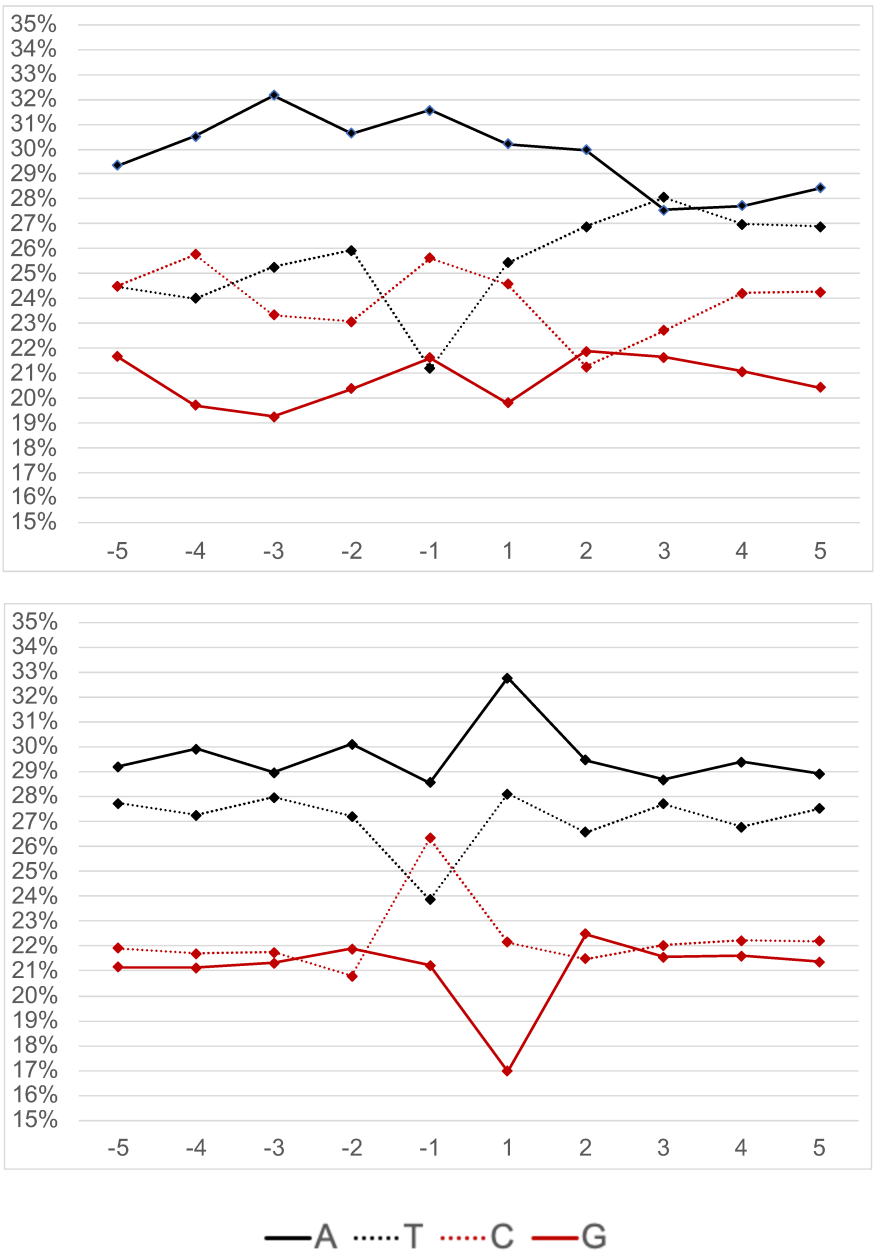


Figura Suplementar 34: Ocorrência de bases nas proximidades do SNP M (A/C) e a frequência basal esperada no genoma, nas regiões 3'UTR em camundongos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

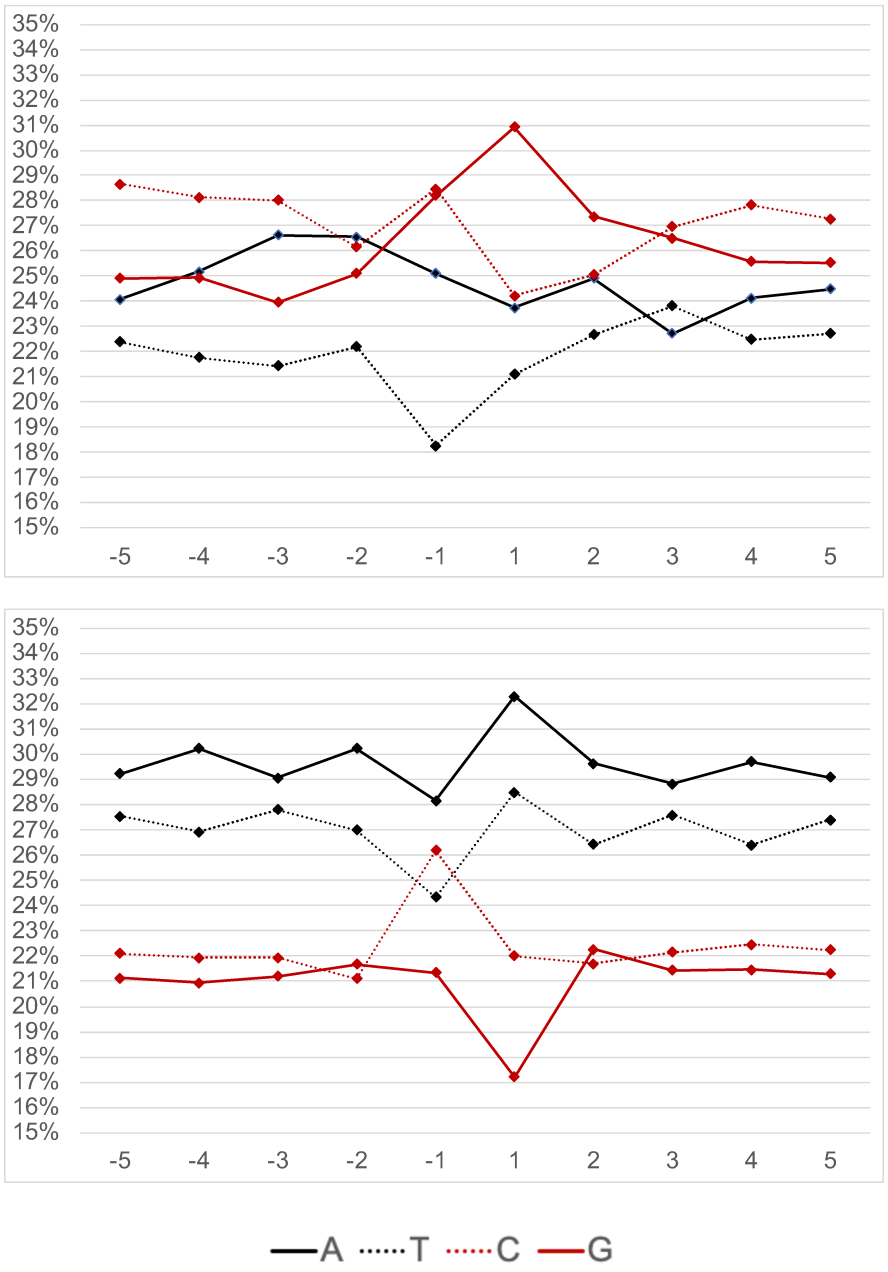


Figura Suplementar 35: Ocorrência de bases nas proximidades do SNP M (A/C) e a frequência basal esperada no genoma, nas regiões 5'UTR em camundongos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

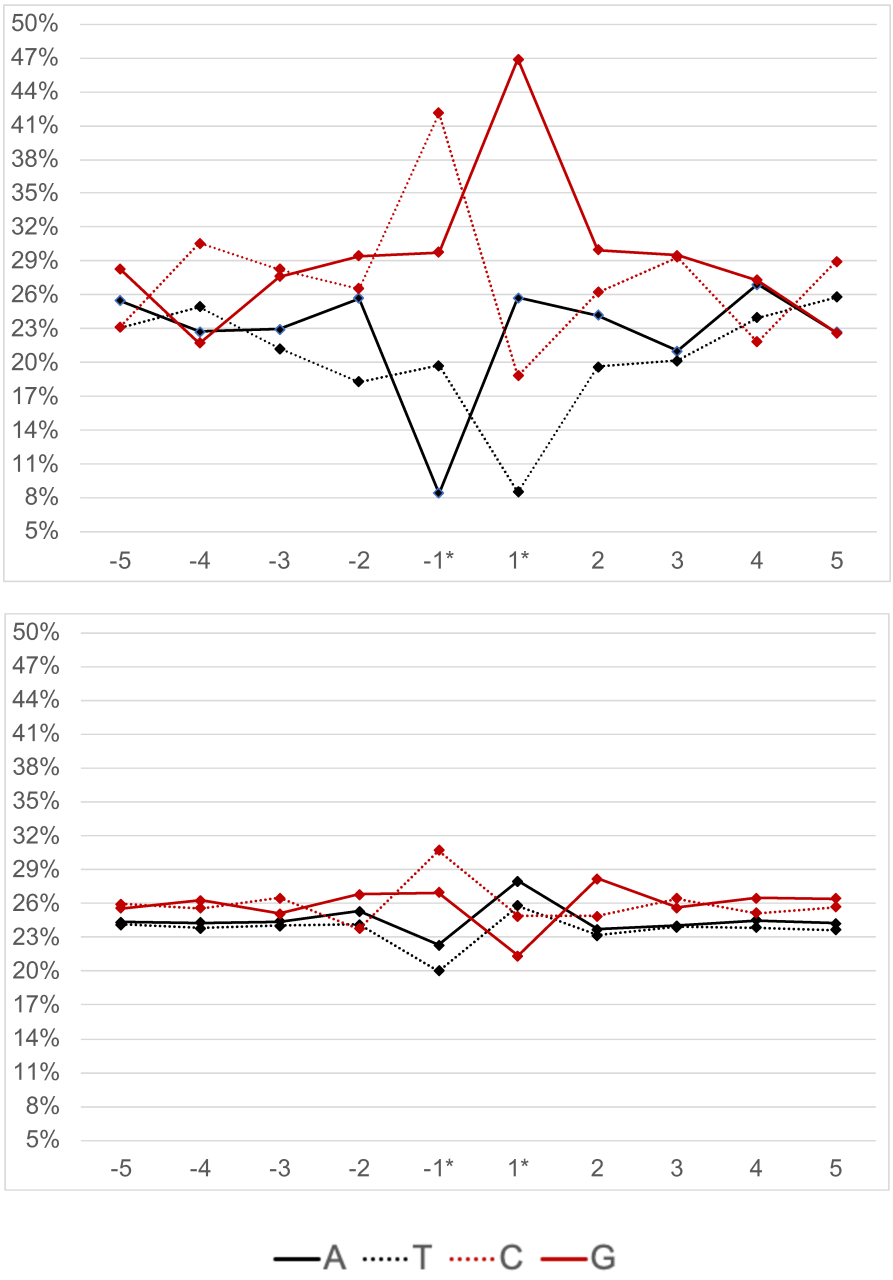


Figura Suplementar 36: Ocorrência de bases nas proximidades do SNP M (A/C) e a frequência basal esperada no genoma, nas regiões sinônimas em camundongos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency^.

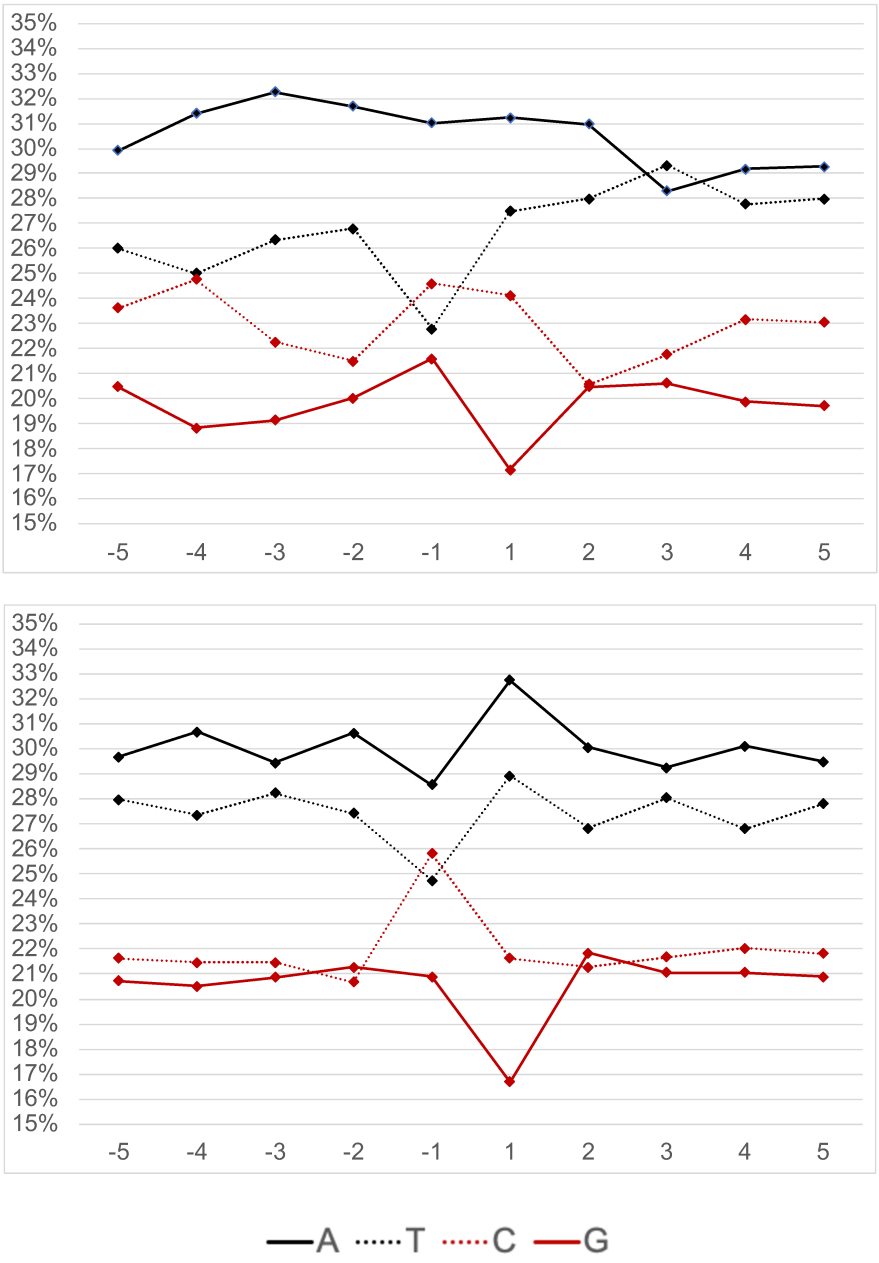


Figura Suplementar 37: Ocorrência de bases nas proximidades do SNP M (A/C) e a frequência basal esperada no genoma, nas regiões intrônicas em camundongos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

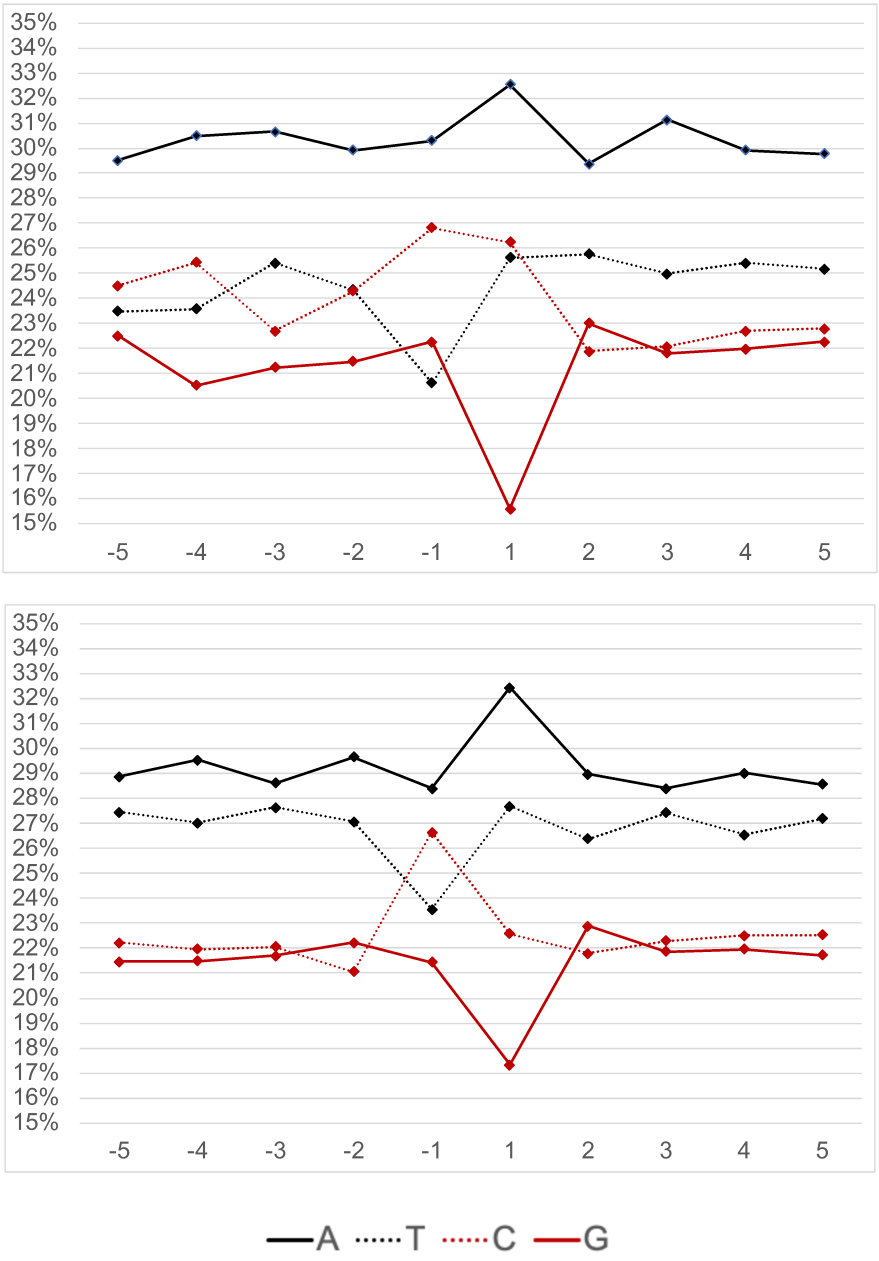


Figura Suplementar 38: Ocorrência de bases nas proximidades do SNP M (A/C) e a frequência basal esperada no genoma, nas regiões 3'UTR em *Rattus norvegicus*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

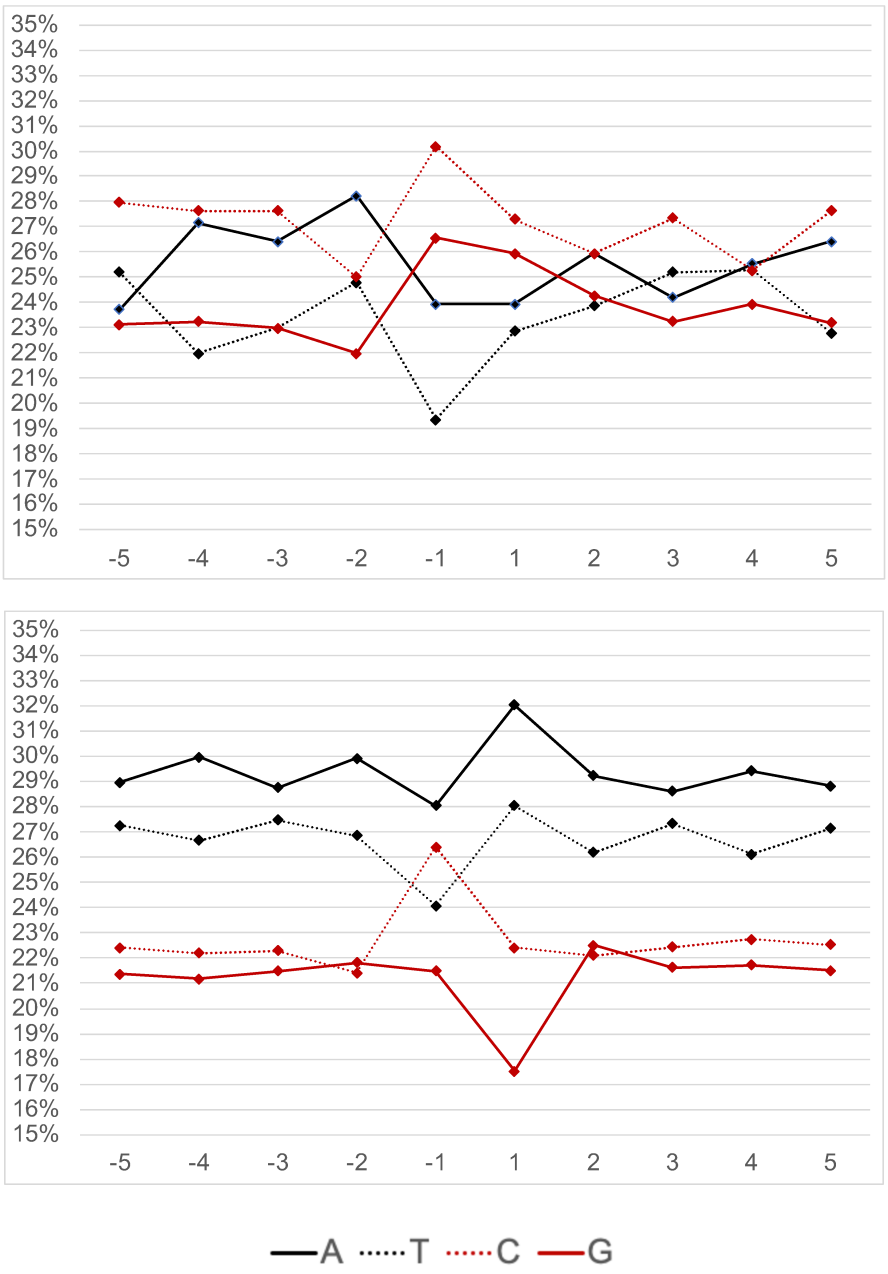


Figura Suplementar 39: Ocorrência de bases nas proximidades do SNP M (A/C) e a frequência basal esperada no genoma, nas regiões 5'UTR em *Rattus norvegicus*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

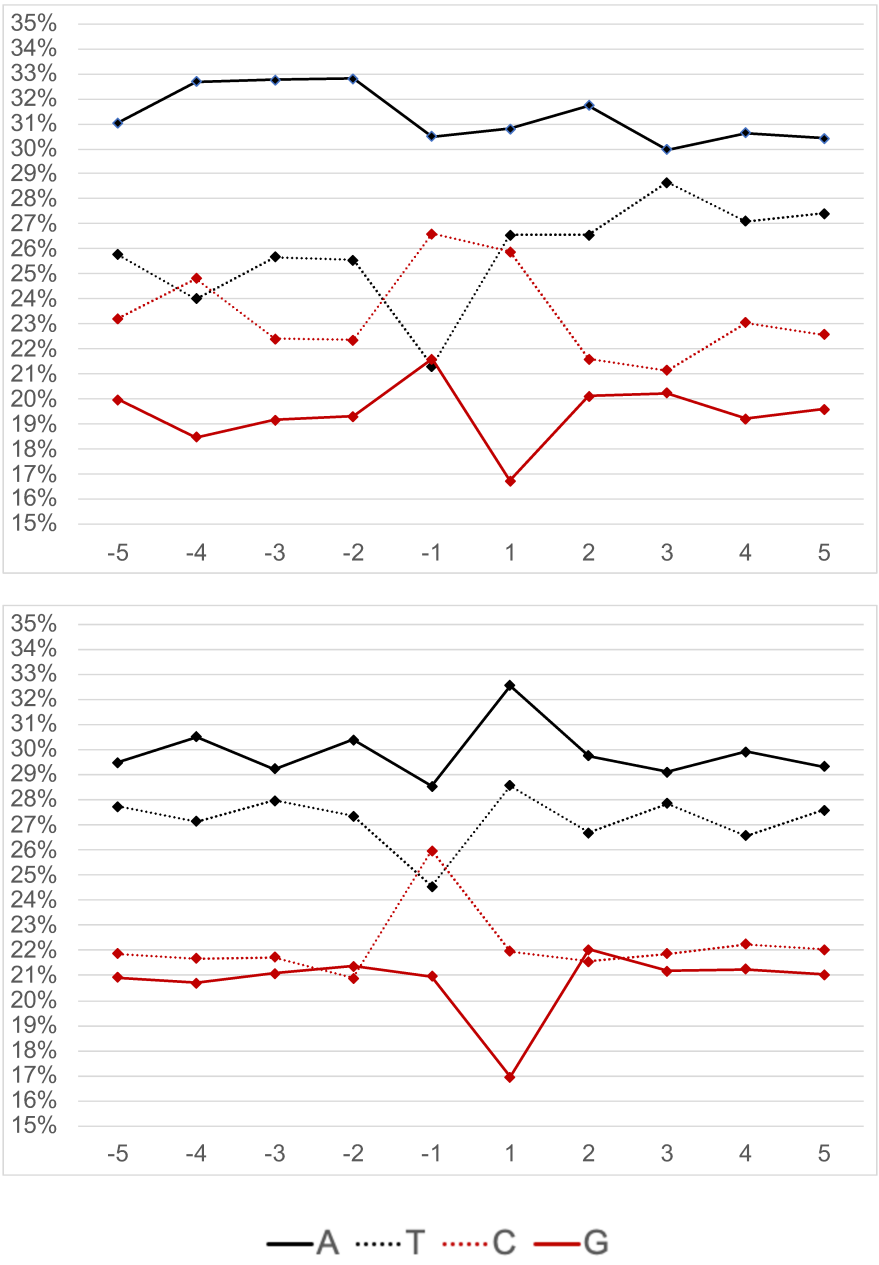


Figura Suplementar 40: Ocorrência de bases nas proximidades do SNP M (A/C) e a frequência basal esperada no genoma, nas regiões intrônicas em *Rattus norvegicus*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

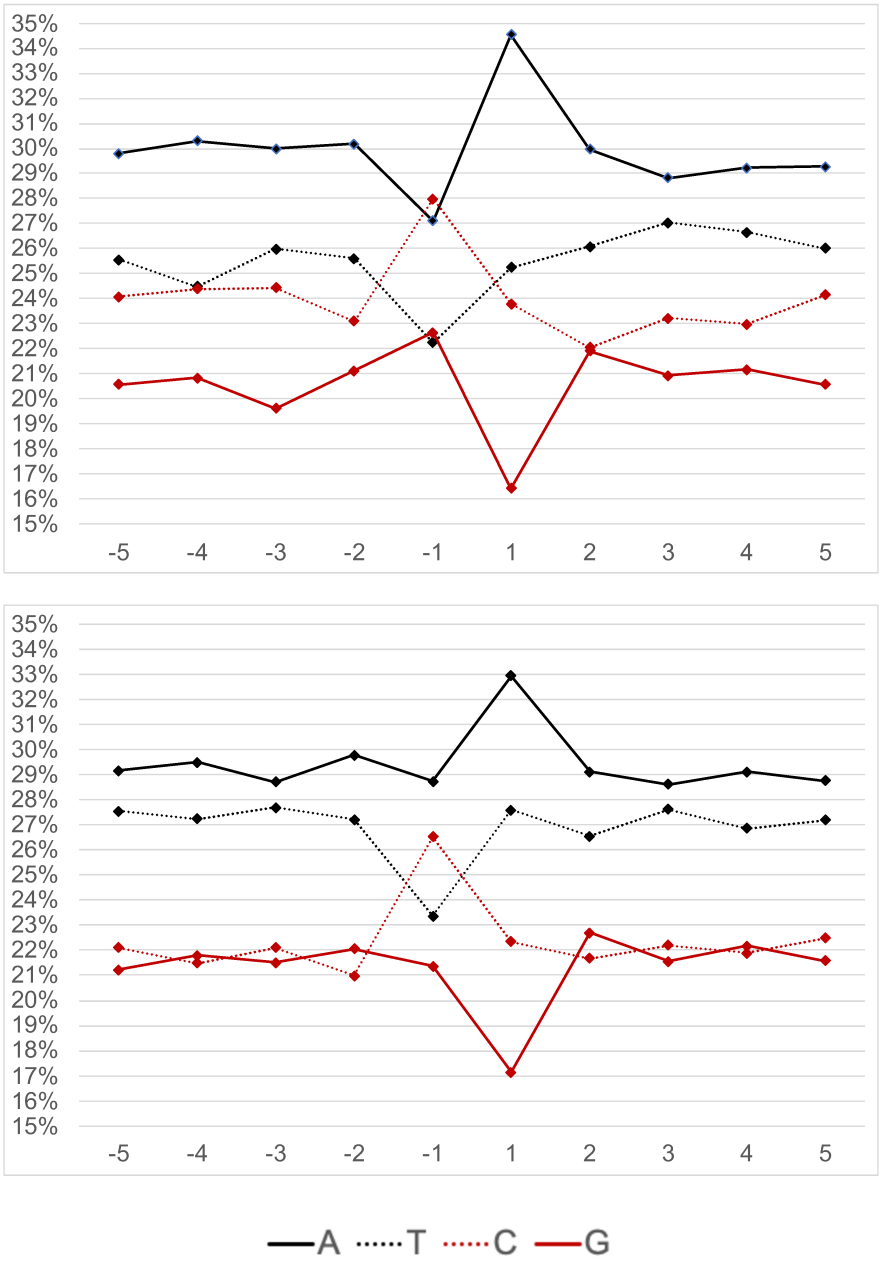


Figura Suplementar 41: Ocorrência de bases nas proximidades do SNP M (A/C) e a frequência basal esperada no genoma, nas regiões 3'UTR em *Sus scrofa*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

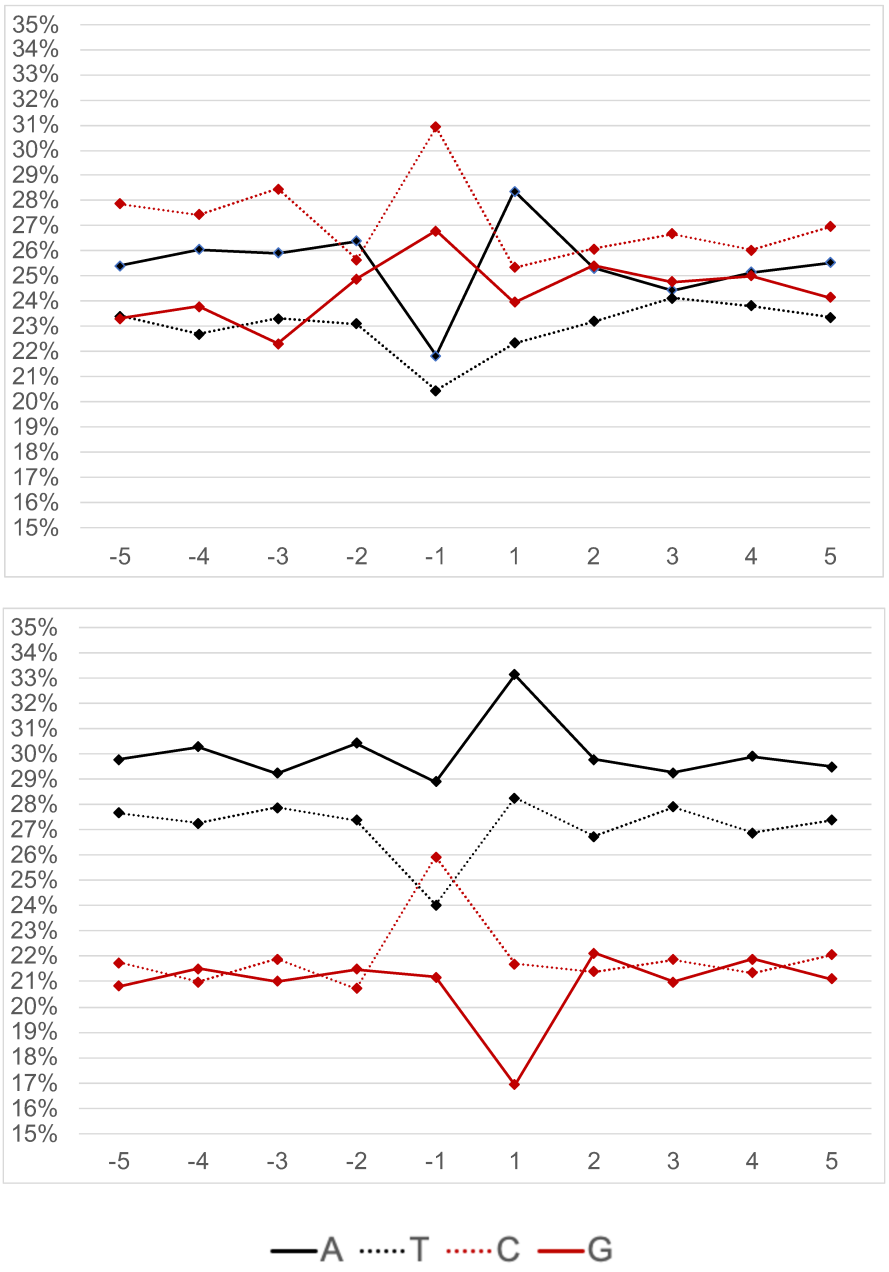


Figura Suplementar 42: Ocorrência de bases nas proximidades do SNP M (A/C) e a frequência basal esperada no genoma, nas regiões 5'UTR em *Sus scrofa*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

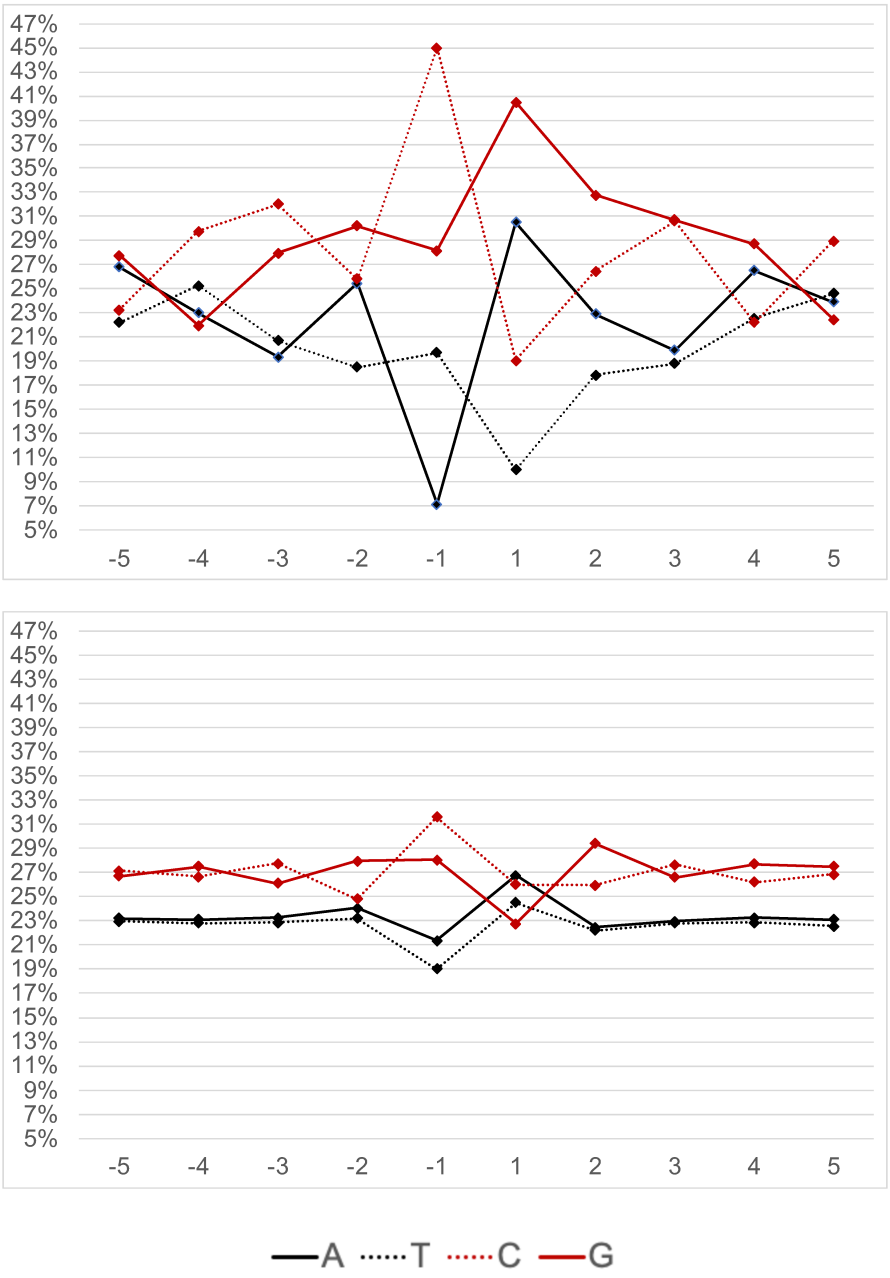


Figura Suplementar 43: Ocorrência de bases nas proximidades do SNP M (A/C) e a frequência basal esperada no genoma, nas regiões sinônimas em *Sus scrofa*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

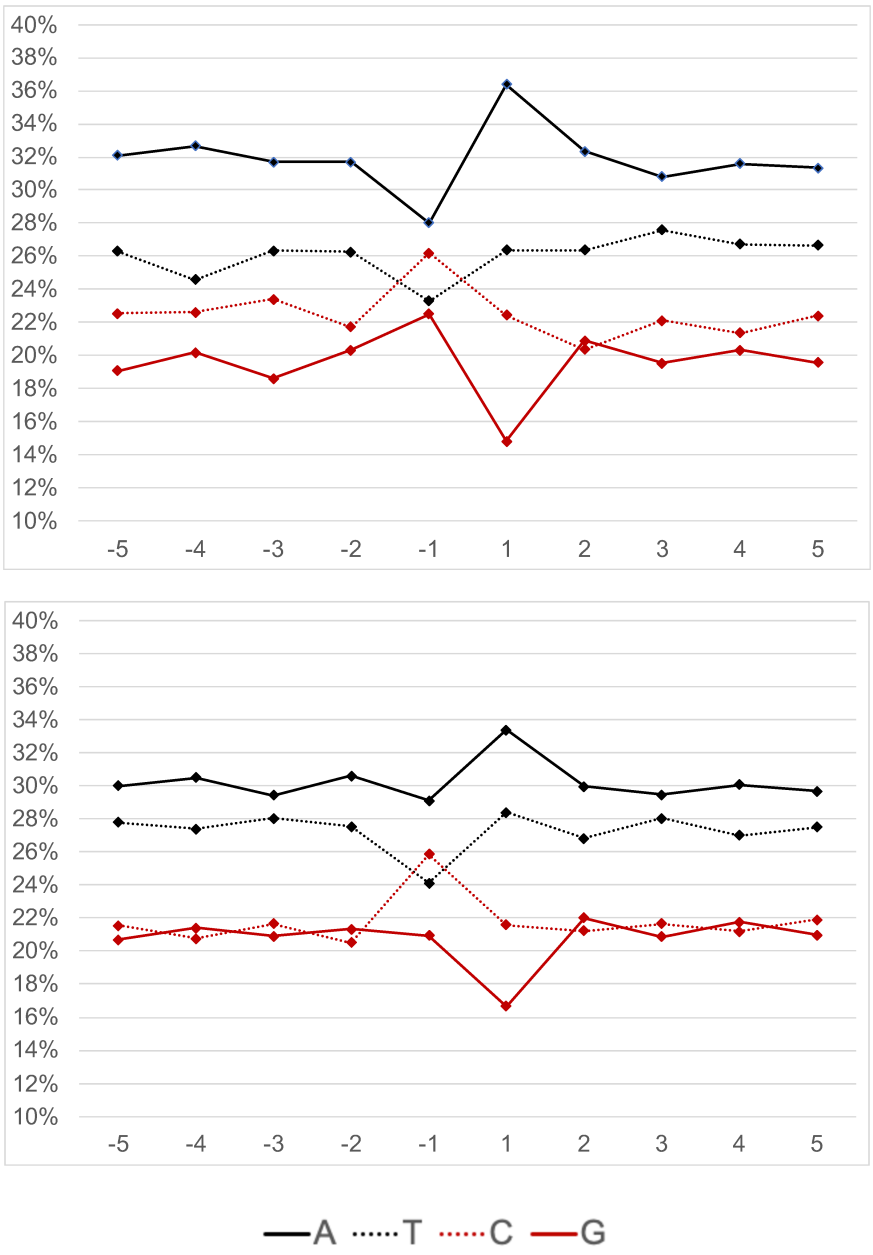


Figura Suplementar 44: Ocorrência de bases nas proximidades do SNP M (A/C) e a frequência basal esperada no genoma, nas regiões intrônicas em *Sus scrofa*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

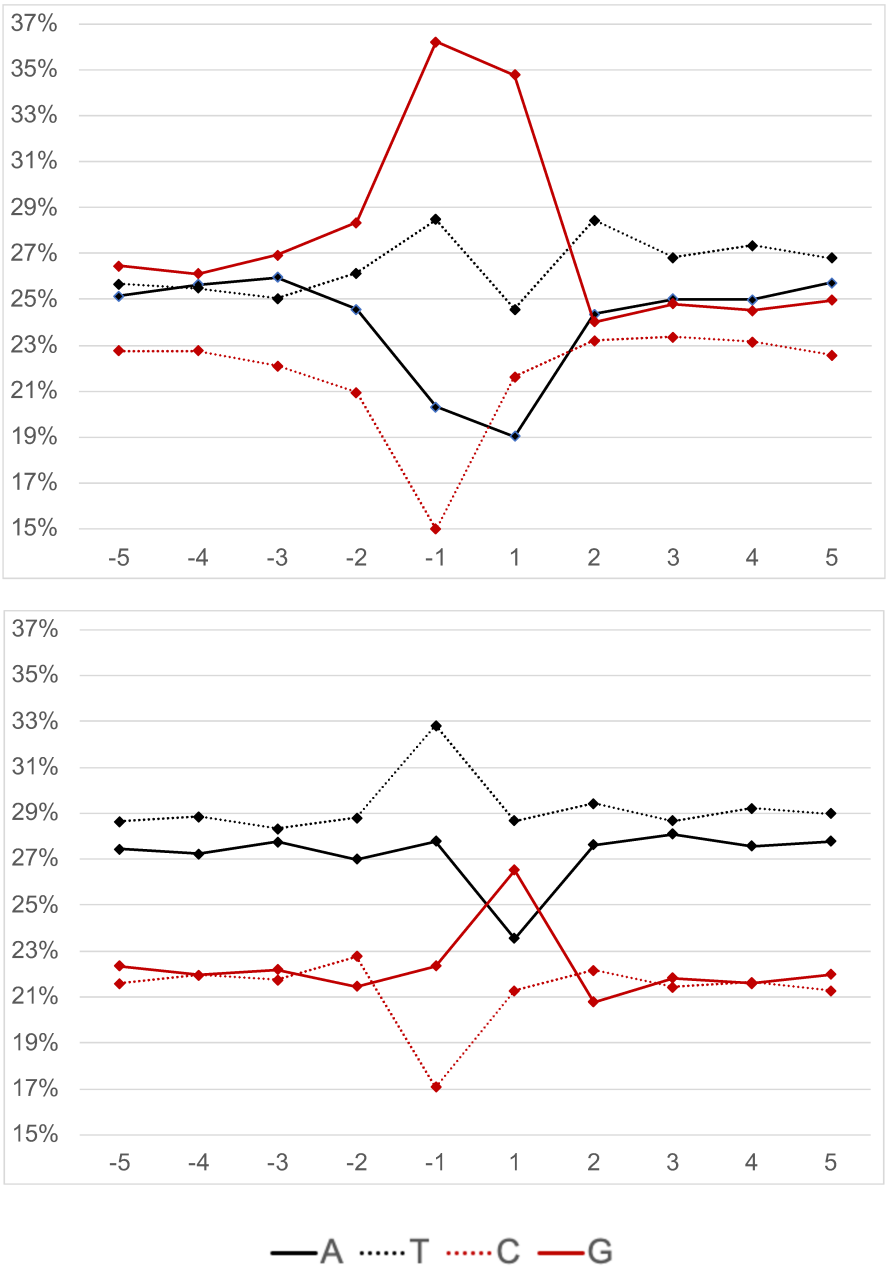


Figura Suplementar 45: Ocorrência de bases nas proximidades do SNP K (G/T) e a frequência basal esperada no genoma, nas regiões 3'UTR em bovinos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). frequencyˆ.

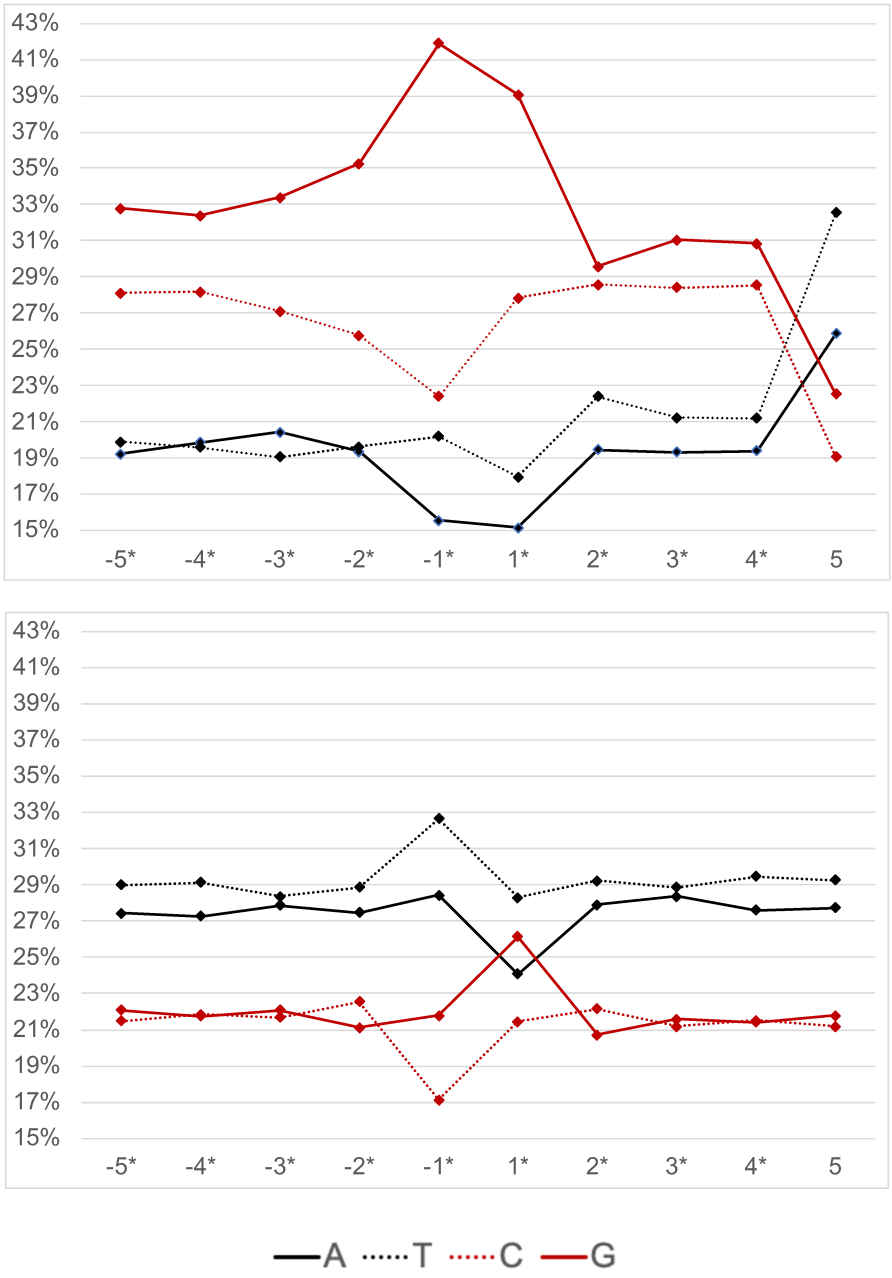


Figura Suplementar 46: Ocorrência de bases nas proximidades do K (G/T) e a frequência basal esperada no genoma, nas regiões 5'UTR em bovinos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ˆbaseline frequencyˆ.

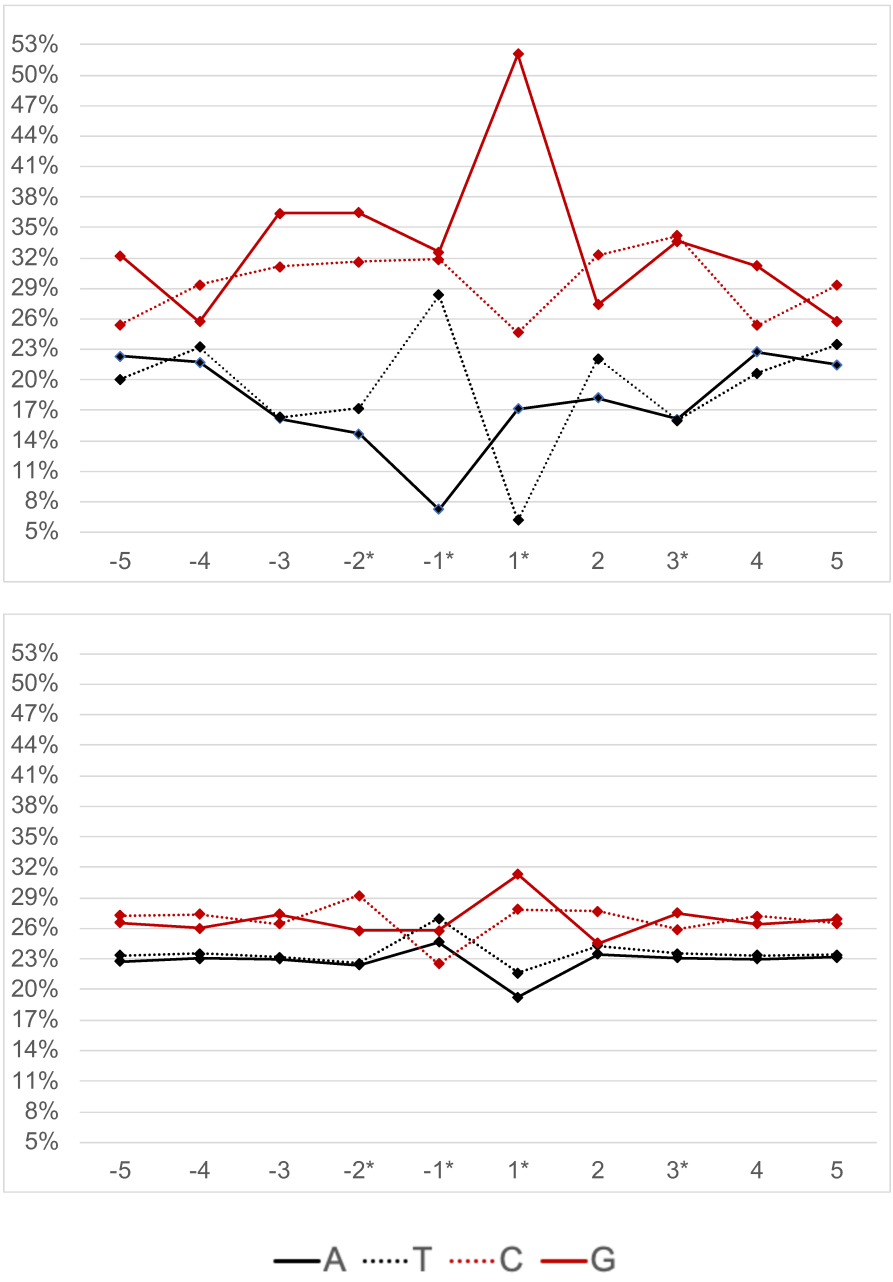


Figura Suplementar 47: Ocorrência de bases nas proximidades do K (G/T) e a frequência basal esperada no genoma, nas regiões sinônimas em bovinos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency^.

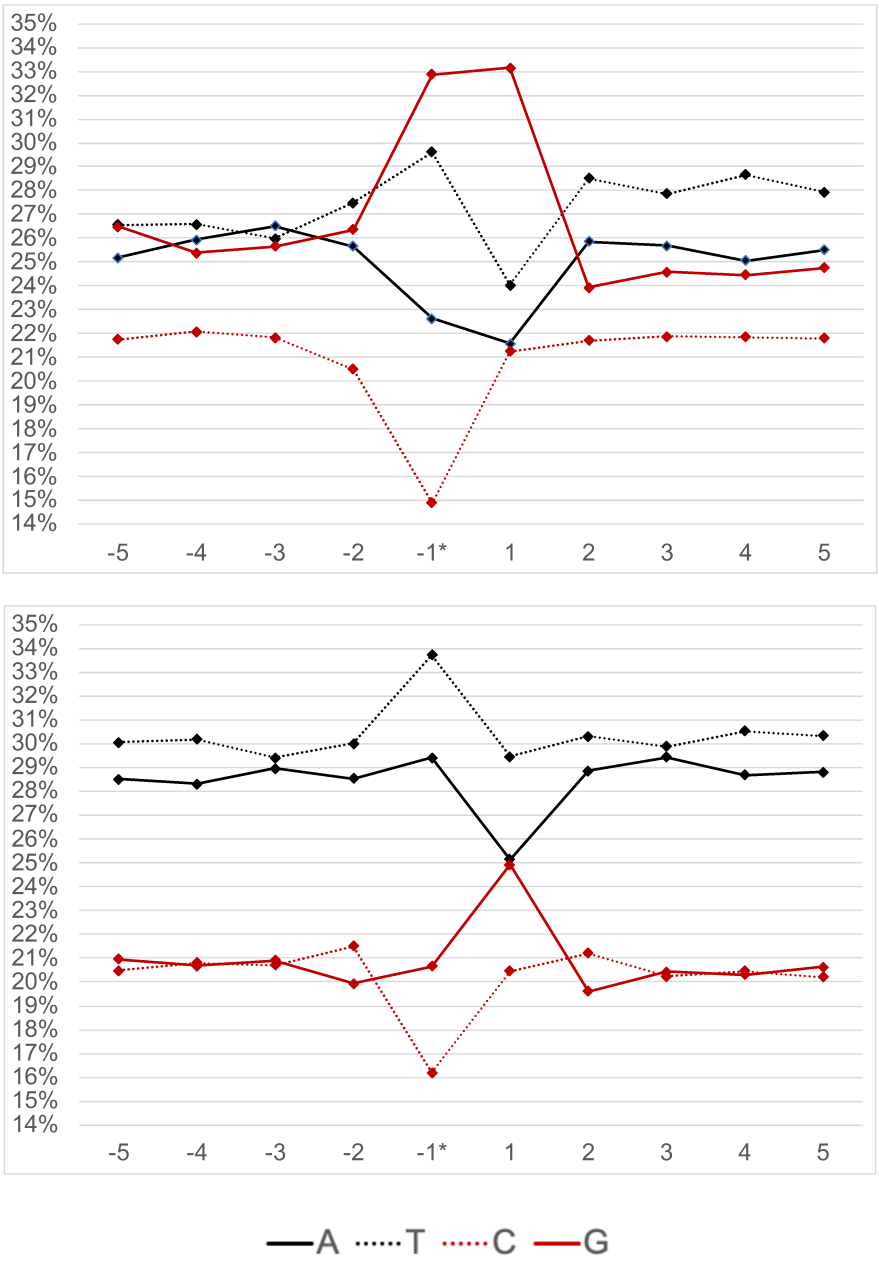


Figura Suplementar 48: Ocorrência de bases nas proximidades do SNP K (G/T) e a frequência basal esperada no genoma, nas regiões intrônicas de bovinos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency^.

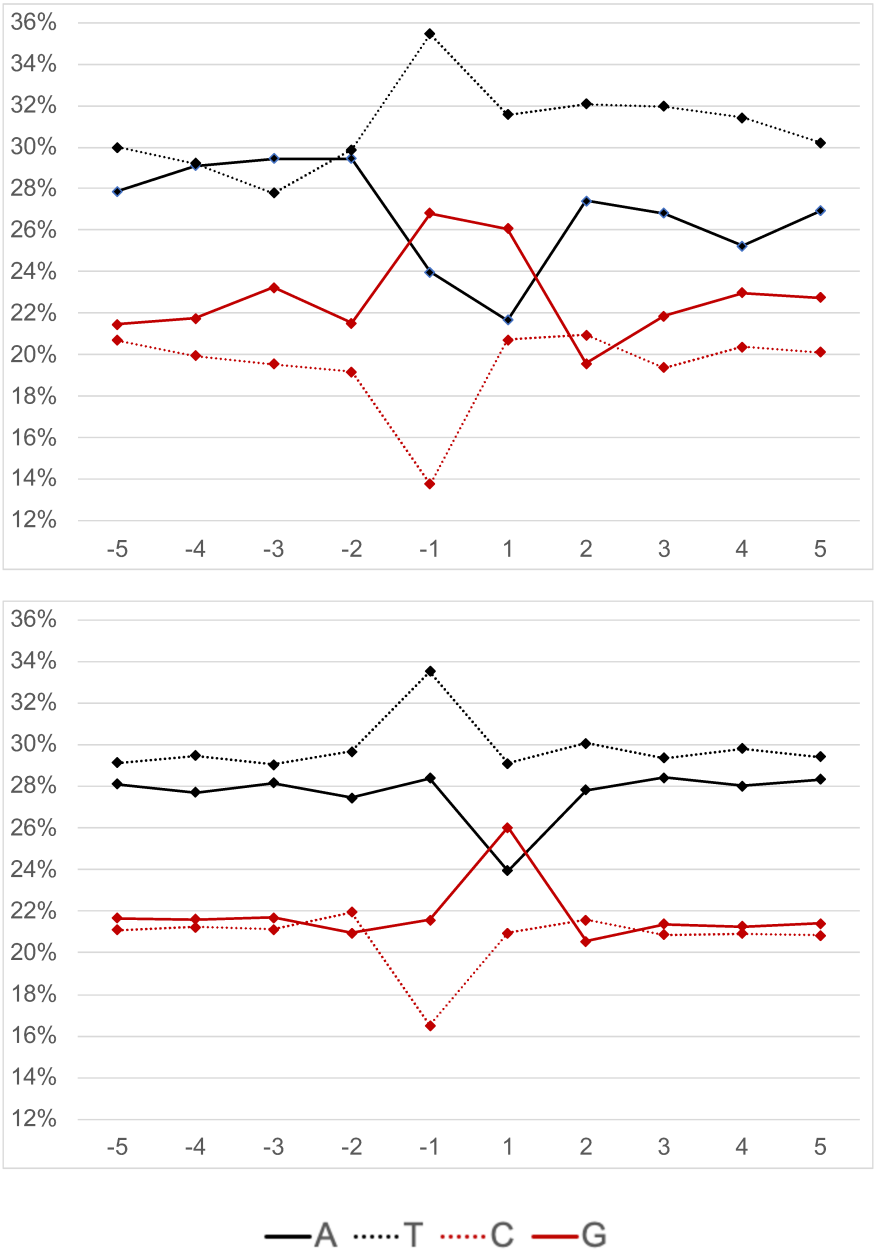


Figura Suplementar 49: Ocorrência de bases nas proximidades do SNP K (G/T) e a frequência basal esperada no genoma, nas regiões 3'UTR em humanos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

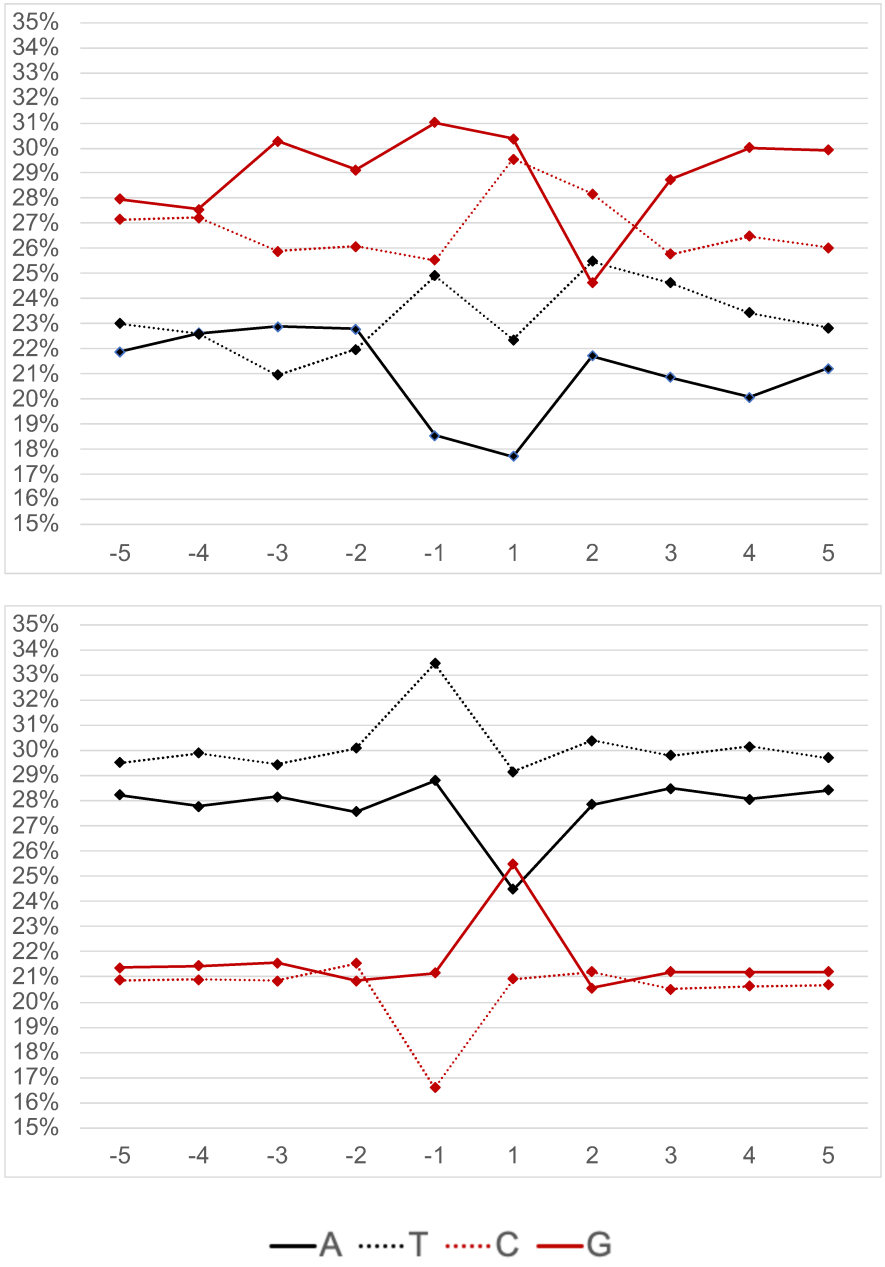


Figura Suplementar 50: Ocorrência de bases nas proximidades do SNP K (G/T) e a frequência basal esperada no genoma, nas regiões 5'UTR em humanos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

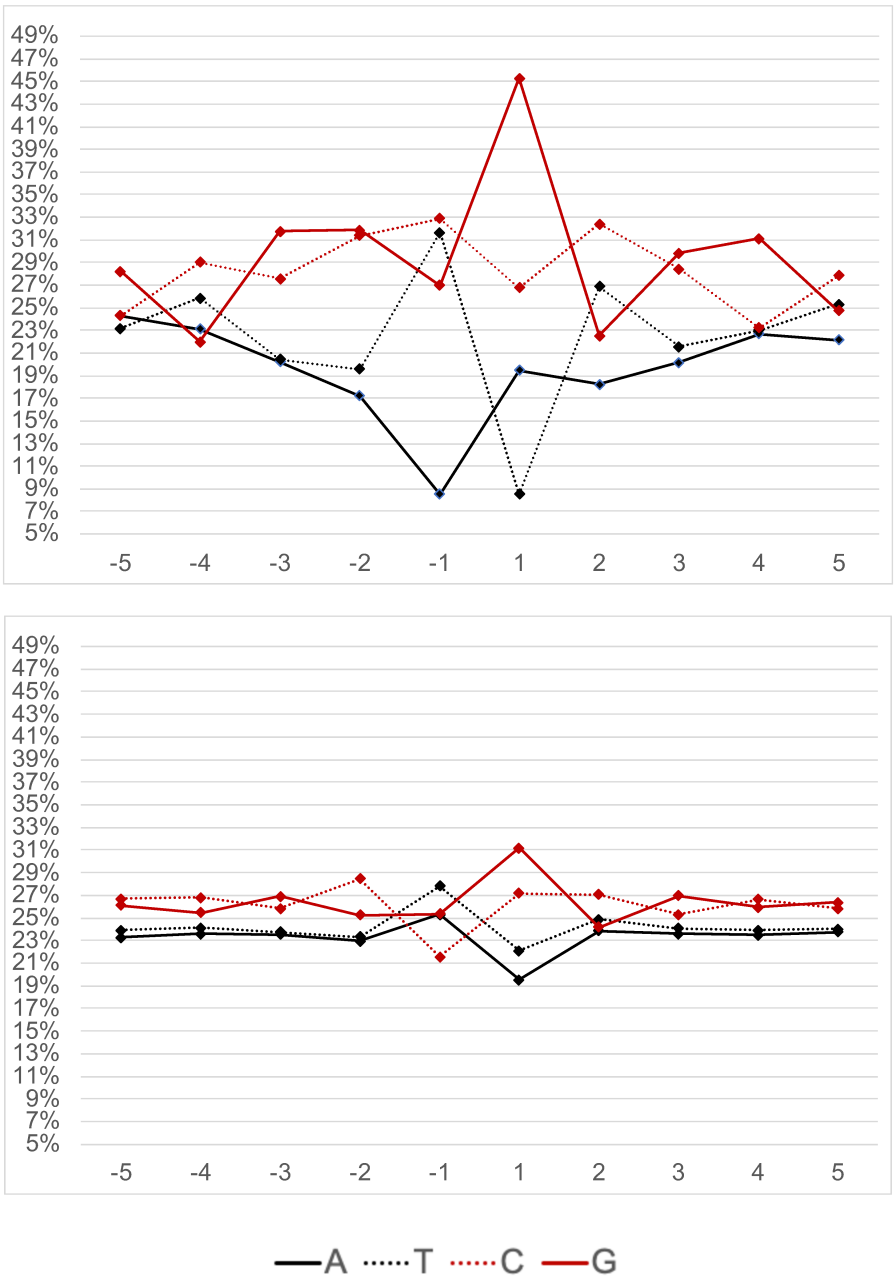


Figura Suplementar 51: Ocorrência de bases nas proximidades do SNP K (G/T) e a frequência basal esperada no genoma, nas regiões sinônimas em humanos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

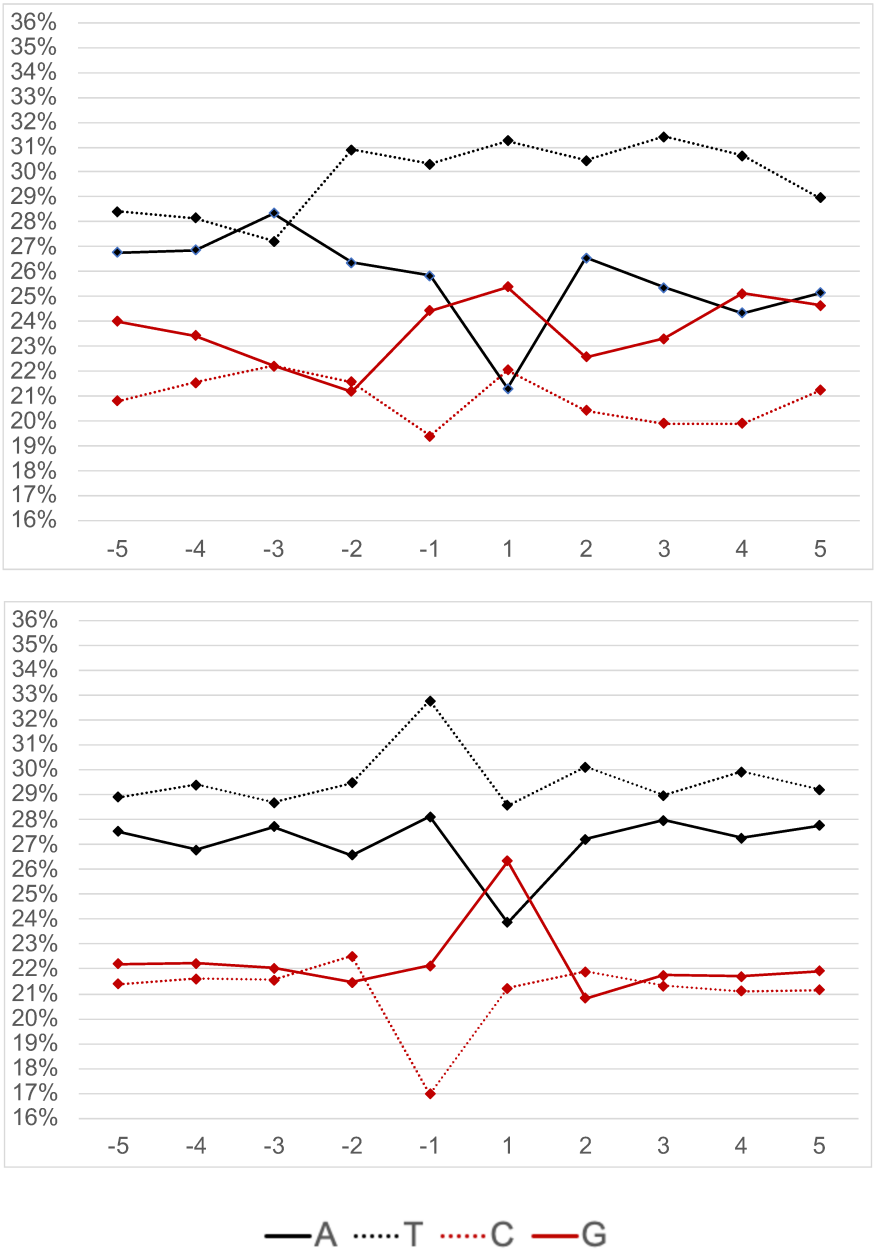


Figura Suplementar 52: Ocorrência de bases nas proximidades do SNP K (G/T) e a frequência basal esperada no genoma, nas regiões 3'UTR em camundongos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

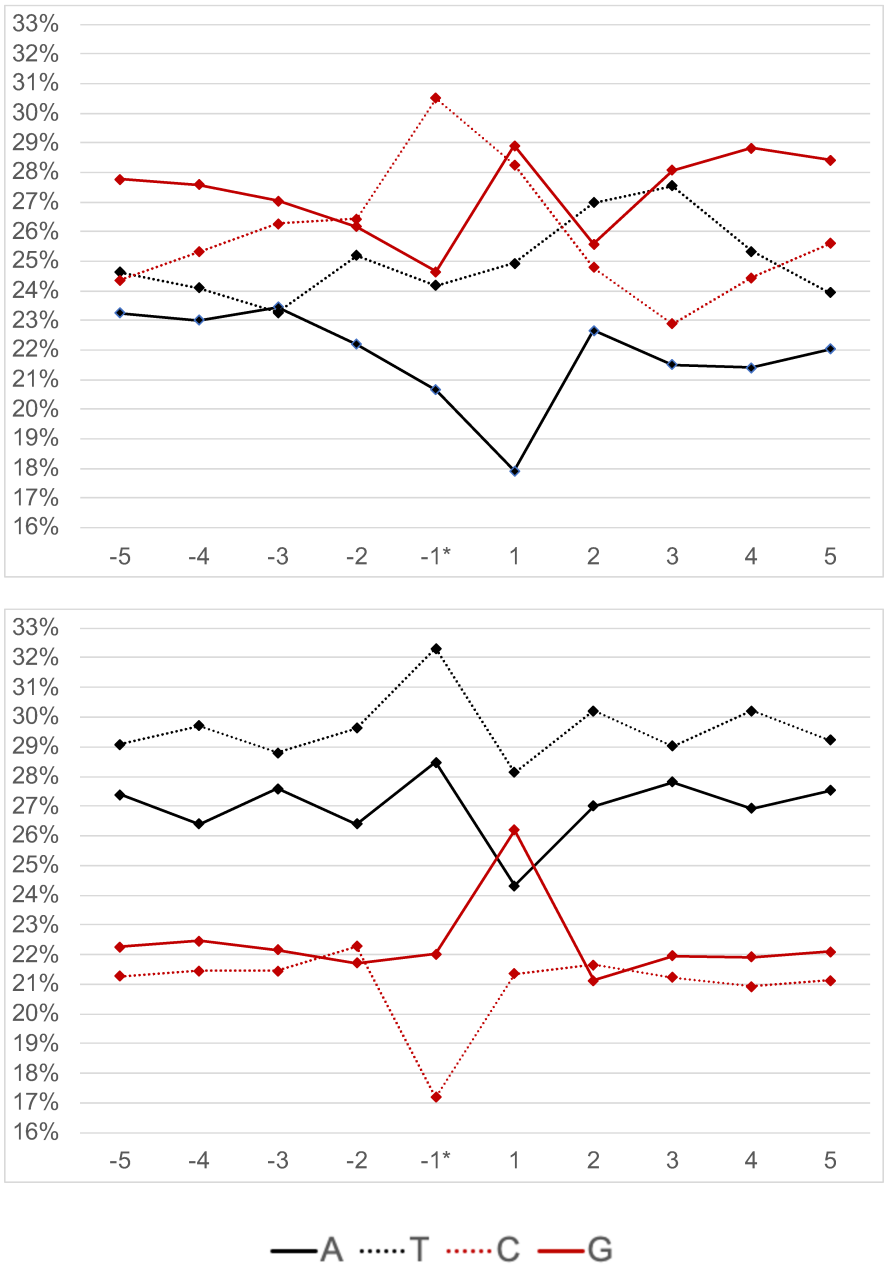


Figura Suplementar 53: Ocorrência de bases nas proximidades do SNP K (G/T) e a frequência basal esperada no genoma, nas regiões 5'UTR em camundongos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency^.

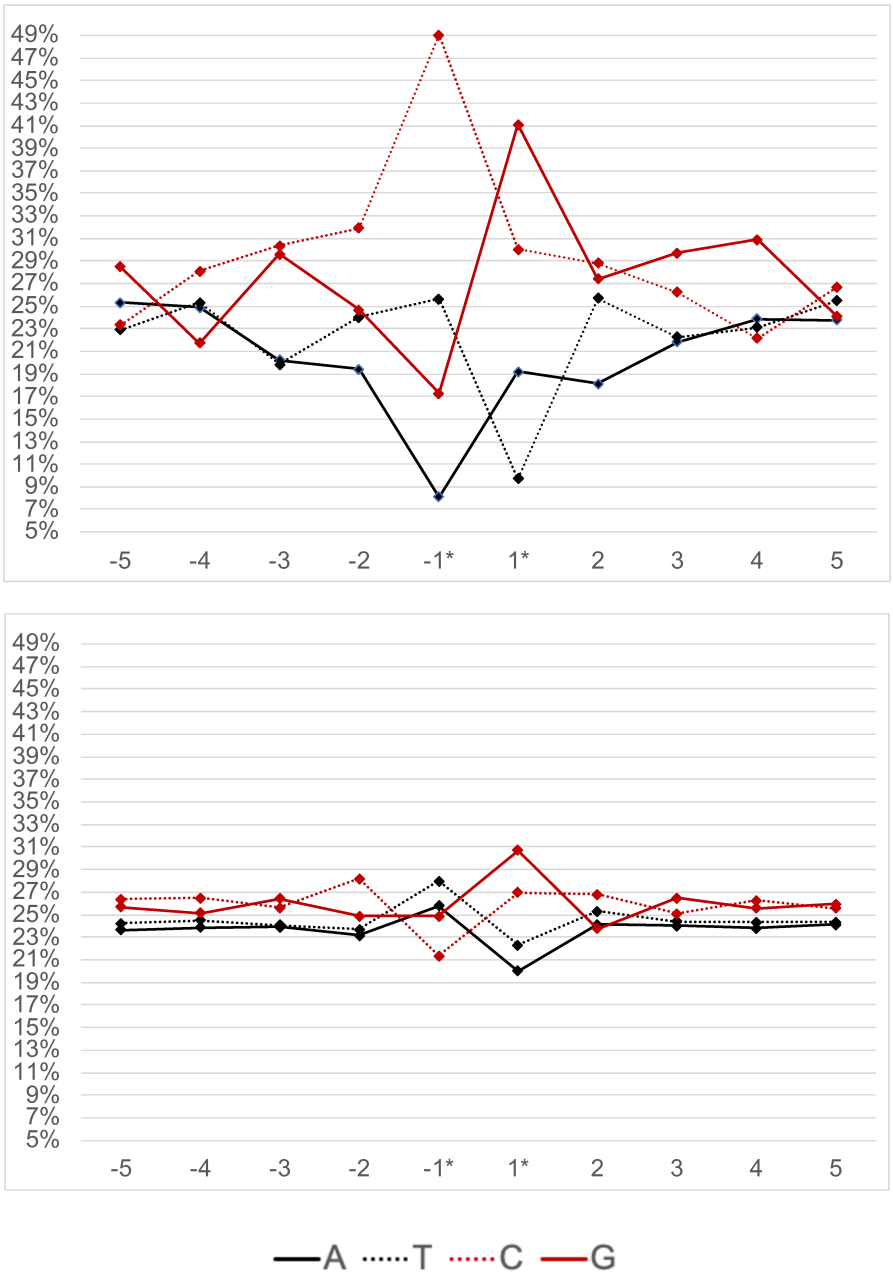


Figura Suplementar 54: Ocorrência de bases nas proximidades do SNP K (G/T) e a frequência basal esperada no genoma, nas regiões sinônimas em camundongos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency^.

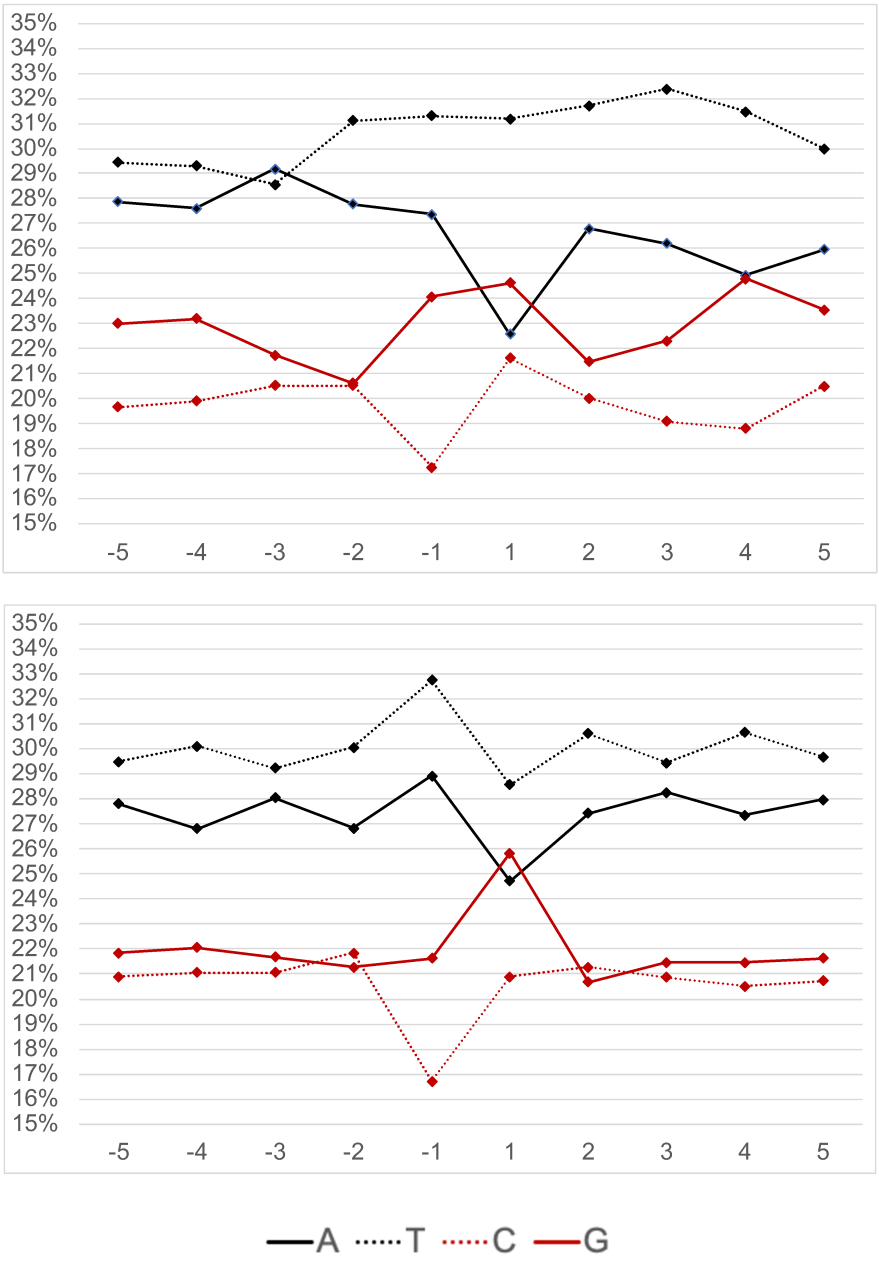


Figura Suplementar 55: Ocorrência de bases nas proximidades do SNP K (G/T) e a frequência basal esperada no genoma, nas regiões intrônicas em camundongos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

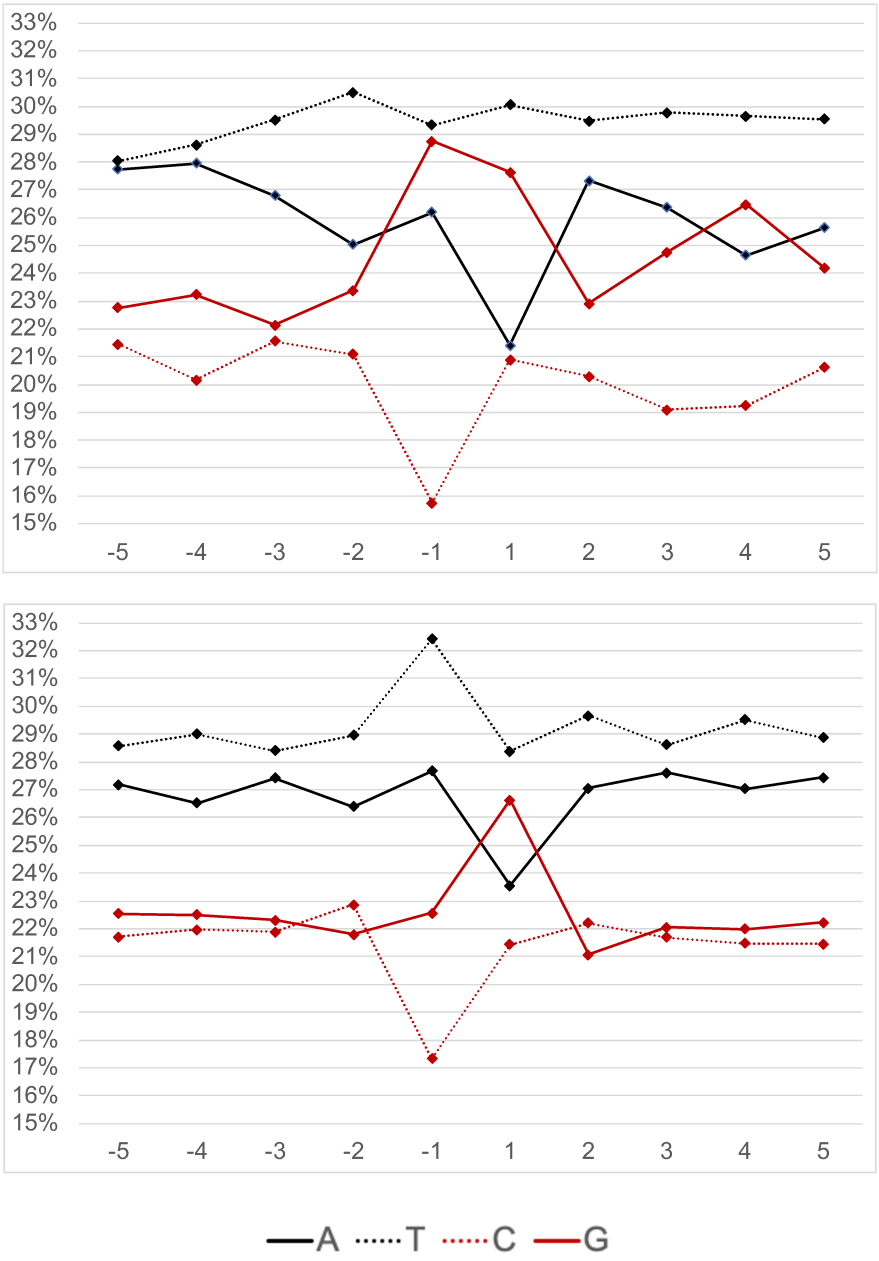


Figura Suplementar 56: Ocorrência de bases nas proximidades do SNP K (G/T) e a frequência basal esperada no genoma, nas regiões 3'UTR em *Rattus norvegicus*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

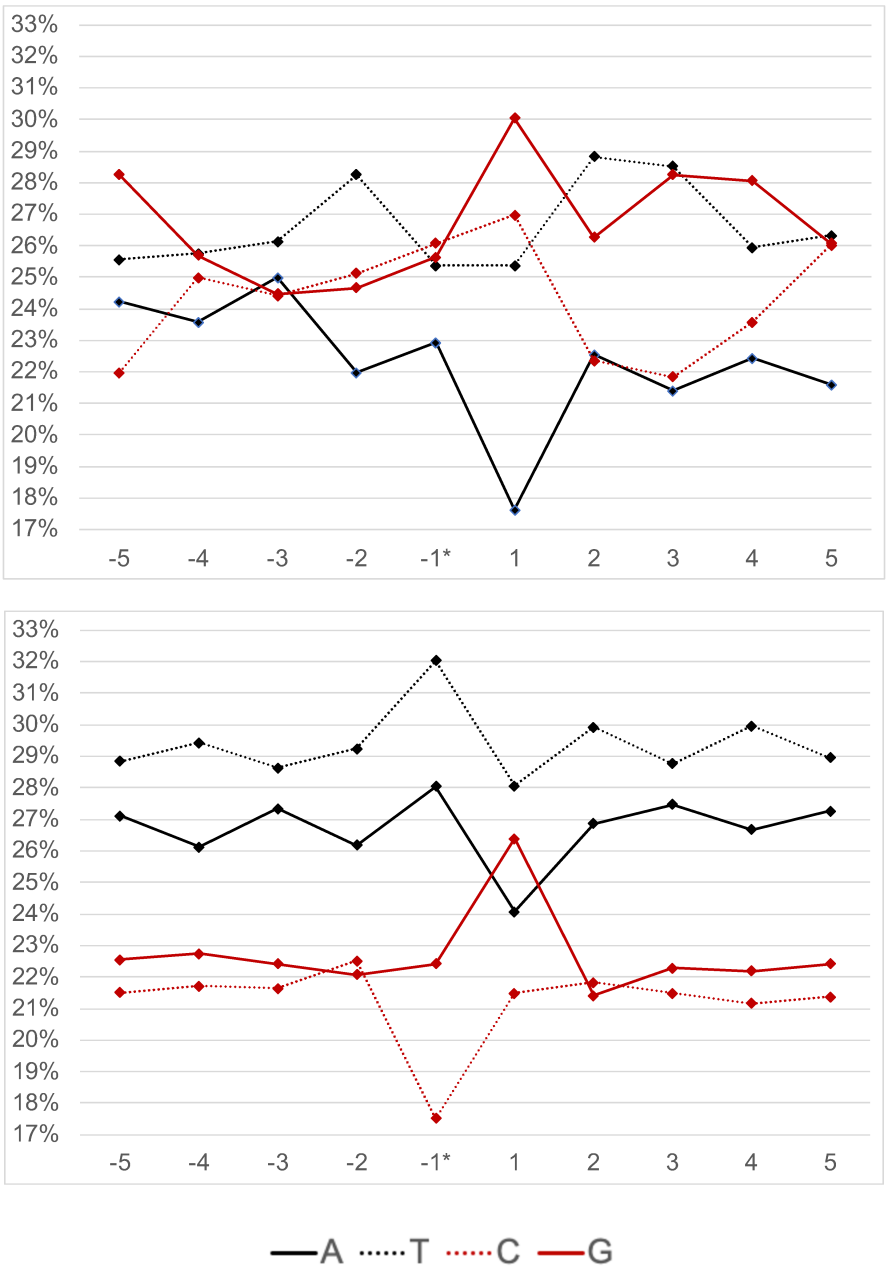


Figura Suplementar 57: Ocorrência de bases nas proximidades do SNP K (G/T) e a frequência basal esperada no genoma, nas regiões 5'UTR em *Rattus norvegicus*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency^.

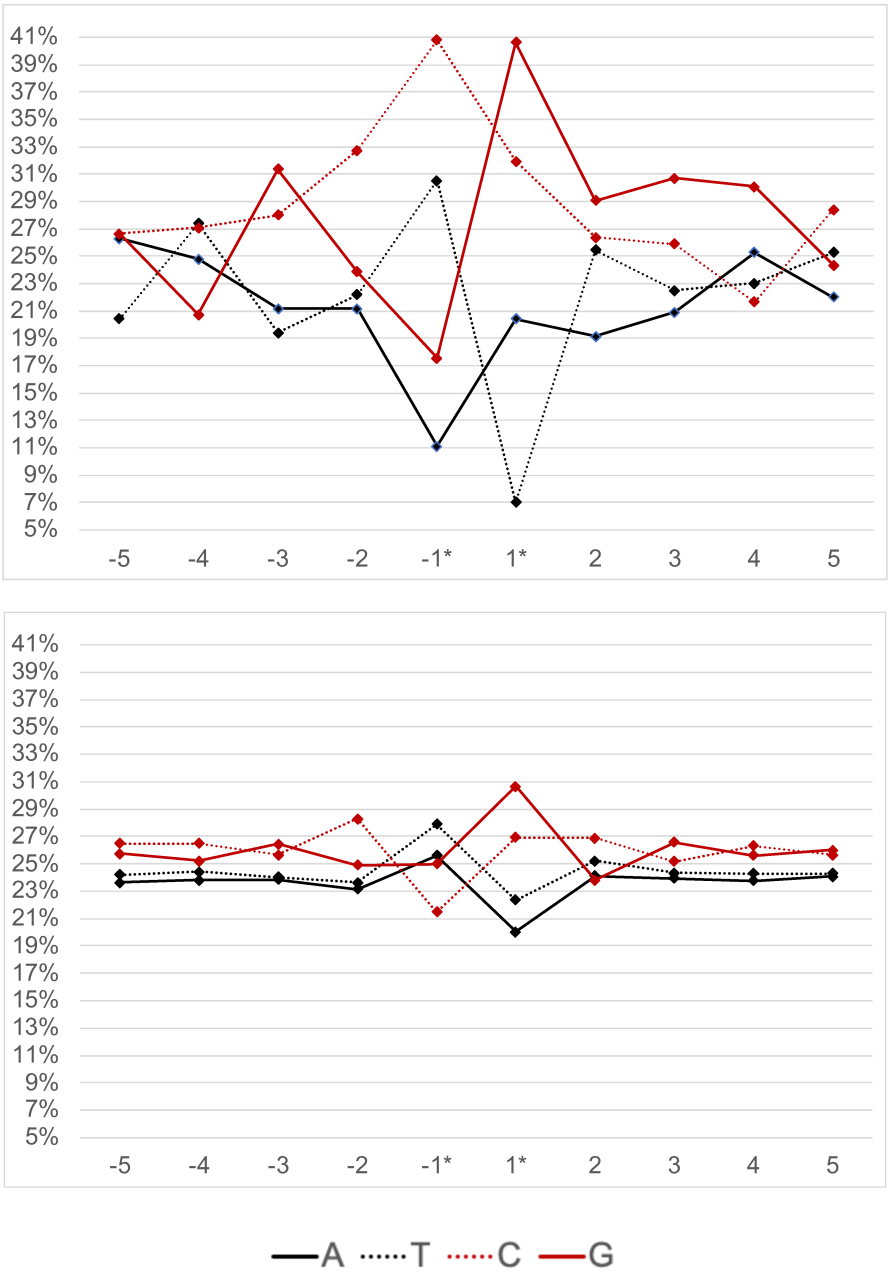


Figura Suplementar 58: Ocorrência de bases nas proximidades do SNP K (G/T) e a frequência basal esperada no genoma, nas regiões sinônimas em *Rattus norvegicus*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ˆbaseline frequencyˆ.

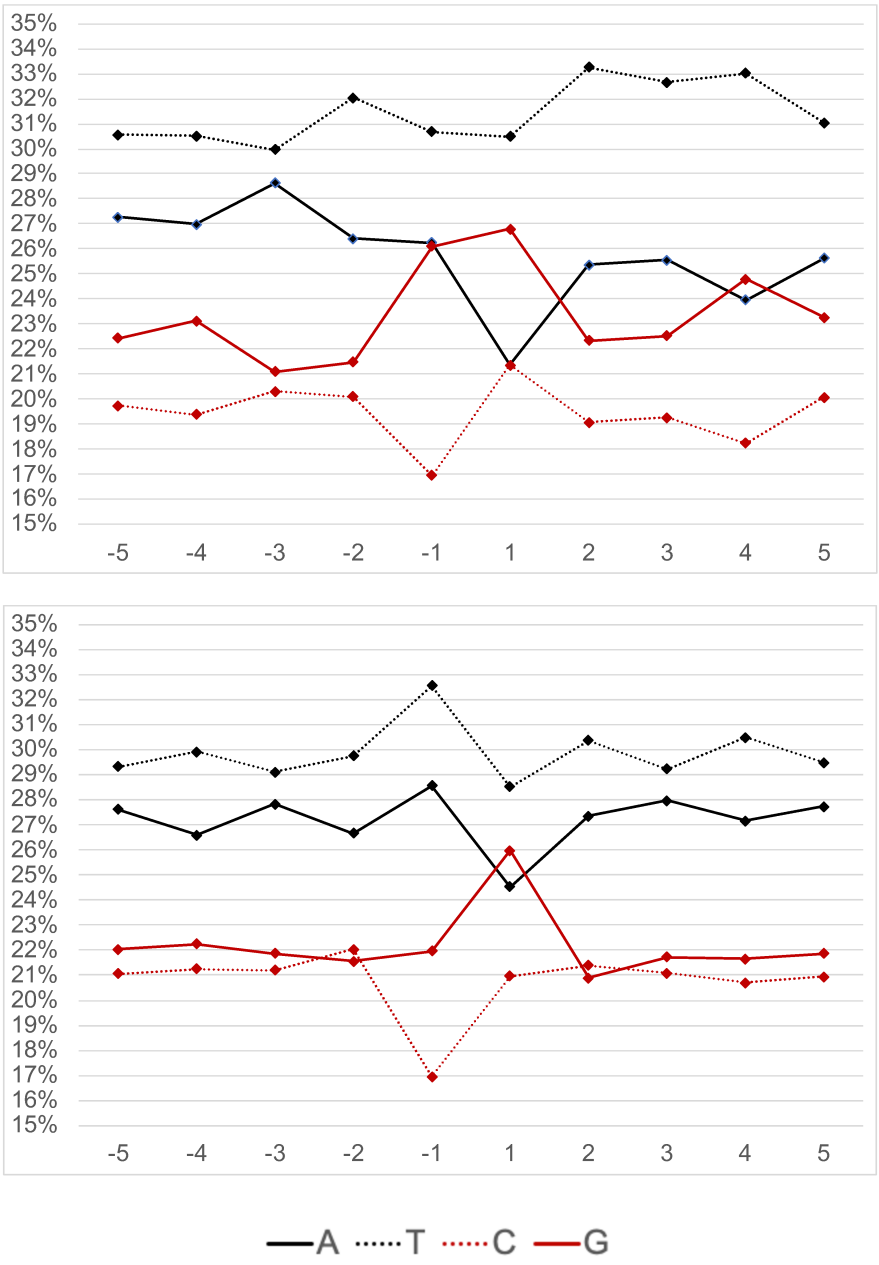


Figura Suplementar 59: Ocorrência de bases nas proximidades do SNP K (G/T) e a frequência basal esperada no genoma, nas regiões intrônicas em *Rattus norvegicus*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

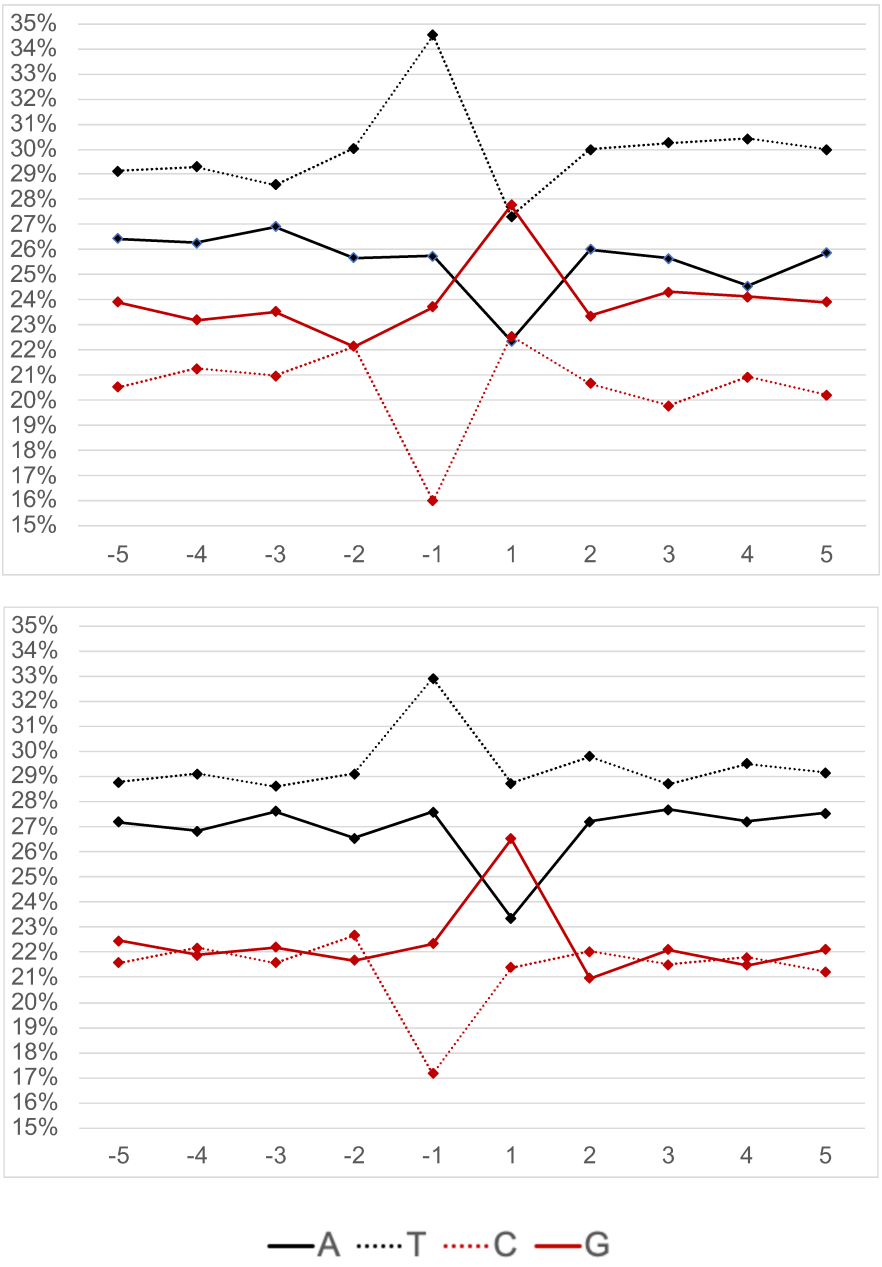


Figura Suplementar 60: Ocorrência de bases nas proximidades do SNP K (G/T) e a frequência basal esperada no genoma, nas regiões 3'UTR em *Sus scrofa*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

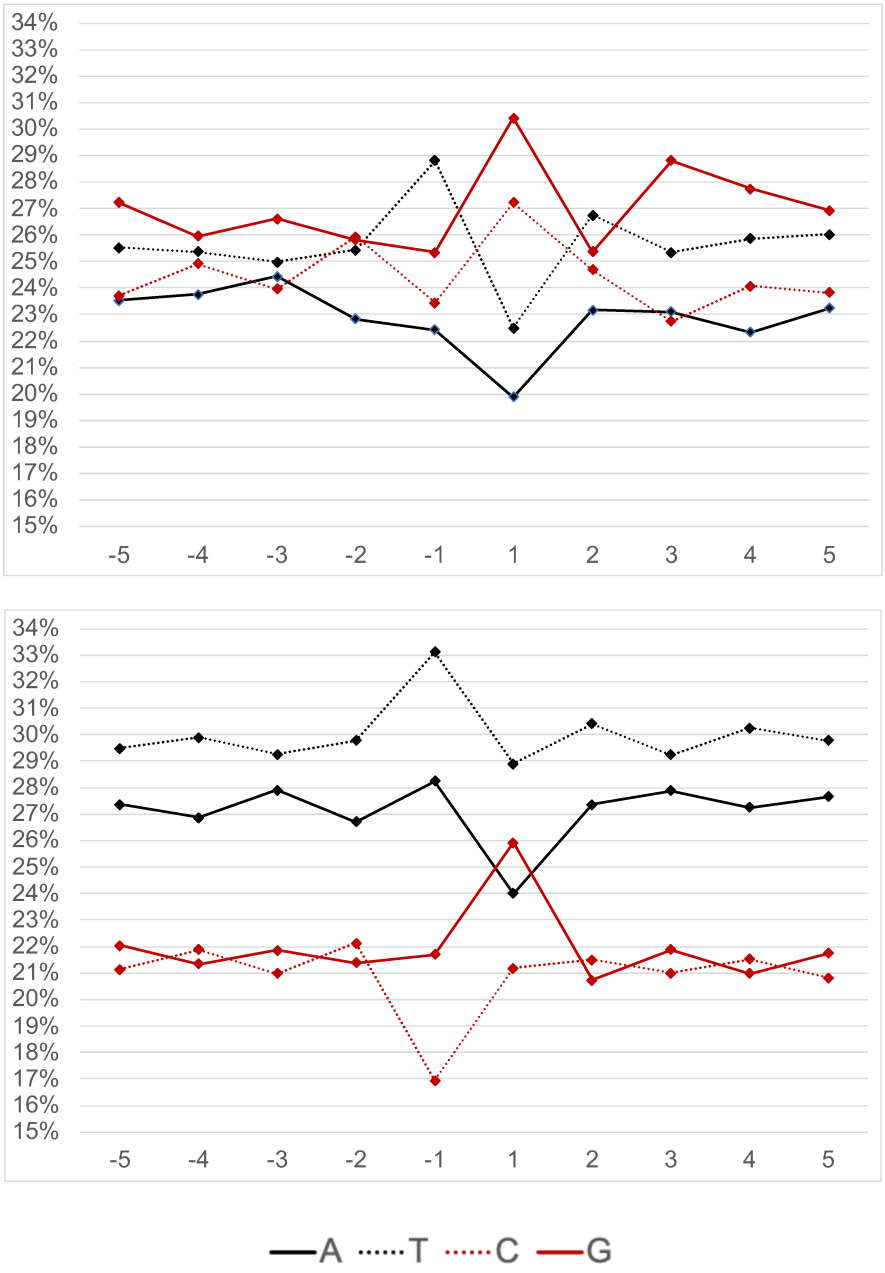


Figura Suplementar 61: Ocorrência de bases nas proximidades do SNP K (G/T) e a frequência basal esperada no genoma, nas regiões 5'UTR em *Sus scrofa*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

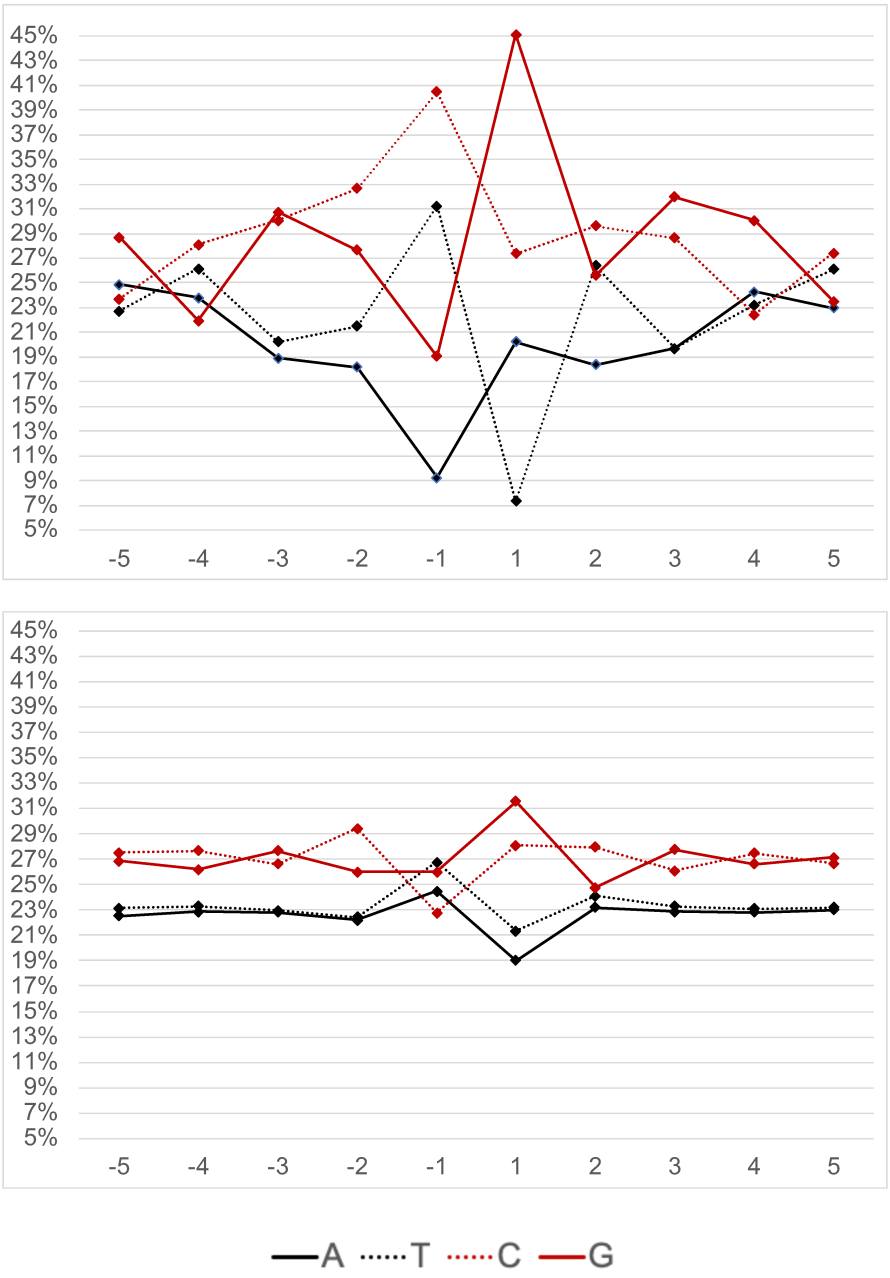


Figura Suplementar 62: Ocorrência de bases nas proximidades do SNP K (G/T) e a frequência basal esperada no genoma, nas regiões sinônimas em *Sus scrofa*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

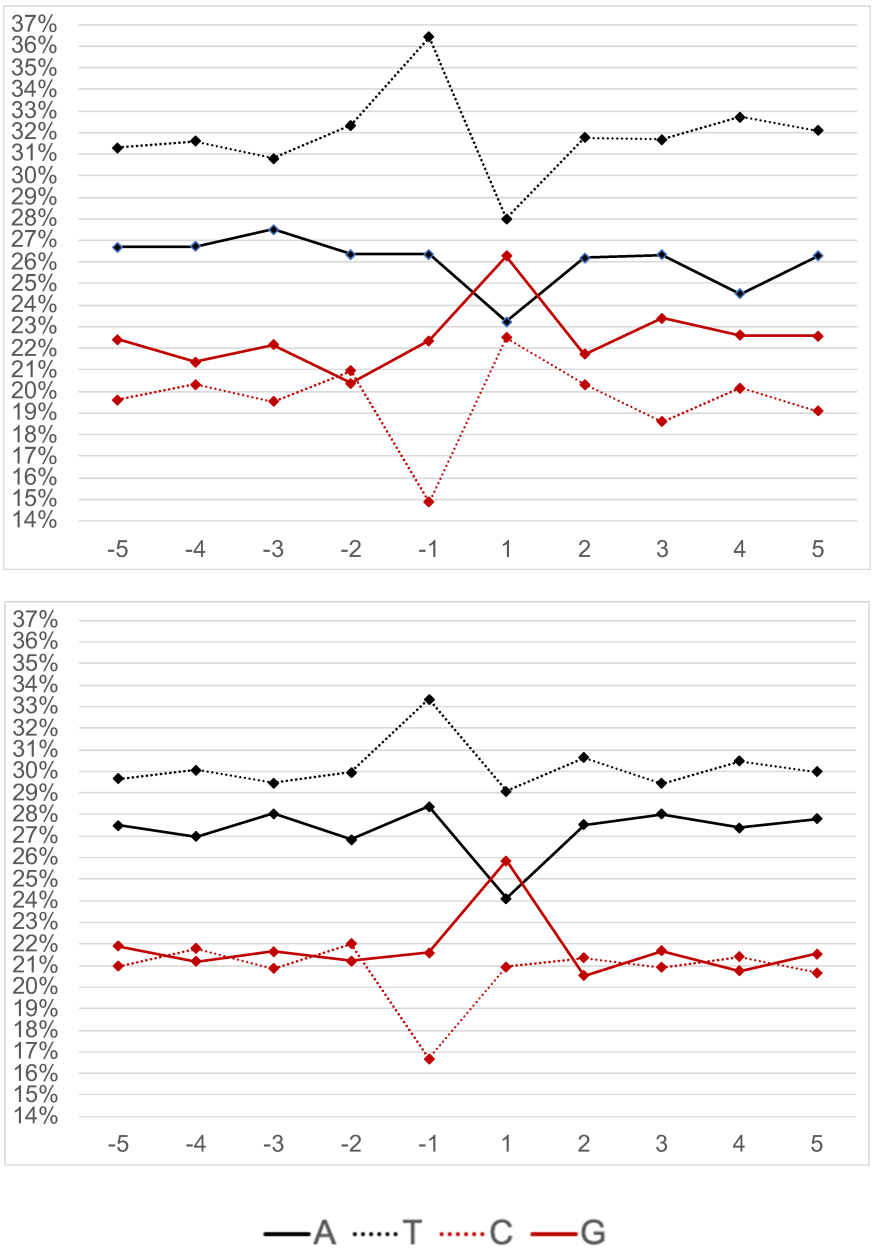


Figura Suplementar 63: Ocorrência de bases nas proximidades do SNP K (G/T) e a frequência basal esperada no genoma, nas regiões intrônicas em *Sus scrofa*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

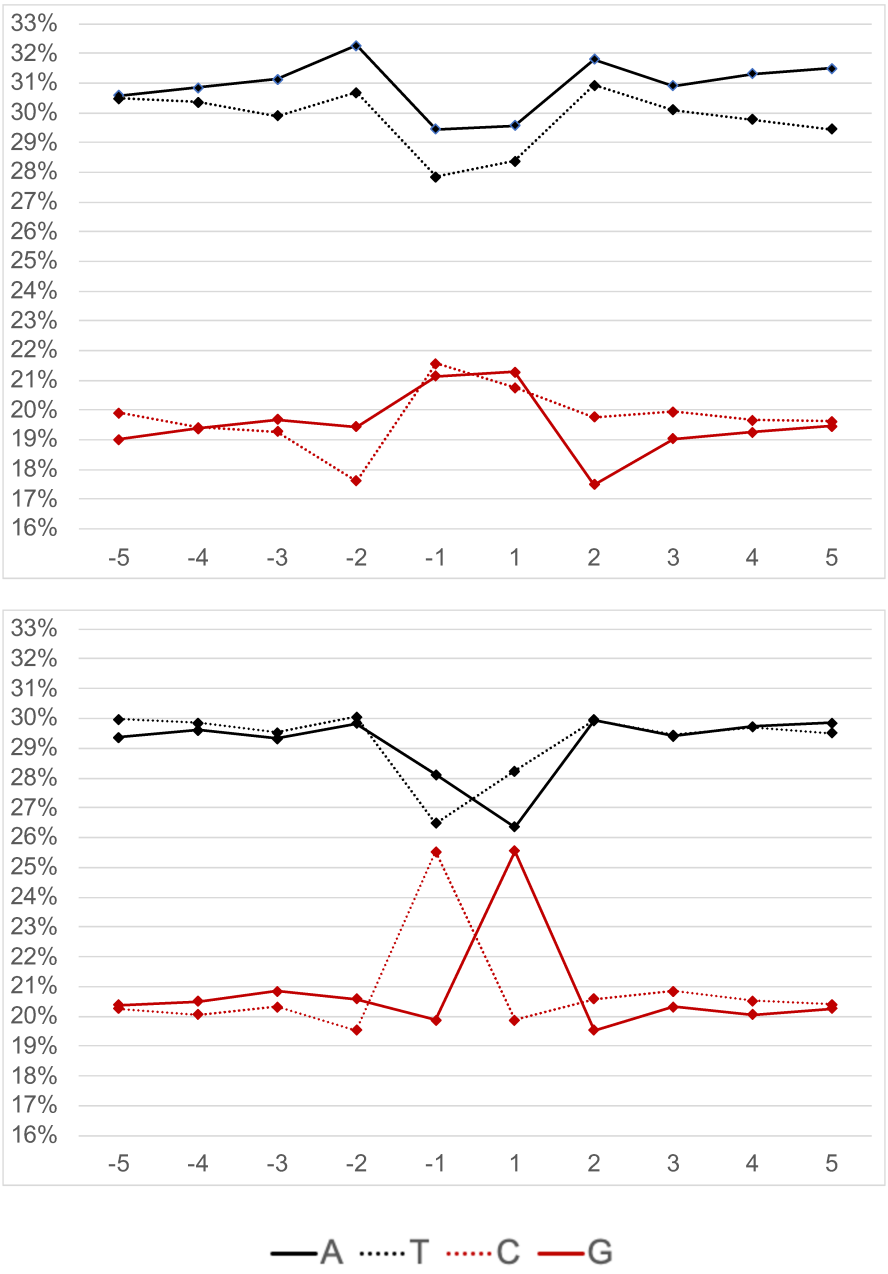


Figura Suplementar 64: Ocorrência de bases nas proximidades do SNP W (A/T) e a frequência basal esperada no genoma, nas regiões 3'UTR em bovinos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

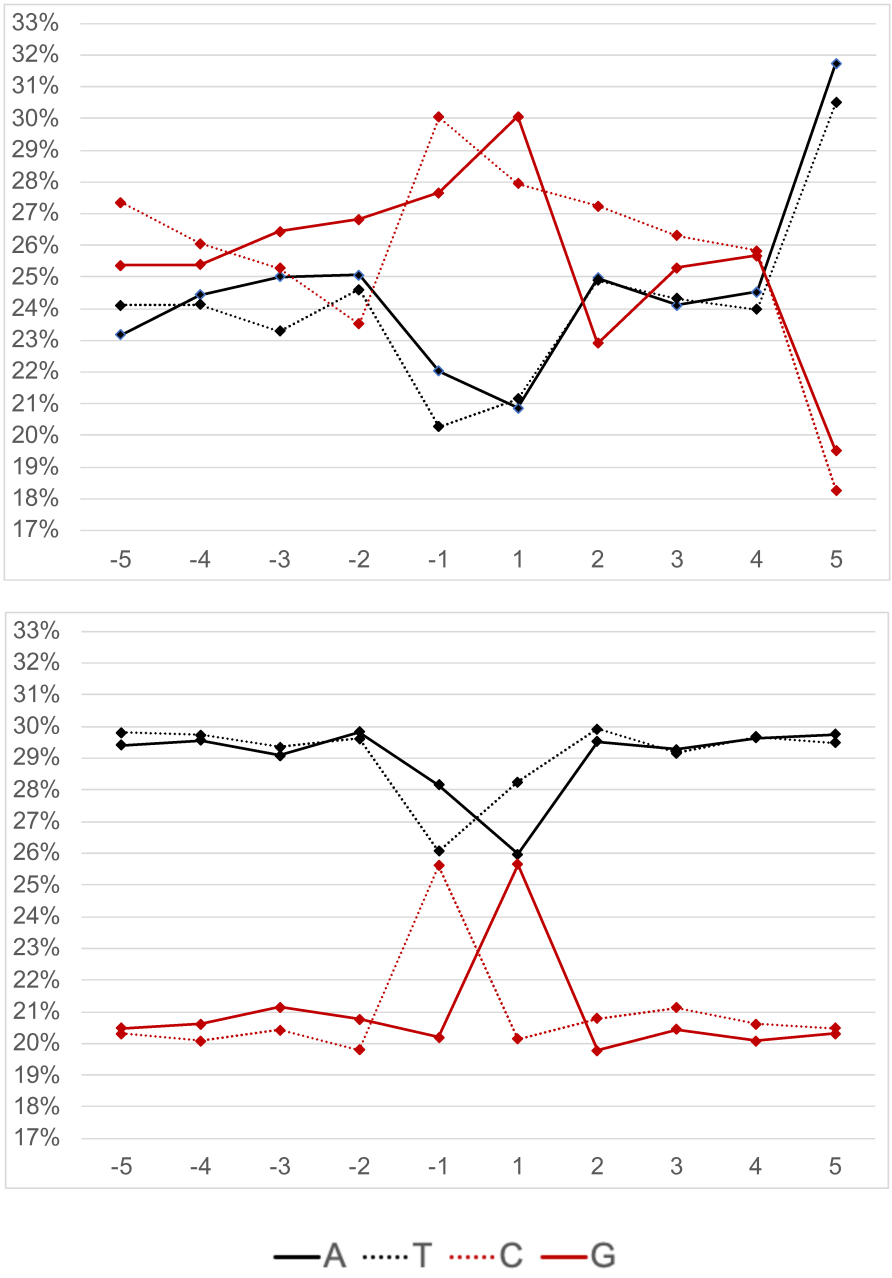


Figura Suplementar 65: Ocorrência de bases nas proximidades do SNP W (A/T) e a frequência basal esperada no genoma, nas regiões 5'UTR em bovinos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

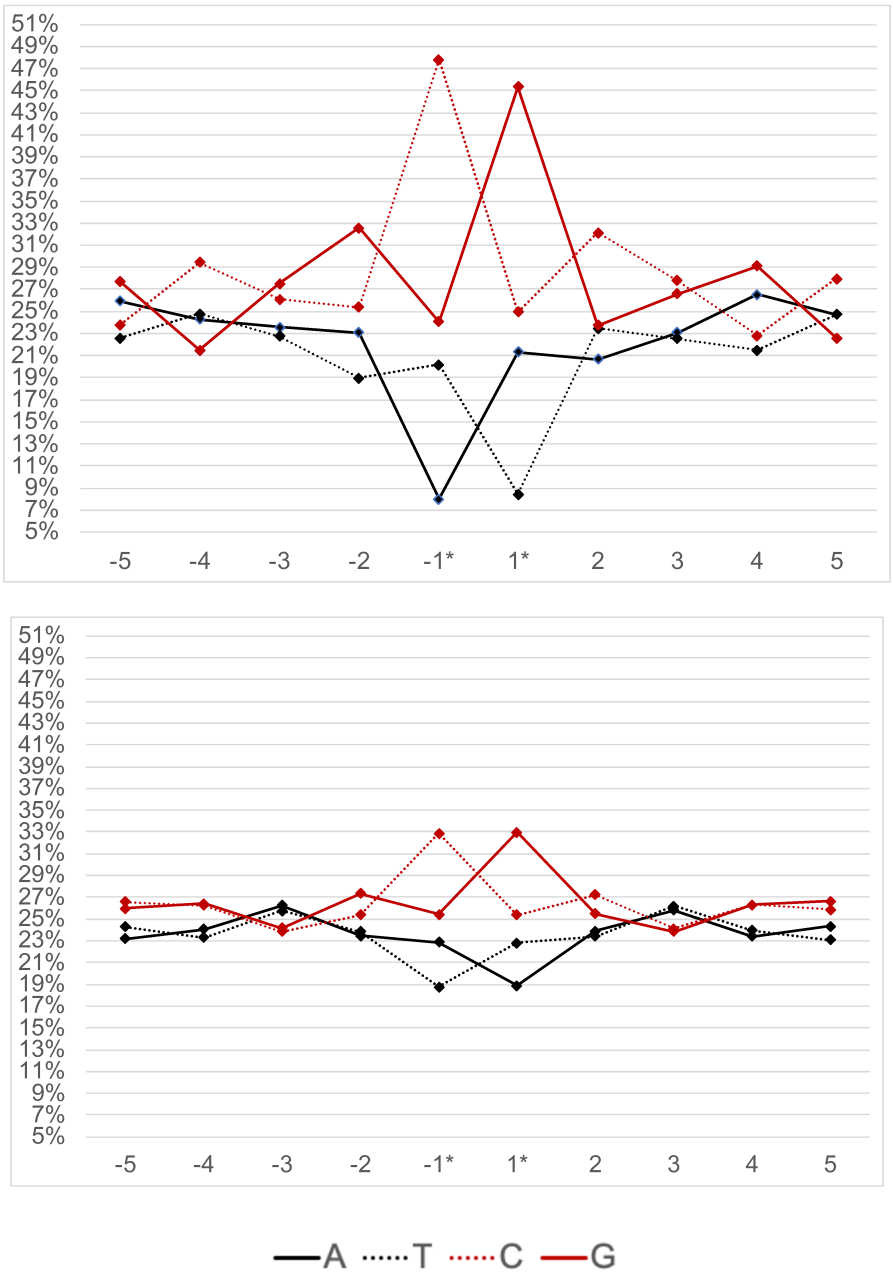


Figura Suplementar 66: Ocorrência de bases nas proximidades do SNP W (A/T) e a frequência basal esperada no genoma, nas regiões sinônimas em bovinos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency^.

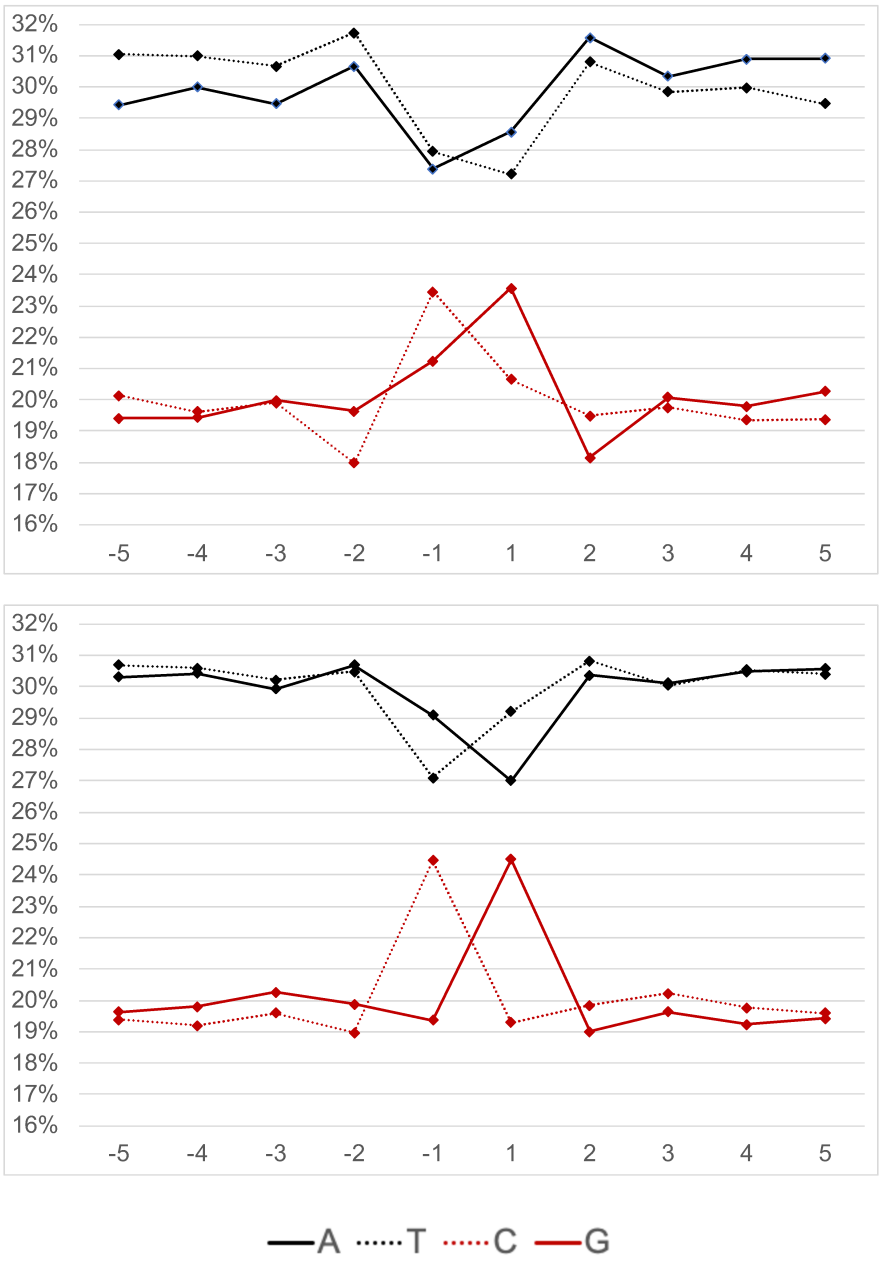


Figura Suplementar 67: Ocorrência de bases nas proximidades do SNP W (A/T) e a frequência basal esperada no genoma, nas regiões intrônicas de bovinos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

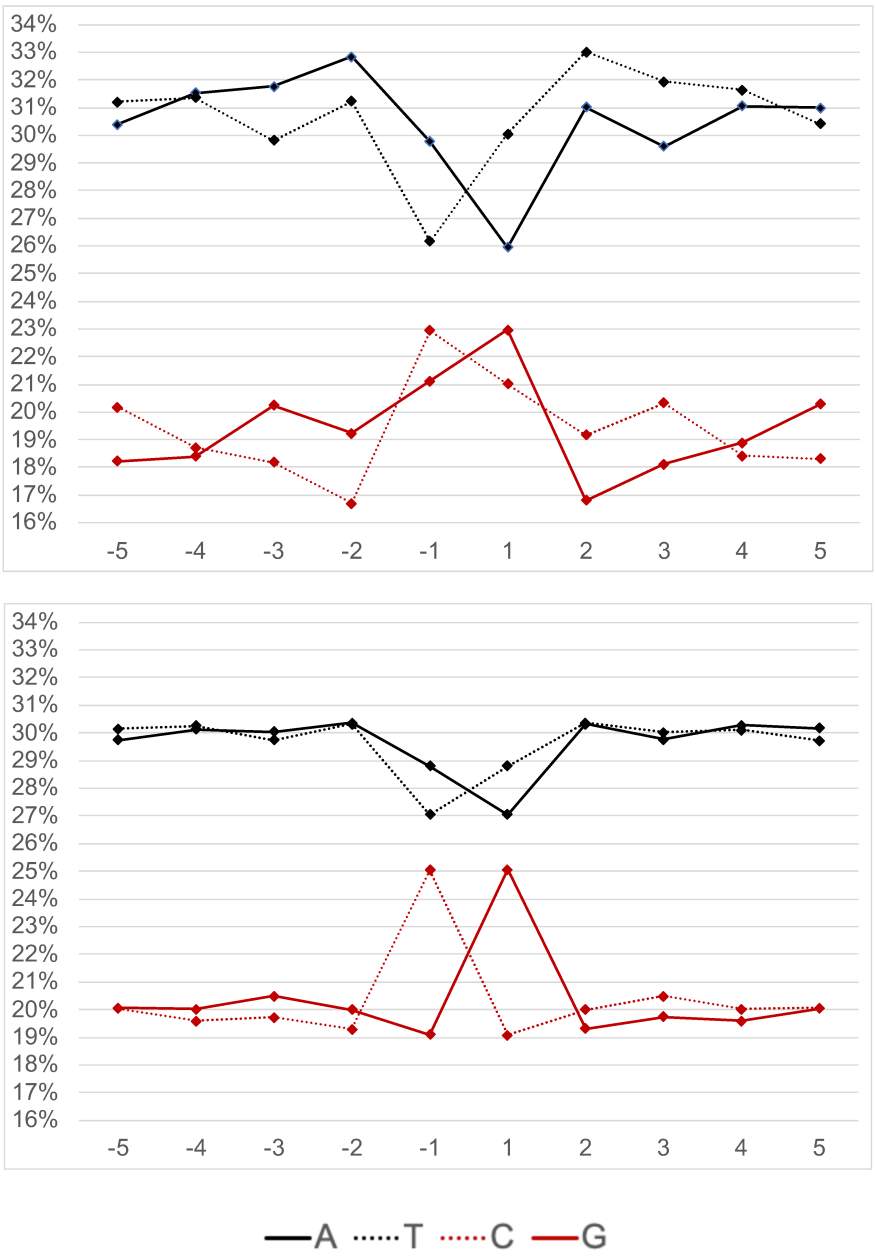


Figura Suplementar 68: Ocorrência de bases nas proximidades do SNP W (A/T) e a frequência basal esperada no genoma, nas regiões 3'UTR em humanos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

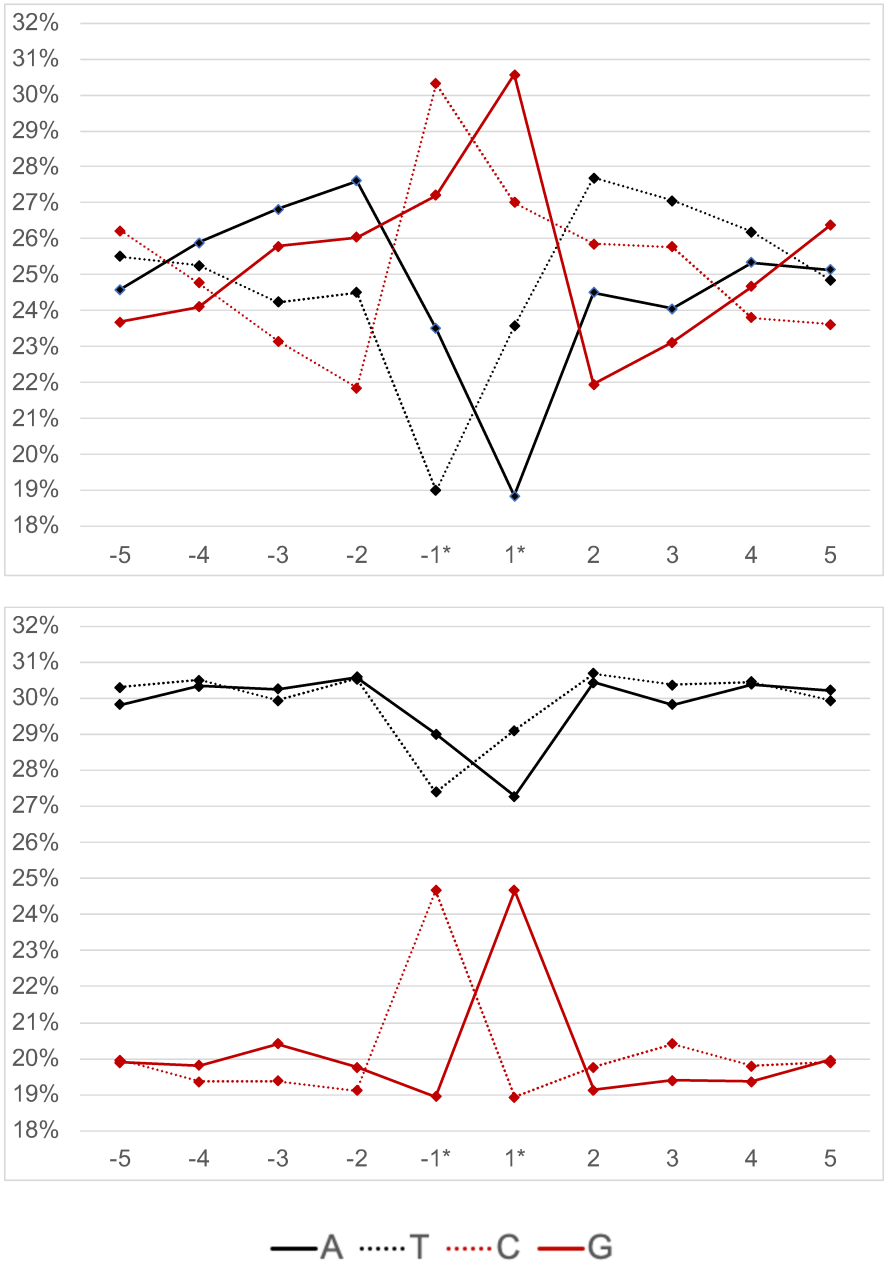


Figura Suplementar 69: Ocorrência de bases nas proximidades do SNP W (A/T) e a frequência basal esperada no genoma, nas regiões 5'UTR em humanos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency^.

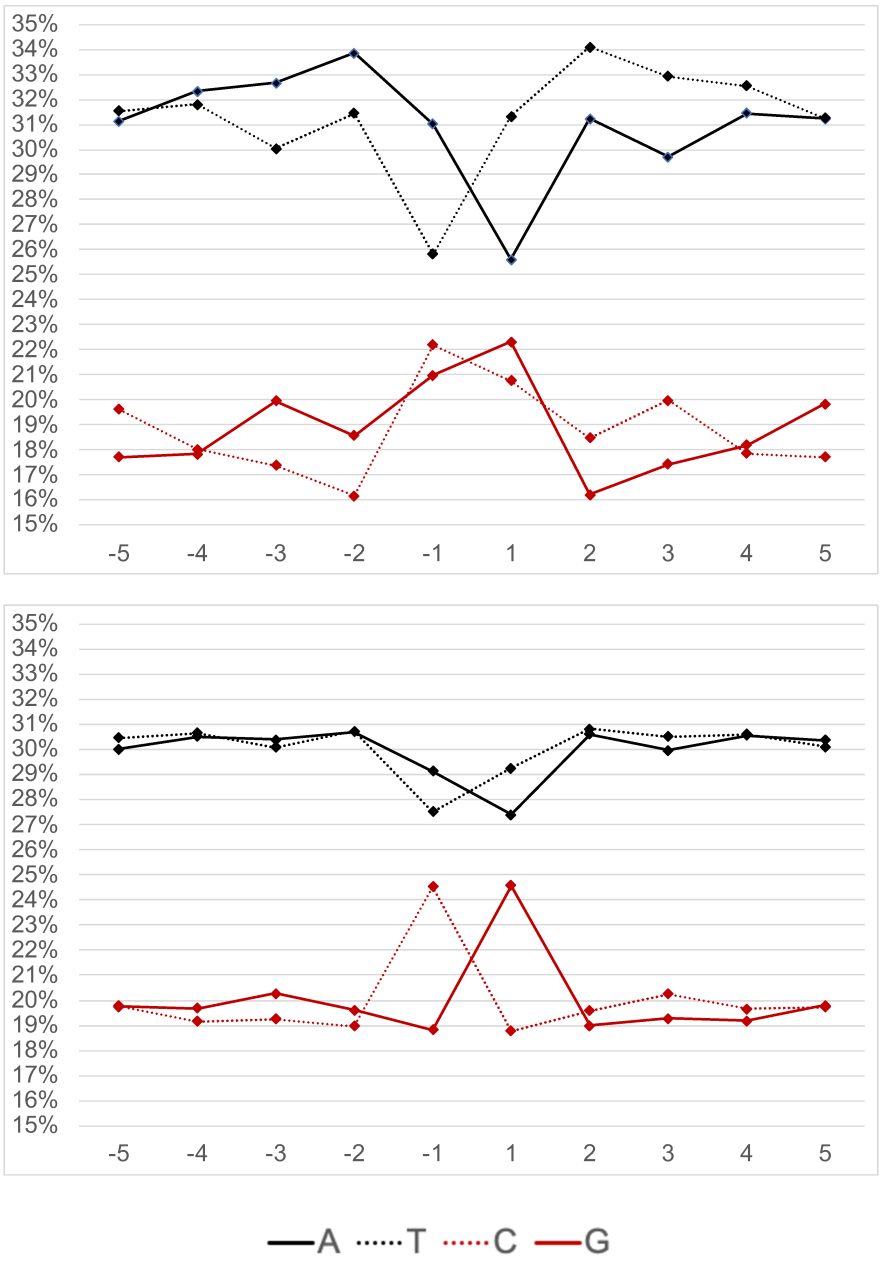


Figura Suplementar 70: Ocorrência de bases nas proximidades do SNP W (A/T) e a frequência basal esperada no genoma, nas regiões intrônicas em humanos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

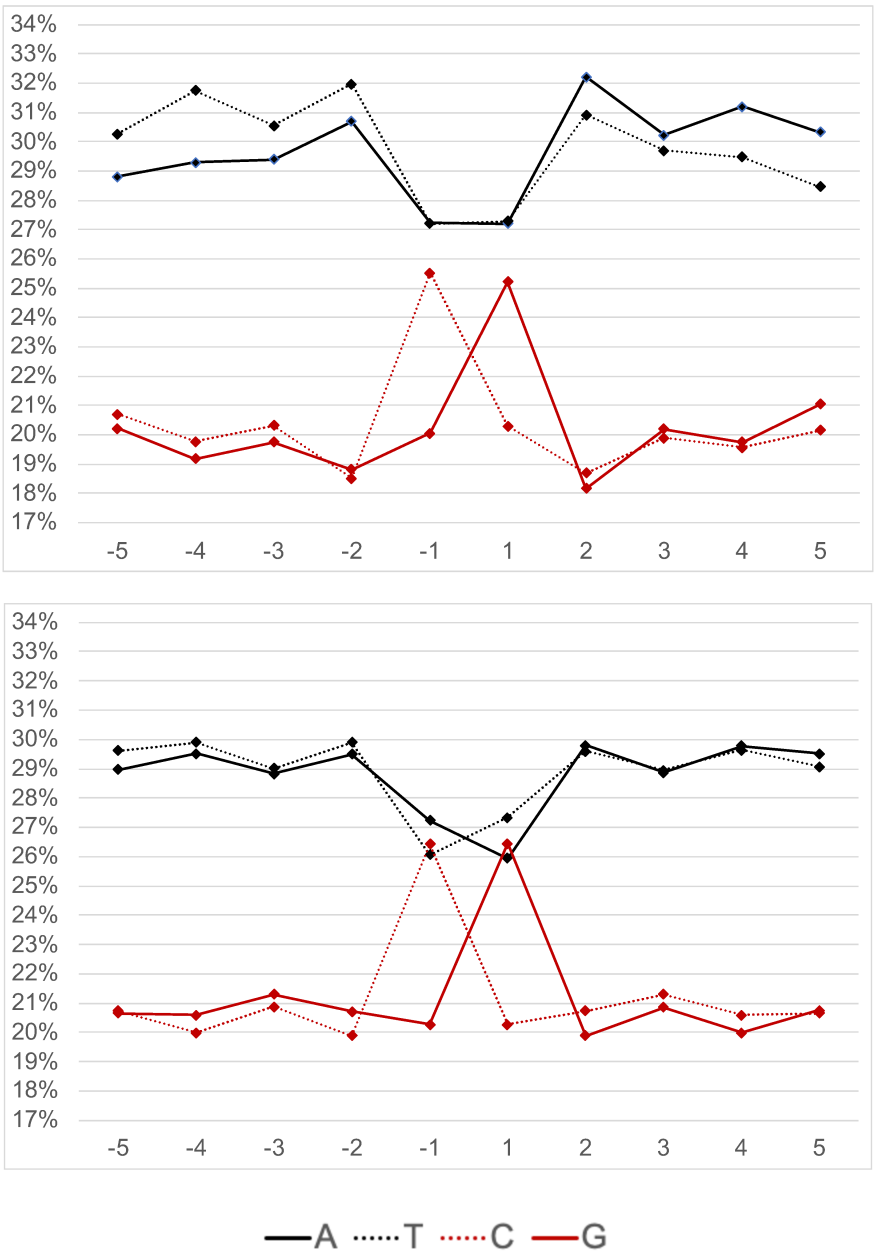


Figura Suplementar 71: Ocorrência de bases nas proximidades do SNP W (A/T) e a frequência basal esperada no genoma, nas regiões 3'UTR em camundongos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

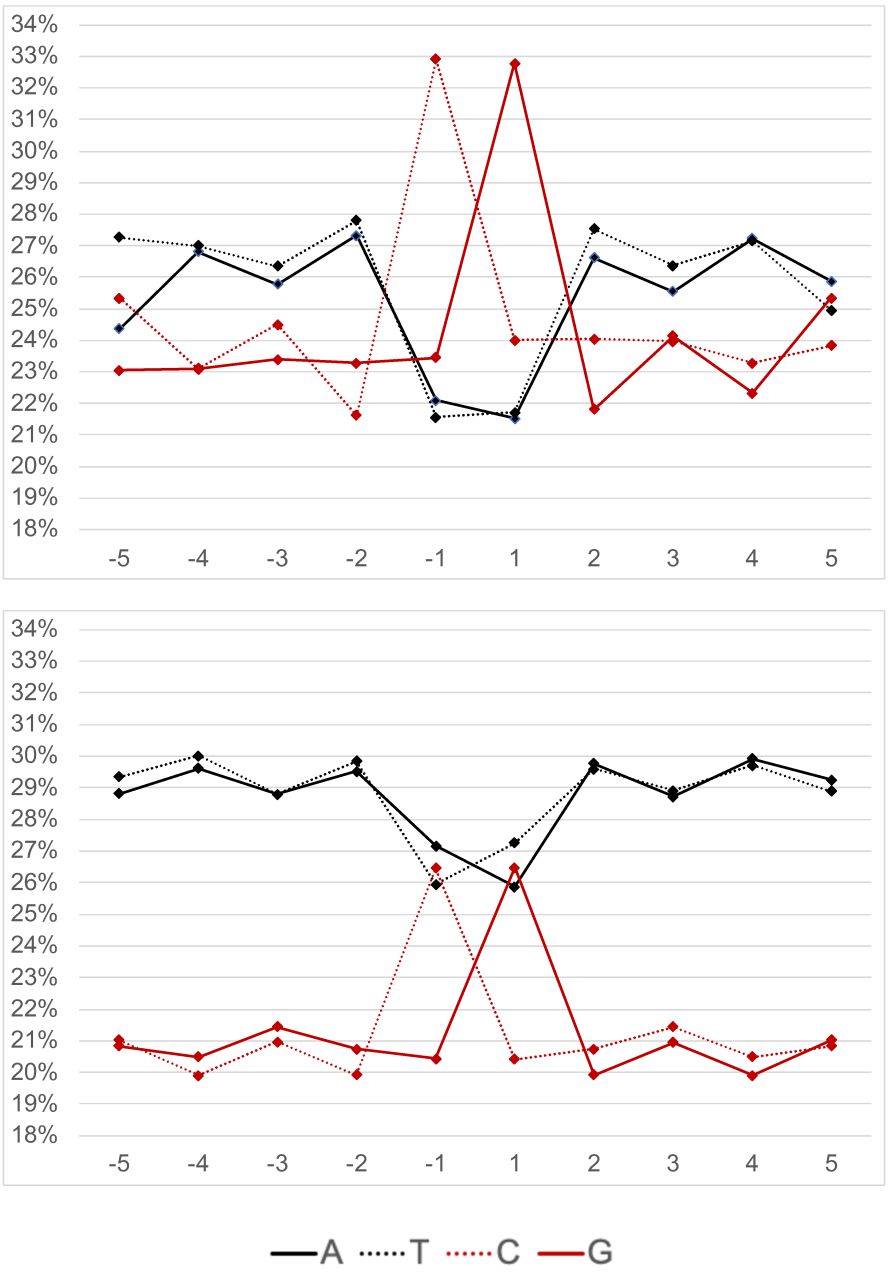


Figura Suplementar 72: Ocorrência de bases nas proximidades do SNP W (A/T) e a frequência basal esperada no genoma, nas regiões 5'UTR em camundongos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

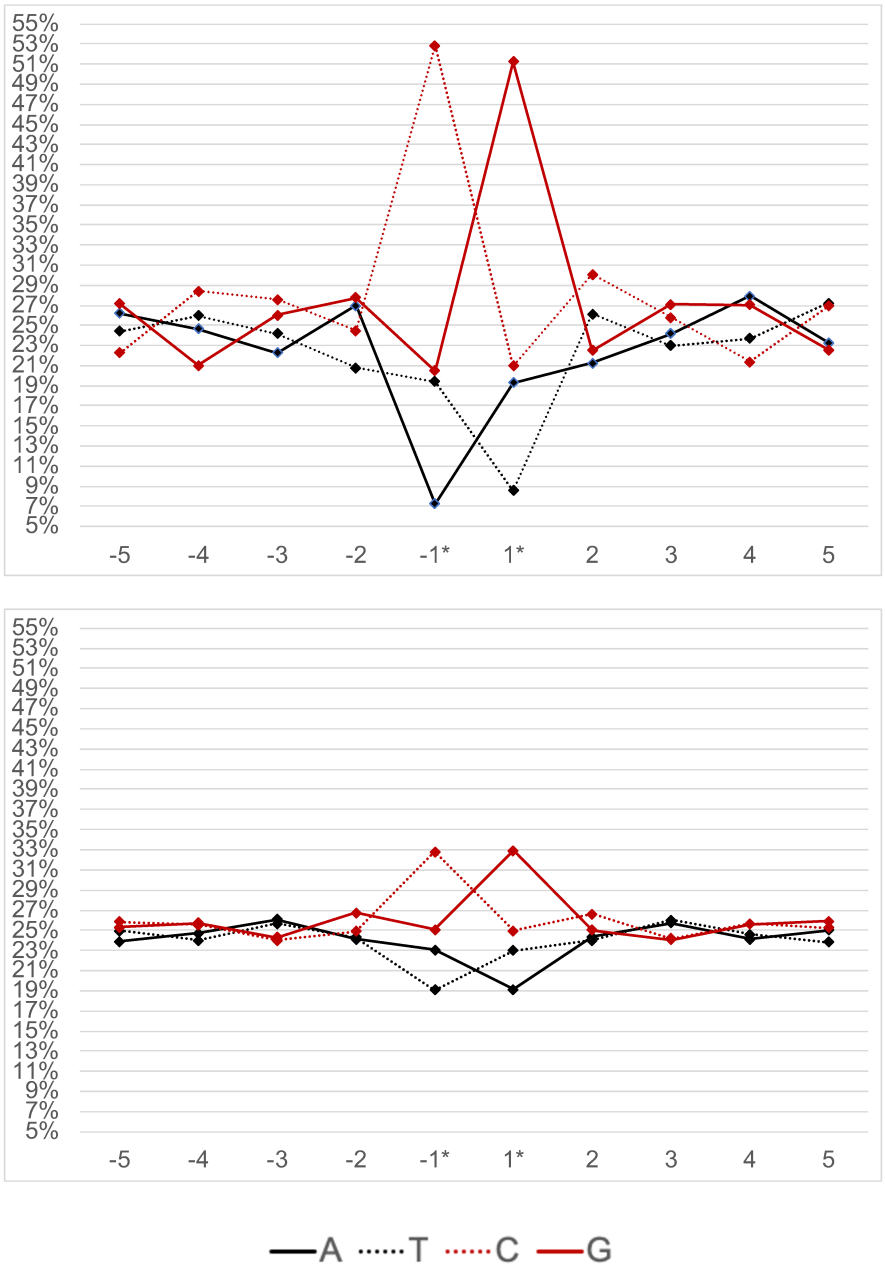


Figura Suplementar 73: Ocorrência de bases nas proximidades do SNP W (A/T) e a frequência basal esperada no genoma, nas regiões sinônimas em camundongos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency^.

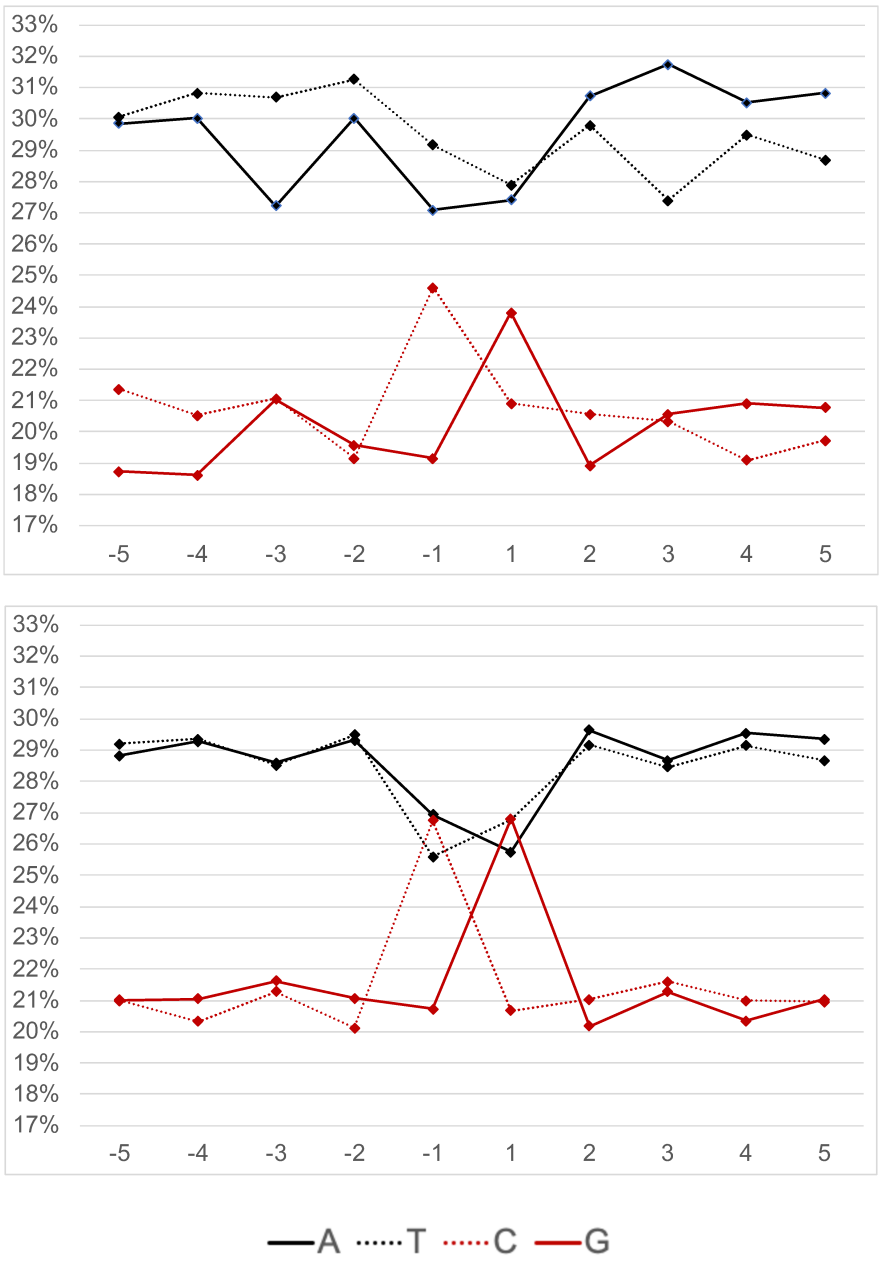


Figura Suplementar 74: Ocorrência de bases nas proximidades do SNP W (A/T) e a frequência basal esperada no genoma, nas regiões 3'UTR em *Rattus norvegicus*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

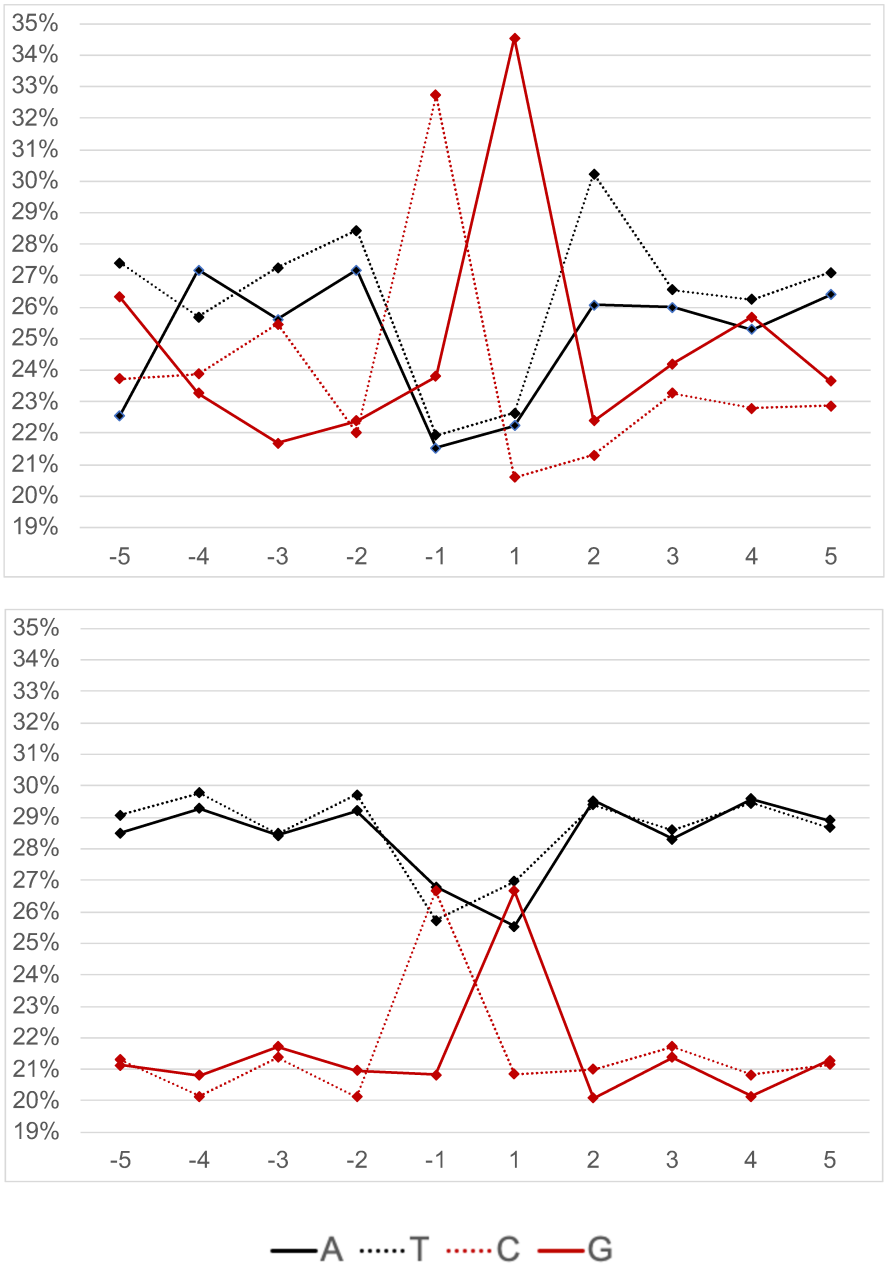


Figura Suplementar 75: Ocorrência de bases nas proximidades do SNP W (A/T) e a frequência basal esperada no genoma, nas regiões 5'UTR em *Rattus norvegicus*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

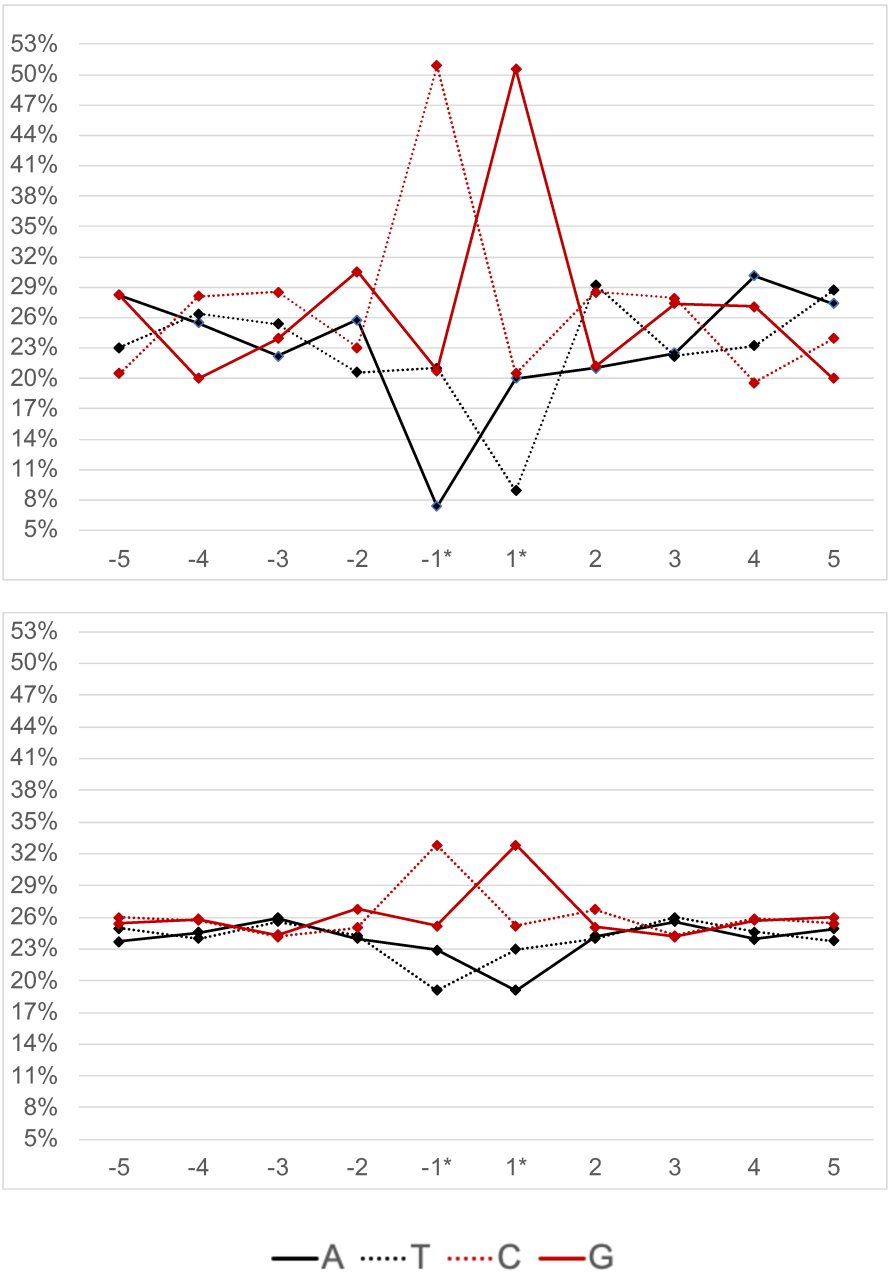


Figura Suplementar 76: Ocorrência de bases nas proximidades do SNP W (A/T) e a frequência basal esperada no genoma, nas regiões sinônimas em *Rattus norvegicus*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ˆbaseline frequencyˆ.

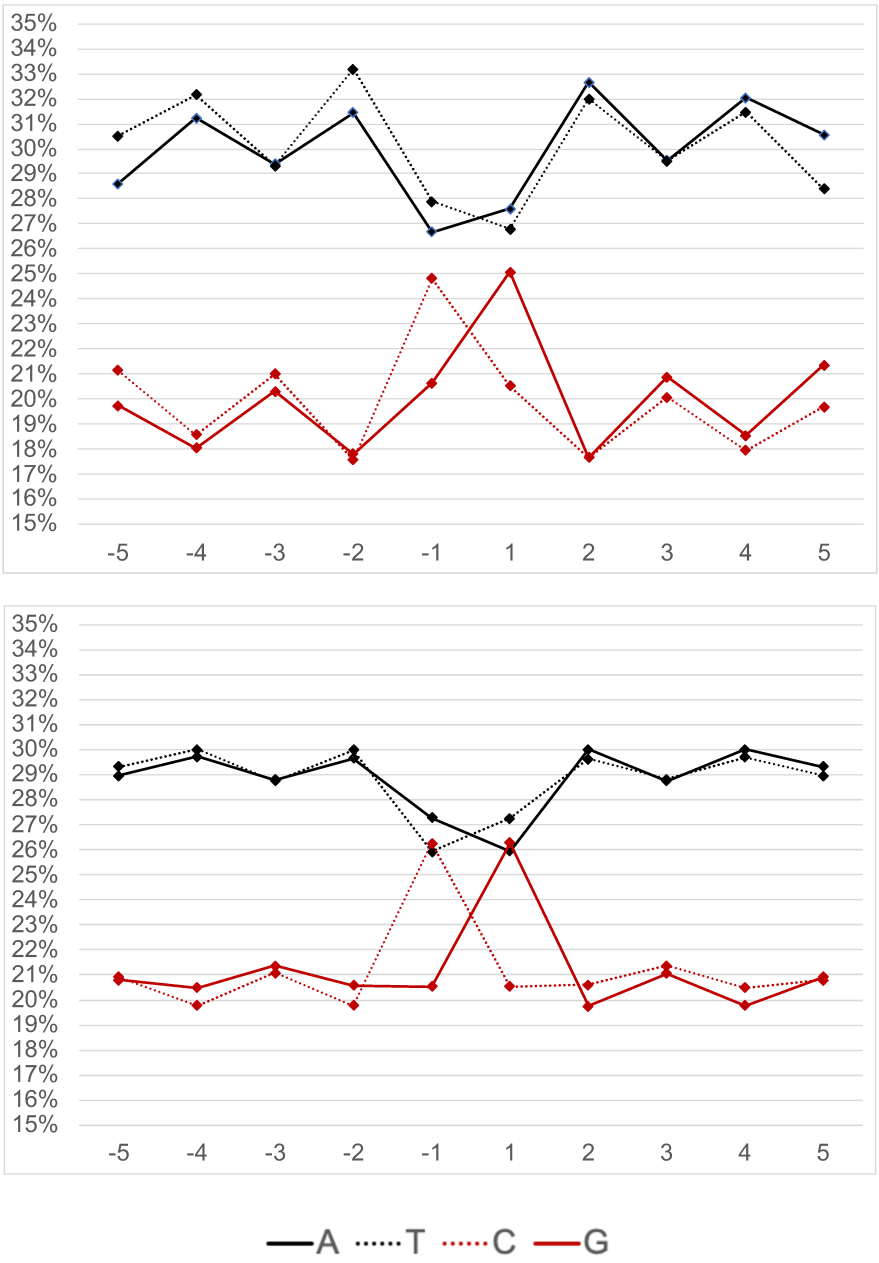


Figura Suplementar 77: Ocorrência de bases nas proximidades do SNP W (A/T) e a frequência basal esperada no genoma, nas regiões intrônicas em *Rattus norvegicus*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

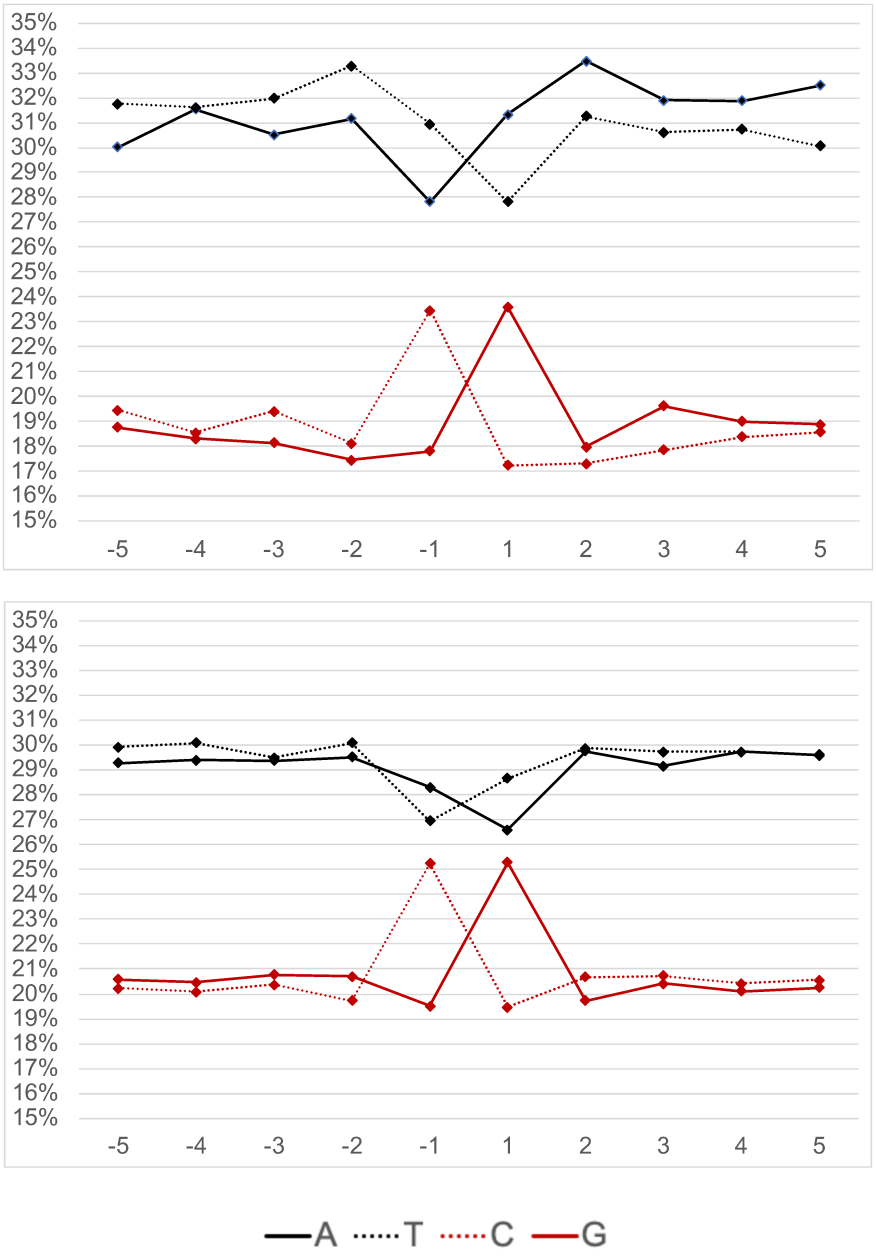


Figura Suplementar 78: Ocorrência de bases nas proximidades do SNP W (A/T) e a frequência basal esperada no genoma, nas regiões 3'UTR em *Sus scrofa*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

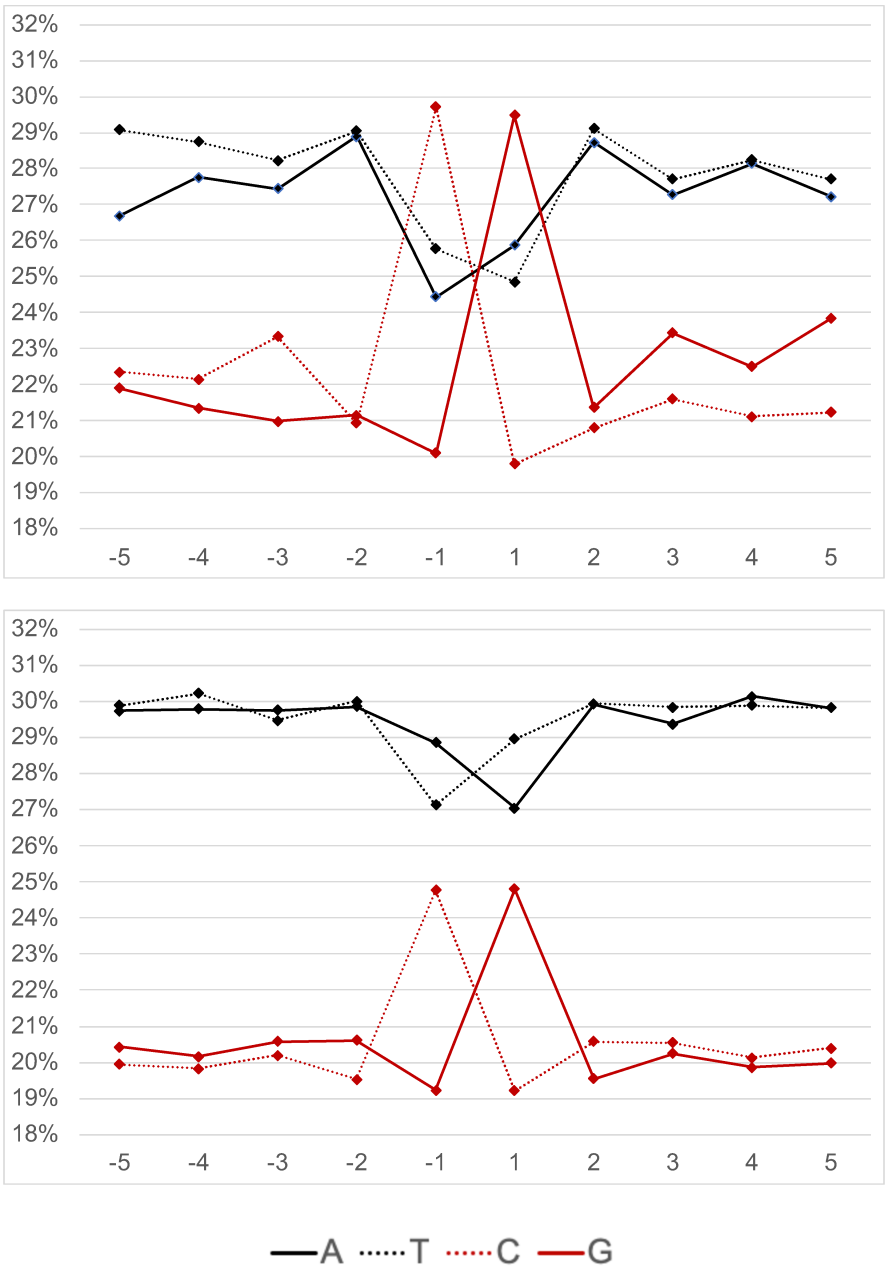


Figura Suplementar 79: Ocorrência de bases nas proximidades do SNP W (A/T) e a frequência basal esperada no genoma, nas regiões 5'UTR em *Sus scrofa*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

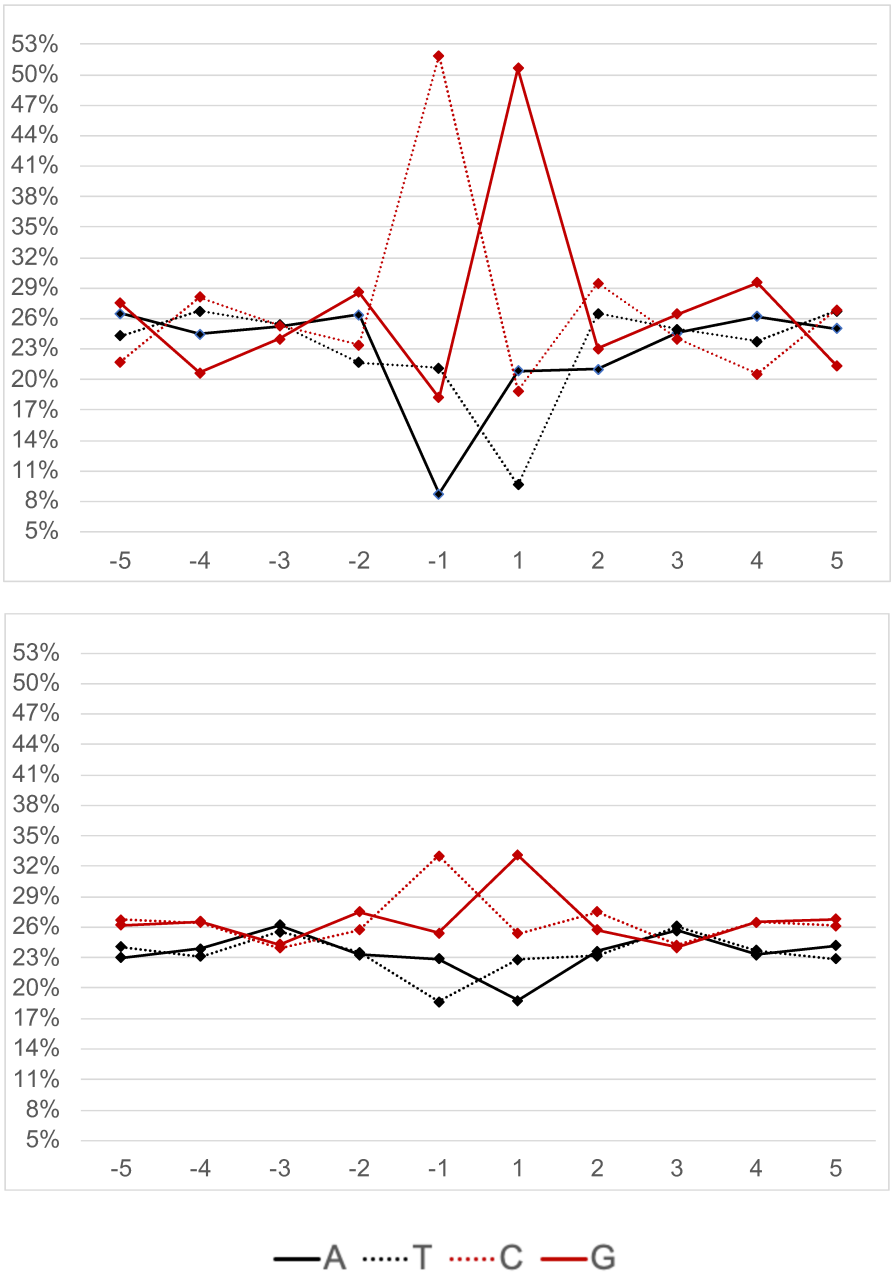


Figura Suplementar 80: Ocorrência de bases nas proximidades do SNP W (A/T) e a frequência basal esperada no genoma, nas regiões sinônimas em *Sus scrofa*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

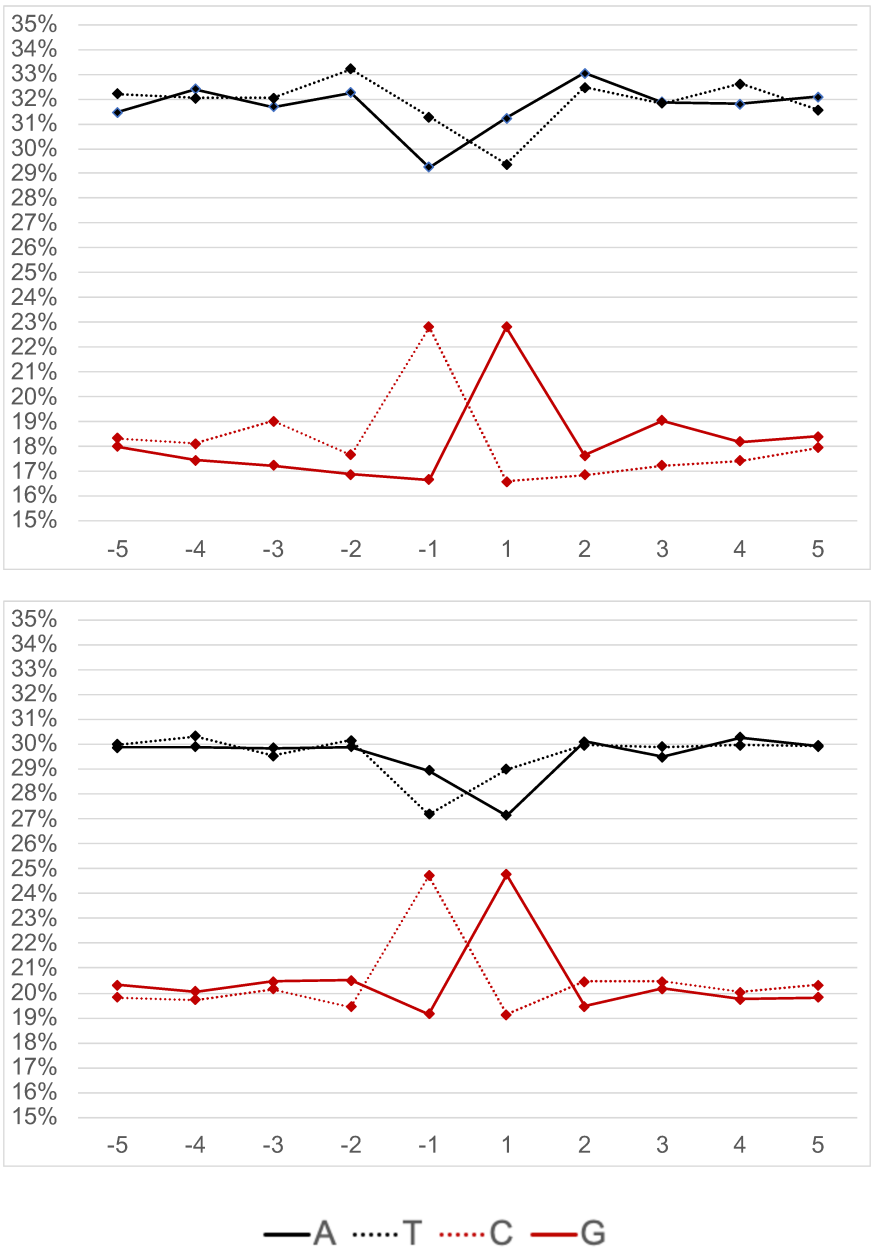


Figura Suplementar 81: Ocorrência de bases nas proximidades do SNP W (A/T) e a frequência basal esperada no genoma, nas regiões intrônicas em *Sus scrofa*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

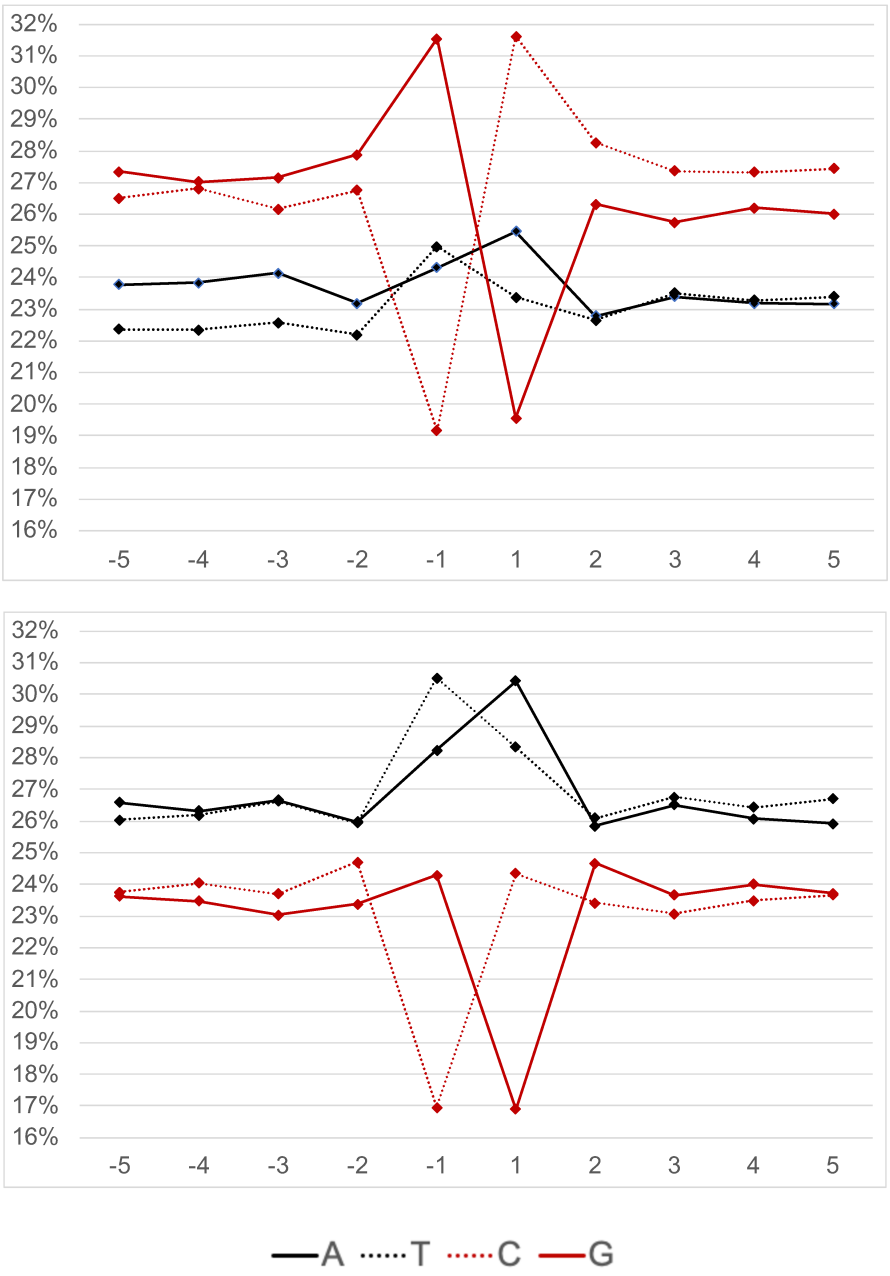


Figura Suplementar 82: Ocorrência de bases nas proximidades do SNP S (C/G) e a frequência basal esperada no genoma, nas regiões 3'UTR em bovinos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

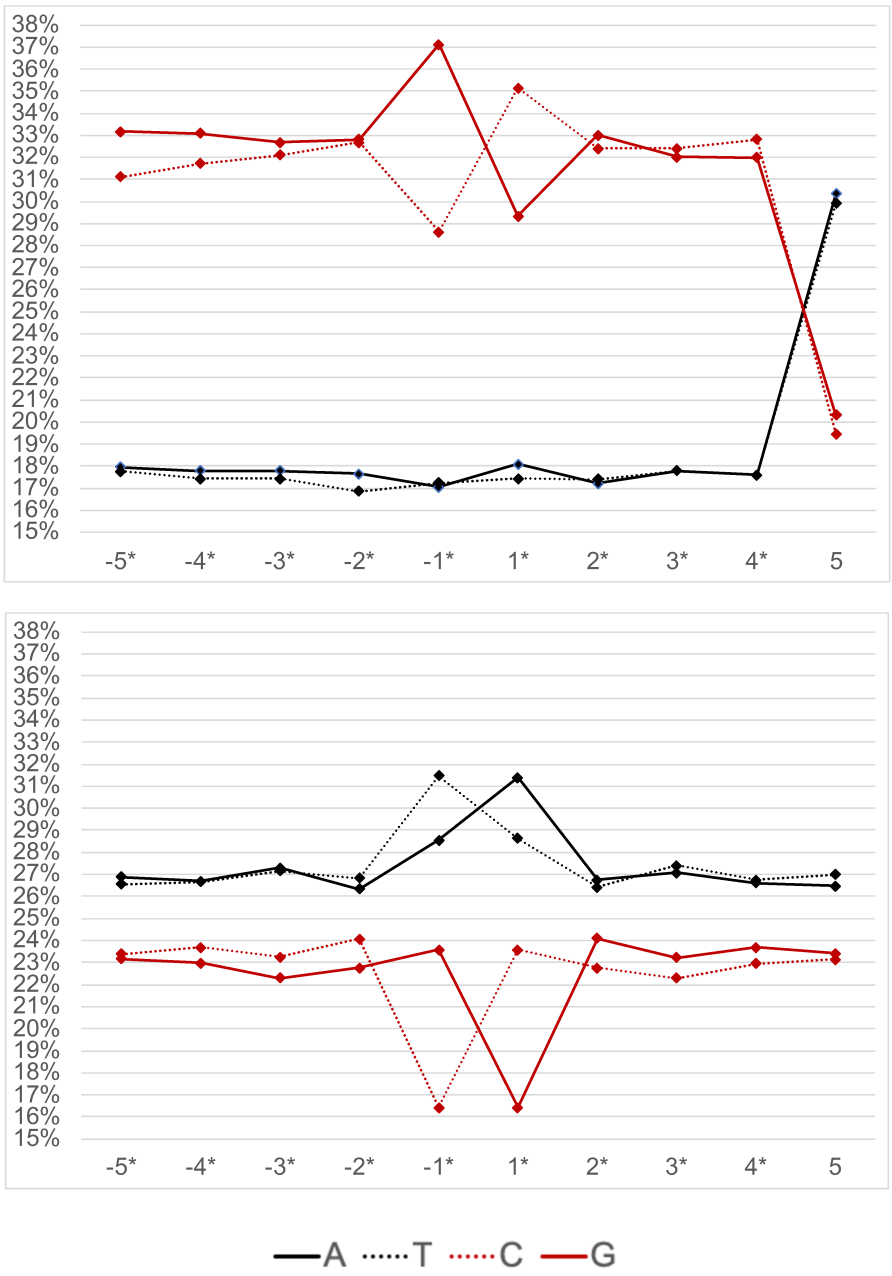


Figura Suplementar 83: Ocorrência de bases nas proximidades do SNP S (C/G) e a frequência basal esperada no genoma, nas regiões 5'UTR em bovinos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency^..

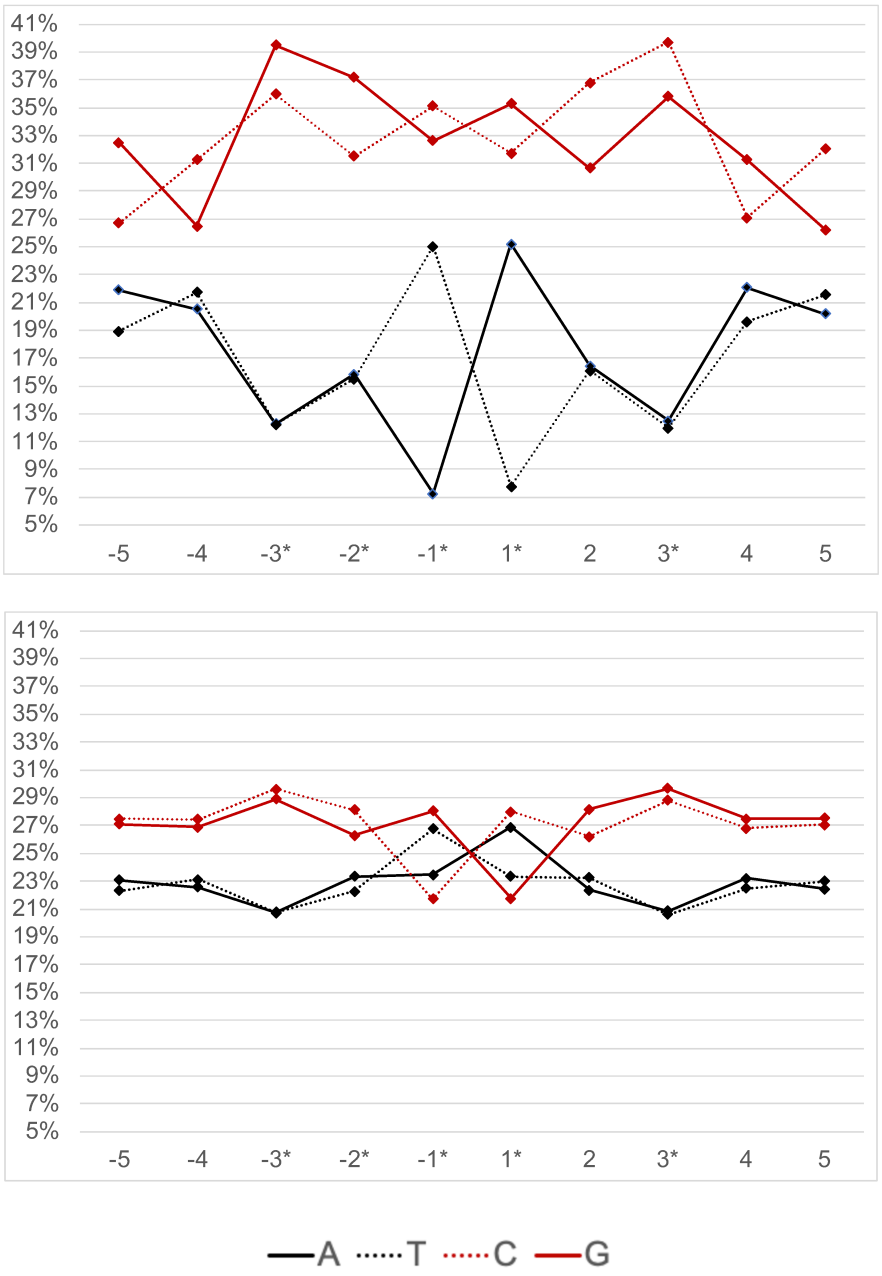


Figura Suplementar 84: Ocorrência de bases nas proximidades do SNP S (C/G) e a frequência basal esperada no genoma, nas regiões sinônimas em bovinos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency^.

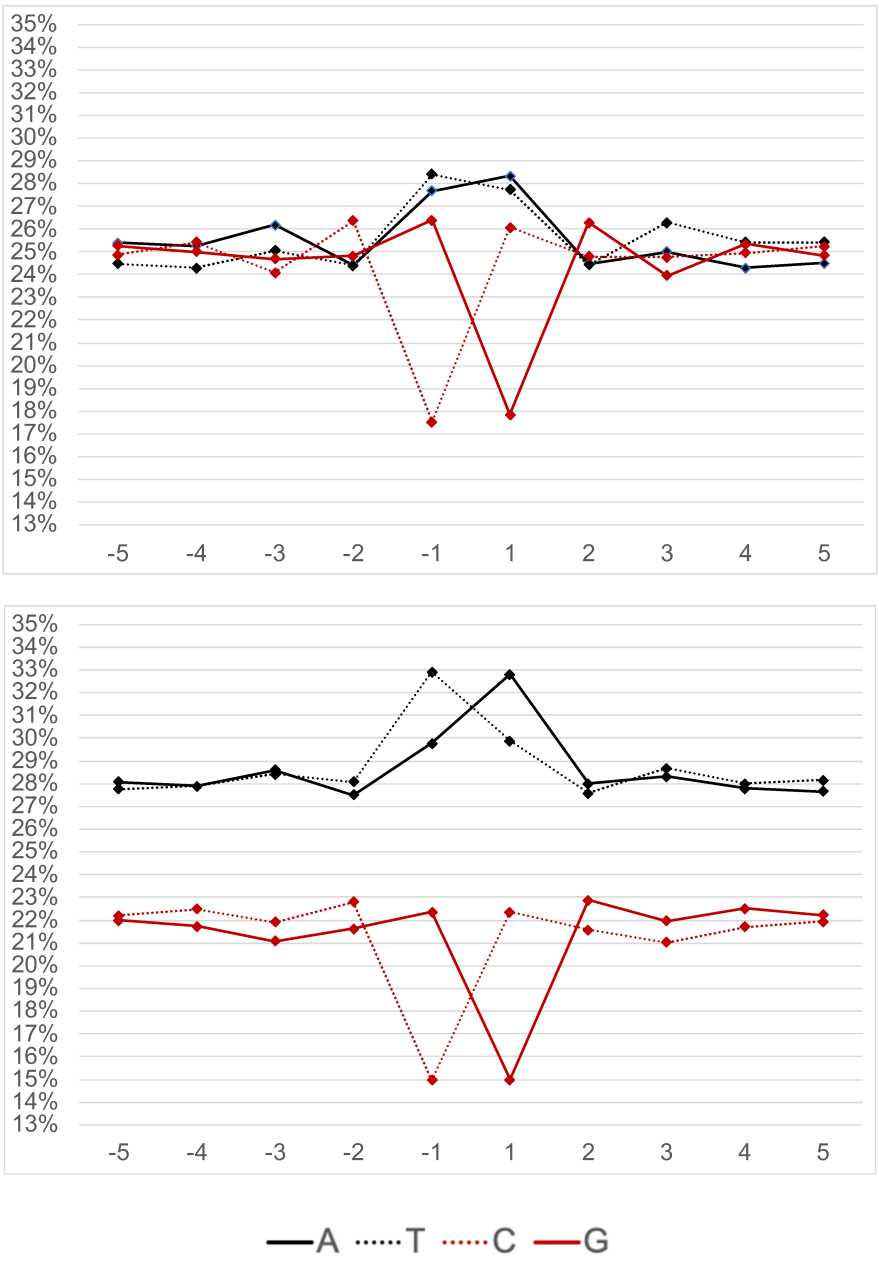


Figura Suplementar 85: Ocorrência de bases nas proximidades do SNP S (C/G) e a frequência basal esperada no genoma, nas regiões intrônicas de bovinos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

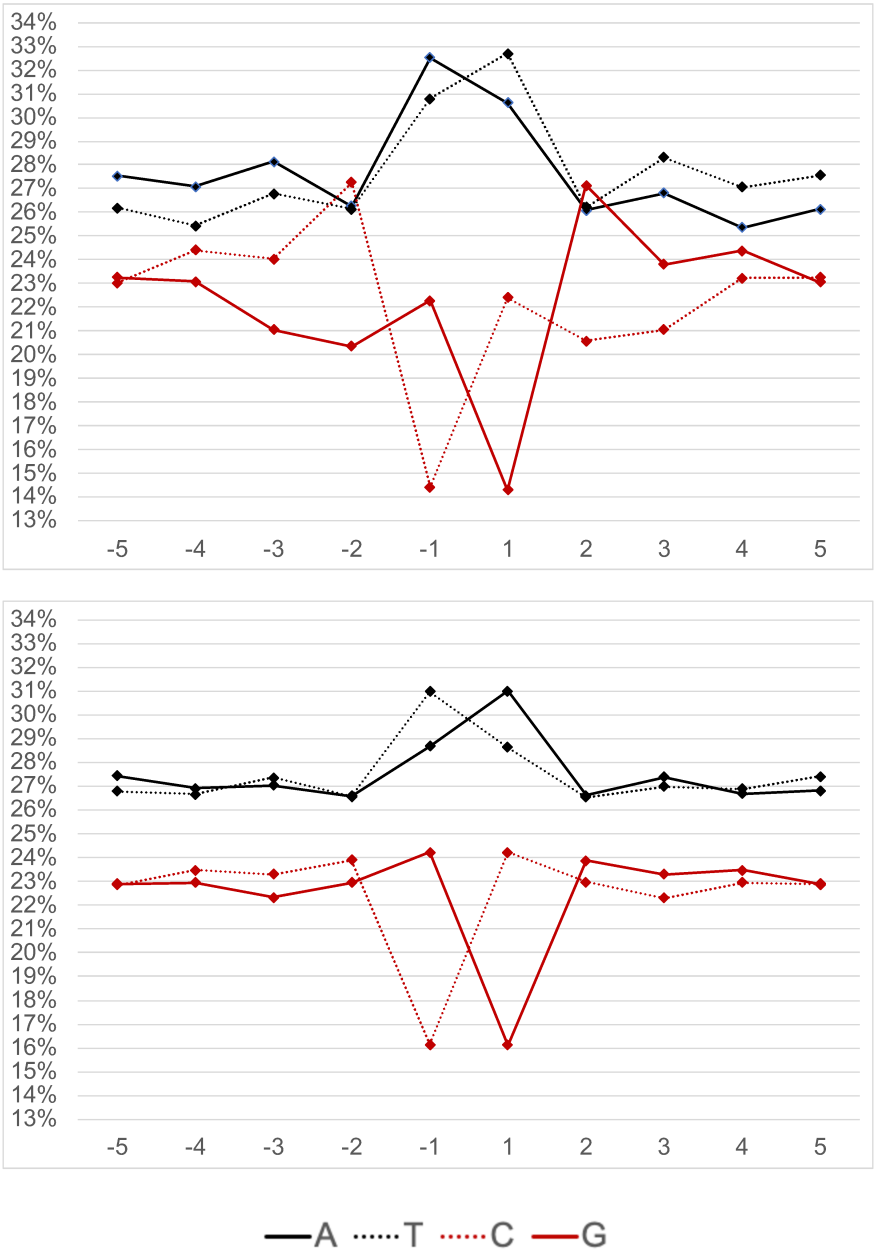


Figura Suplementar 86: Ocorrência de bases nas proximidades do SNP S (C/G) e a frequência basal esperada no genoma, nas regiões 3'UTR em humanos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

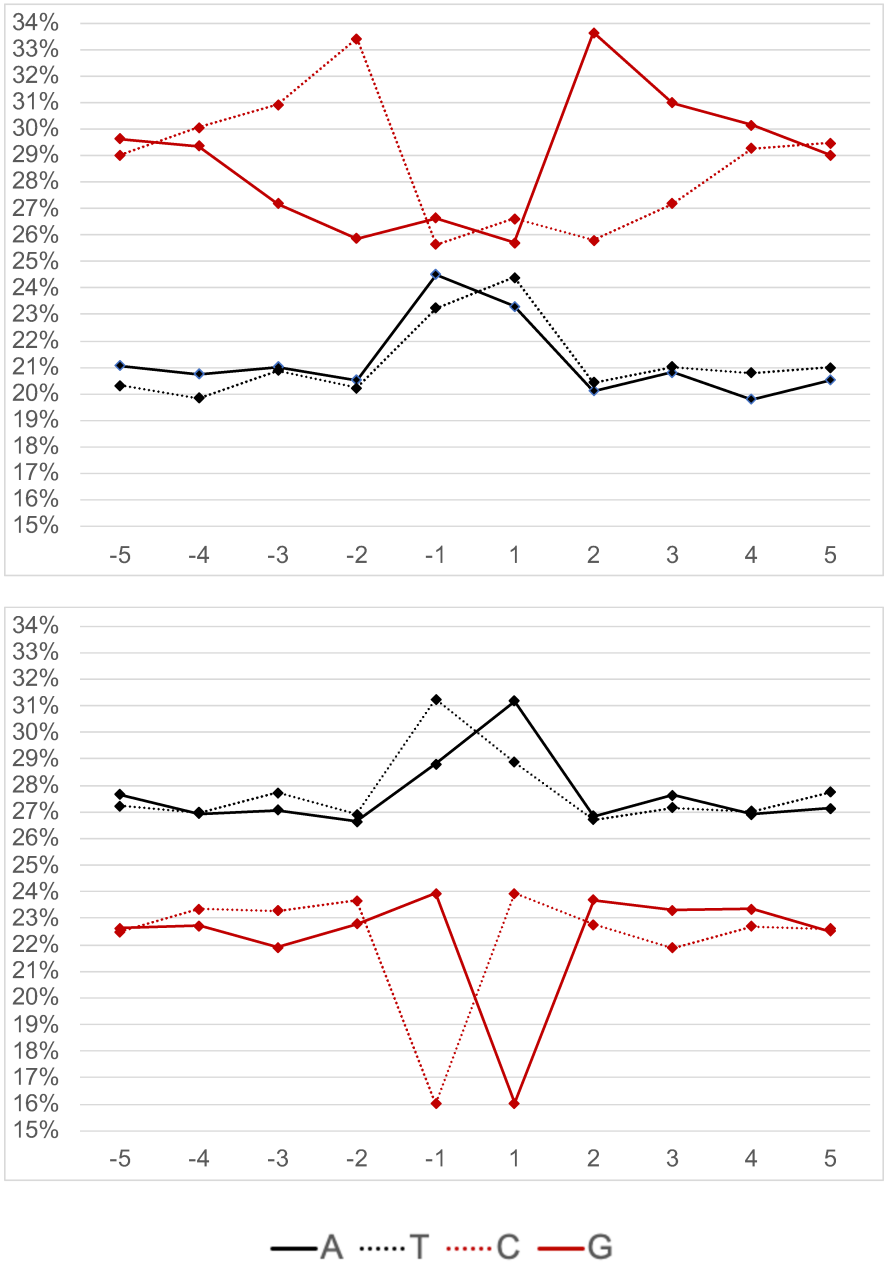


Figura Suplementar 87: Ocorrência de bases nas proximidades do SNP S (C/G) e a frequência basal esperada no genoma, nas regiões 5'UTR em humanos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

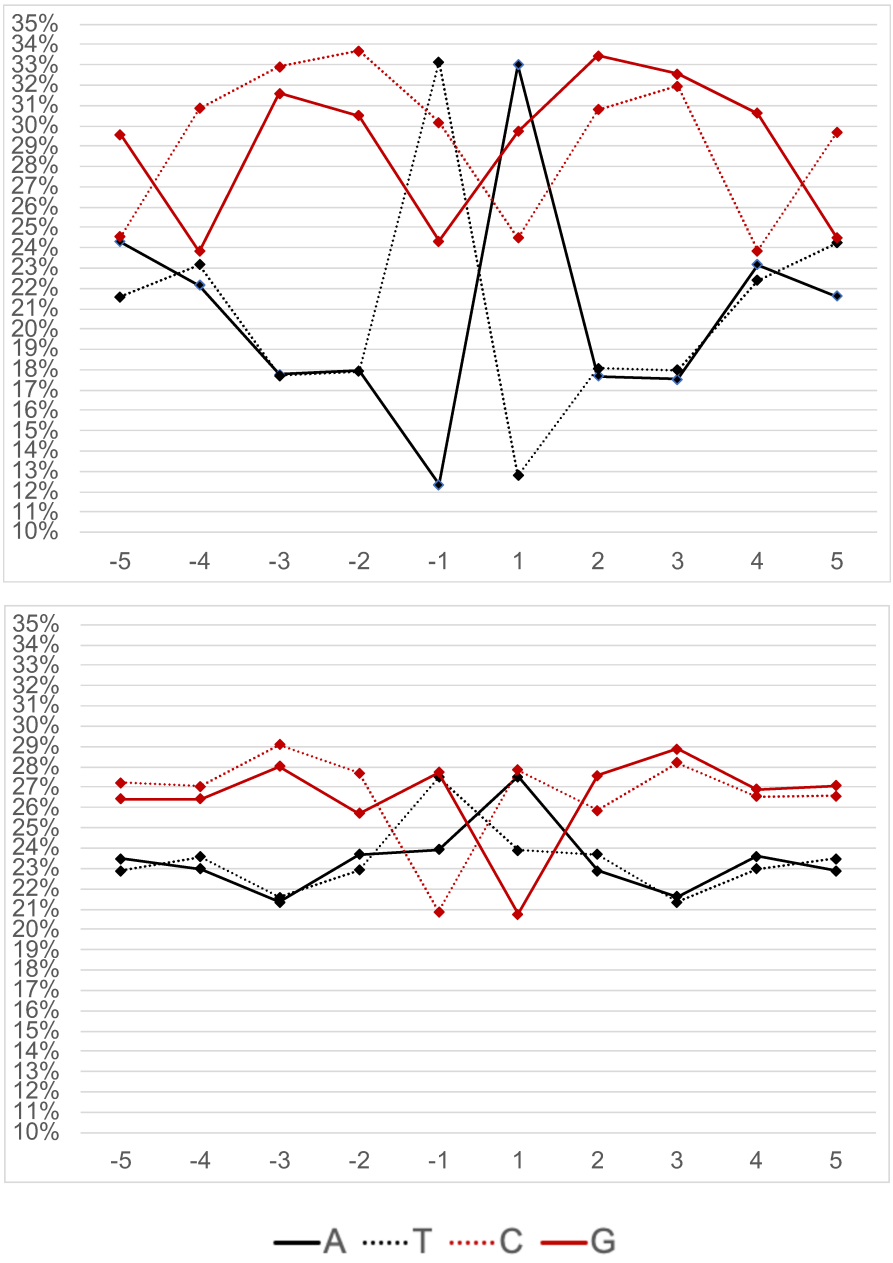


Figura Suplementar 88: Ocorrência de bases nas proximidades do SNP S (C/G) e a frequência basal esperada no genoma, nas regiões sinônimas em humanos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

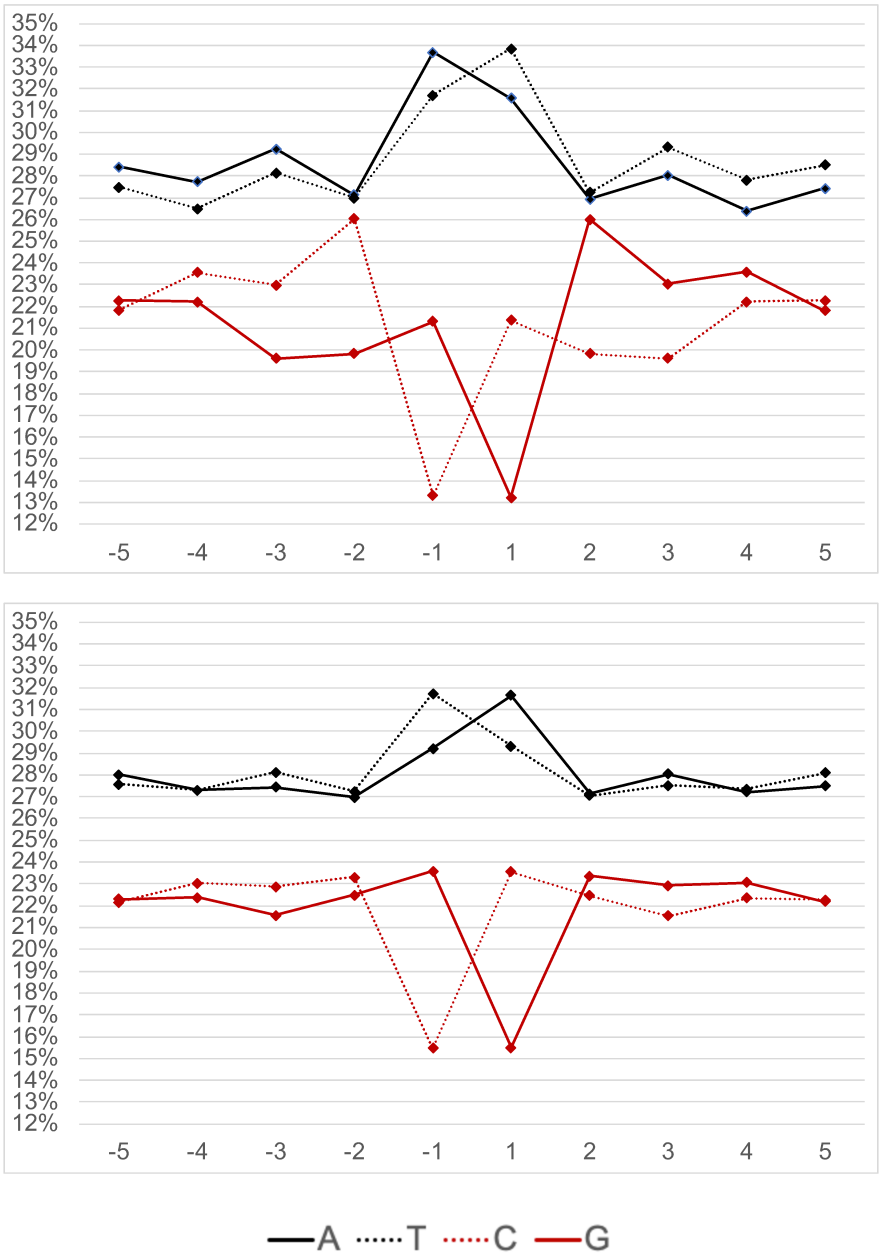


Figura Suplementar 89: Ocorrência de bases nas proximidades do SNP S (C/G) e a frequência basal esperada no genoma, nas regiões intrônicas em humanos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

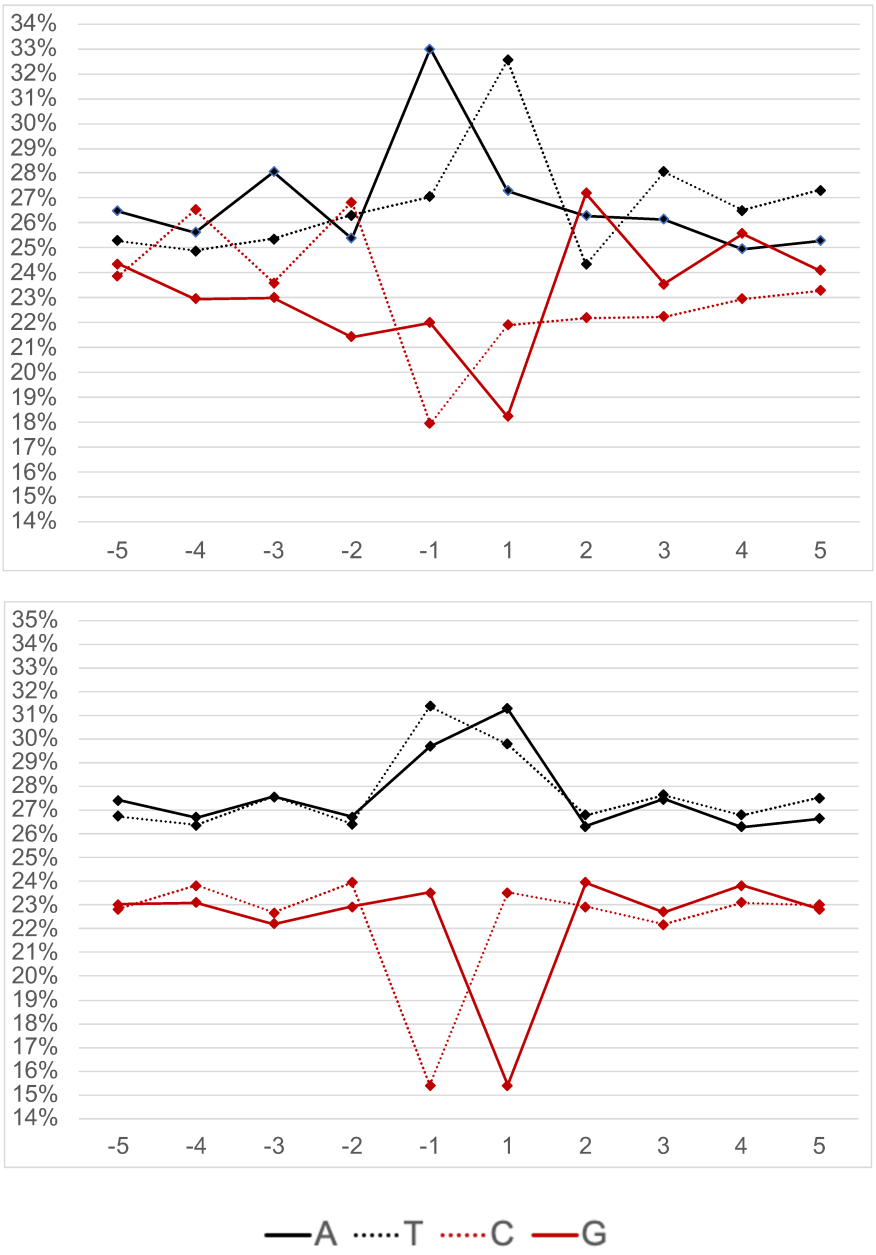


Figura Suplementar 90: Ocorrência de bases nas proximidades do SNP S (C/G) e a frequência basal esperada no genoma, nas regiões 3'UTR em camundongos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

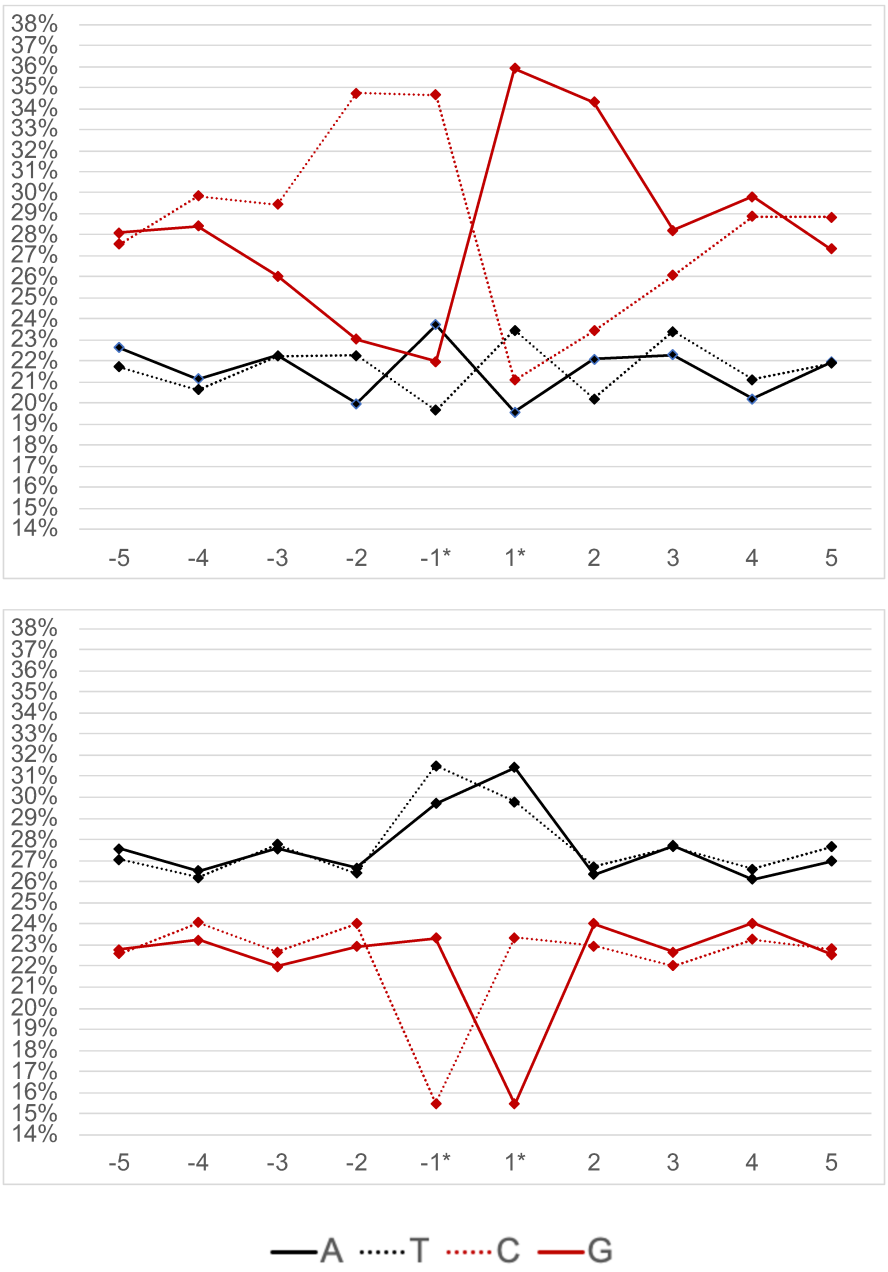


Figura Suplementar 91: Ocorrência de bases nas proximidades do SNP S (C/G) e a frequência basal esperada no genoma, nas regiões 5'UTR em camundongos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency^.

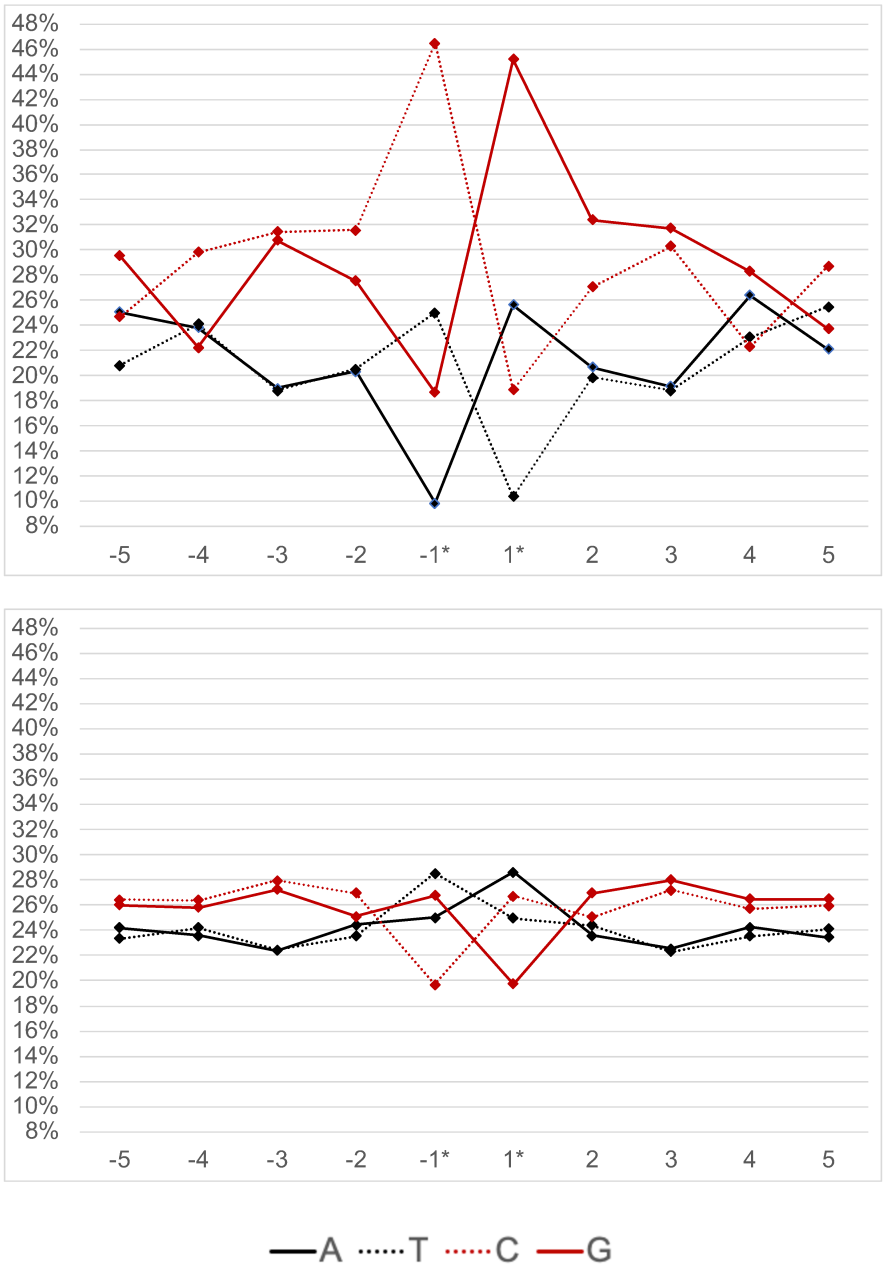


Figura Suplementar 92: Ocorrência de bases nas proximidades do SNP S (C/G) e a frequência basal esperada no genoma, nas regiões sinônimas em camundongos. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency^.

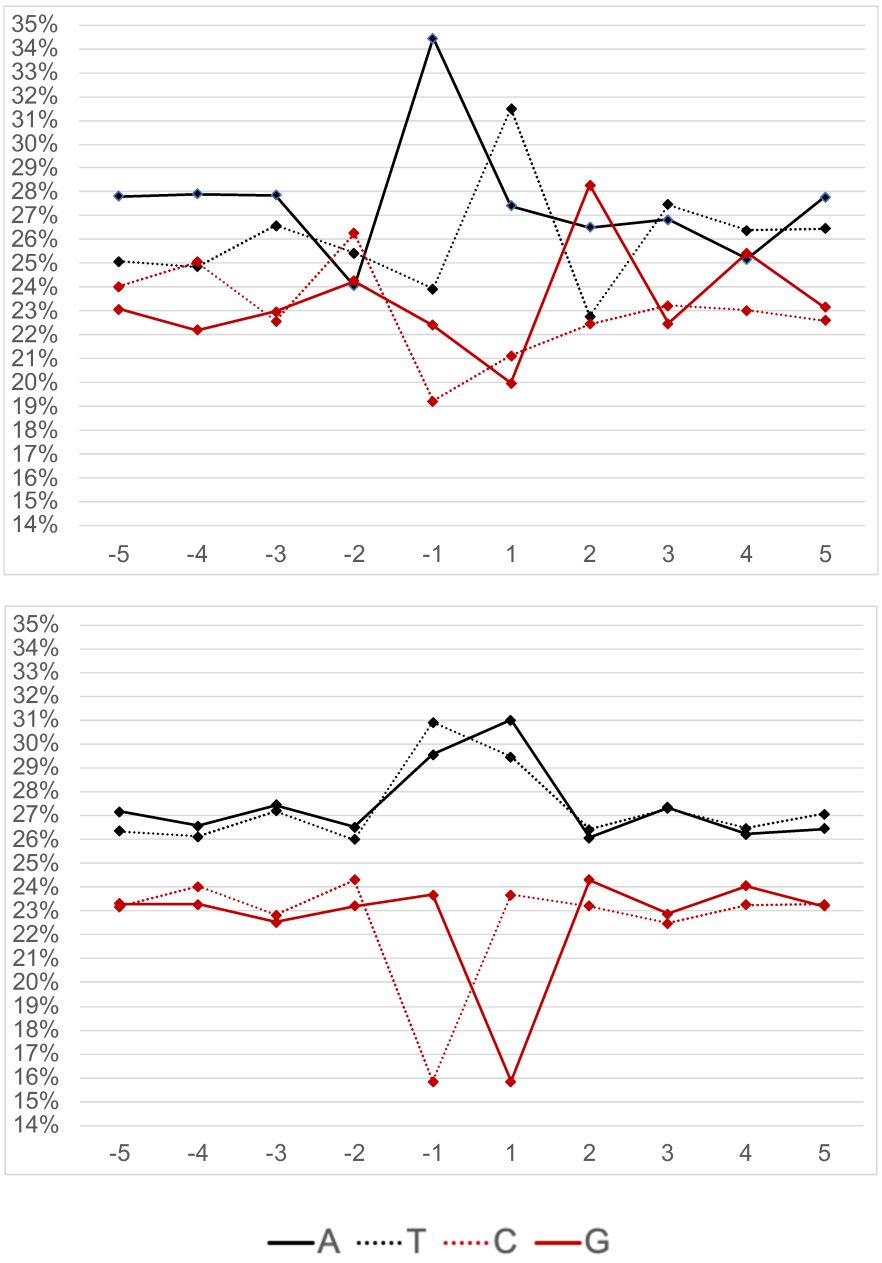


Figura Suplementar 93: Ocorrência de bases nas proximidades do SNP S (C/G) e a frequência basal esperada no genoma, nas regiões 3'UTR em *Rattus norvegicus*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

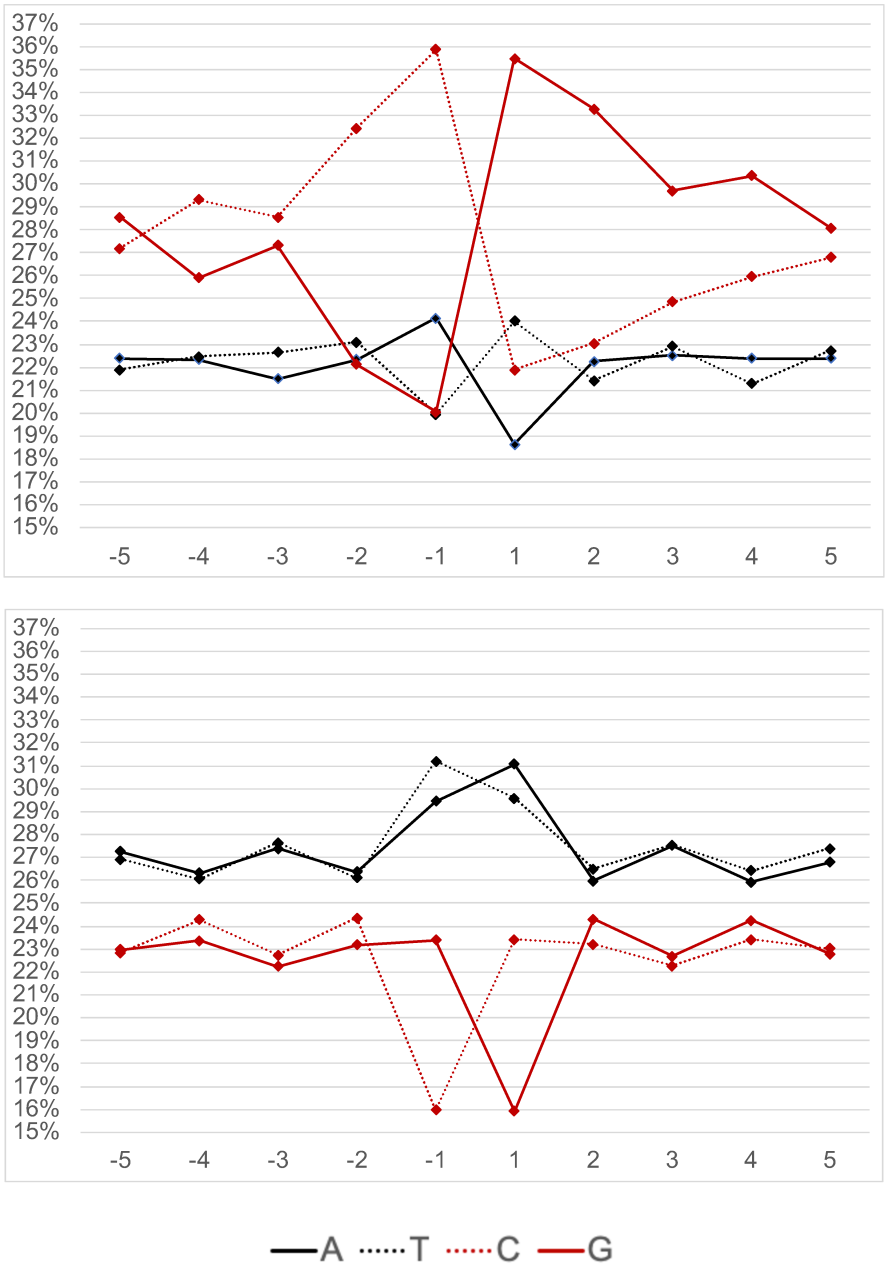


Figura Suplementar 94: Ocorrência de bases nas proximidades do SNP S (C/G) e a frequência basal esperada no genoma, nas regiões 5'UTR em *Rattus norvegicus*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

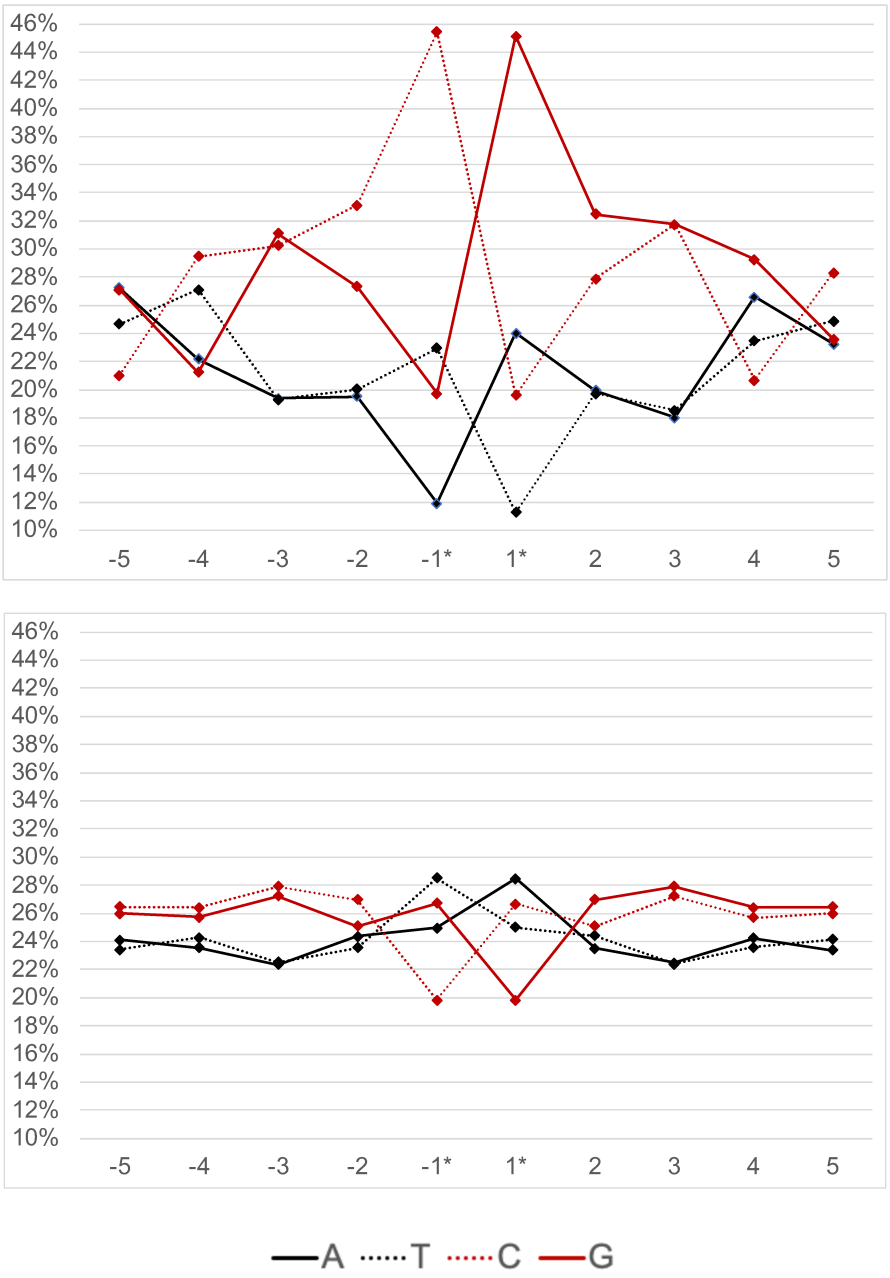


Figura Suplementar 95: Ocorrência de bases nas proximidades do SNP S (C/G) e a frequência basal esperada no genoma, nas regiões sinônimas em *Rattus norvegicus*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ˆbaseline frequencyˆ.

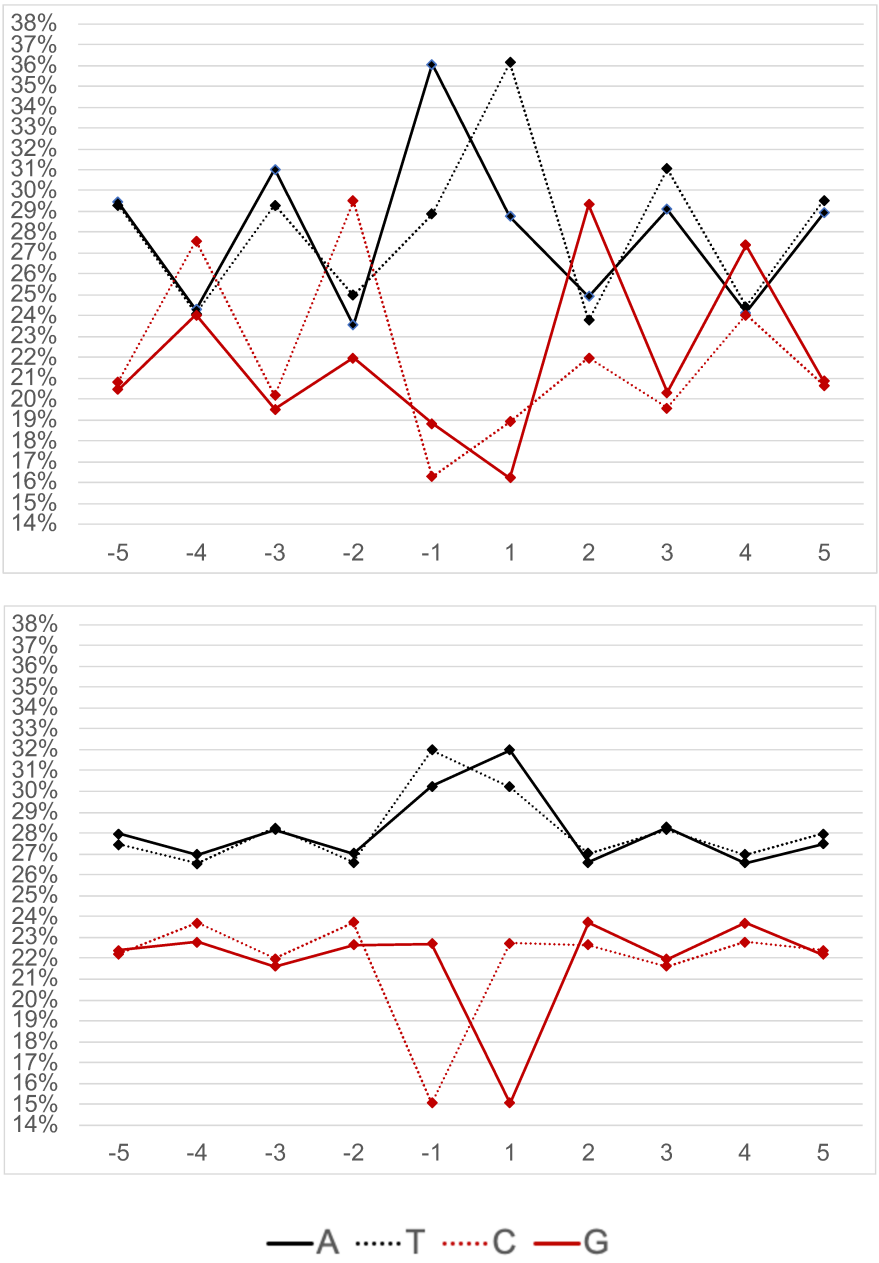


Figura Suplementar 96: Ocorrência de bases nas proximidades do SNP S (C/G) e a frequência basal esperada no genoma, nas regiões intrônicas em *Rattus norvegicus*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

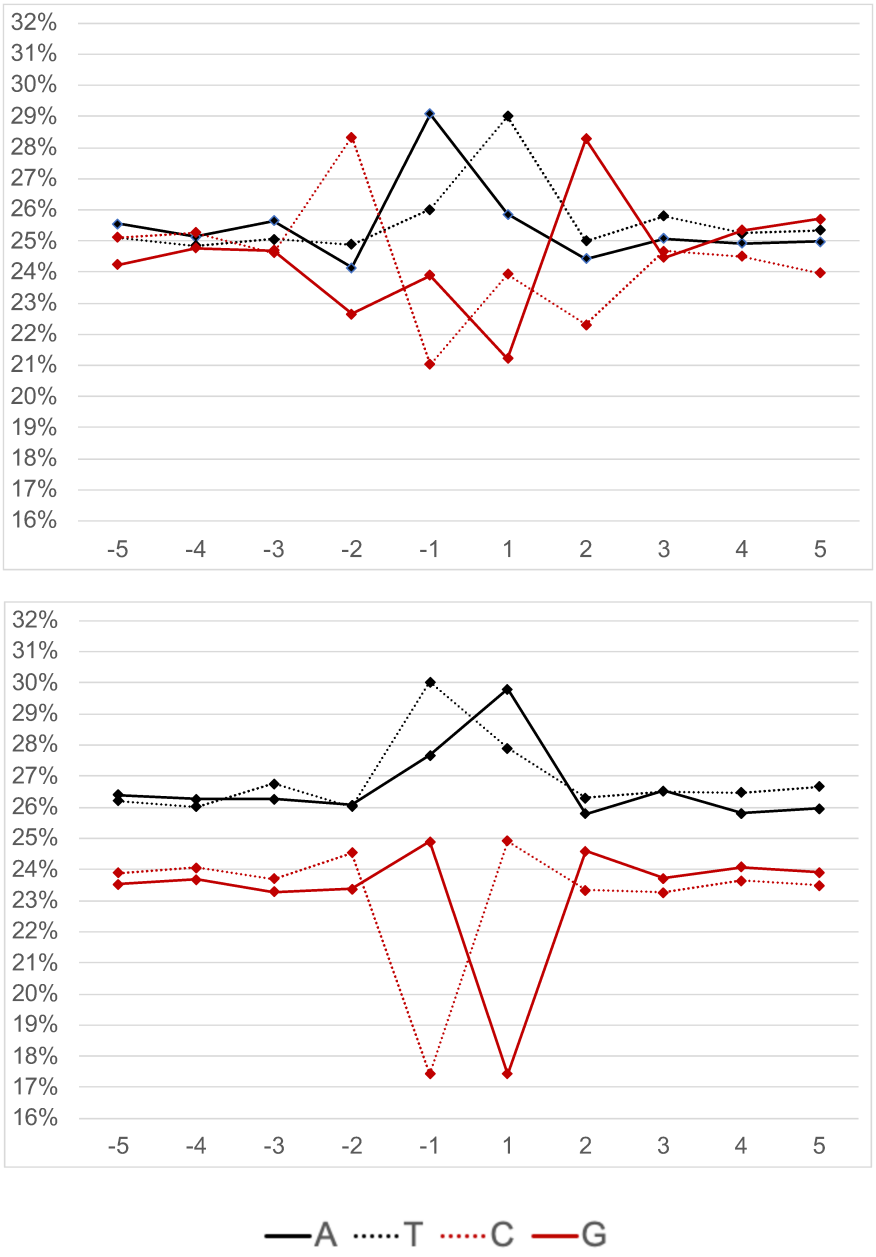


Figura Suplementar 97: Ocorrência de bases nas proximidades do SNP S (C/G) e a frequência basal esperada no genoma, nas regiões 3'UTR em *Sus scrofa*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

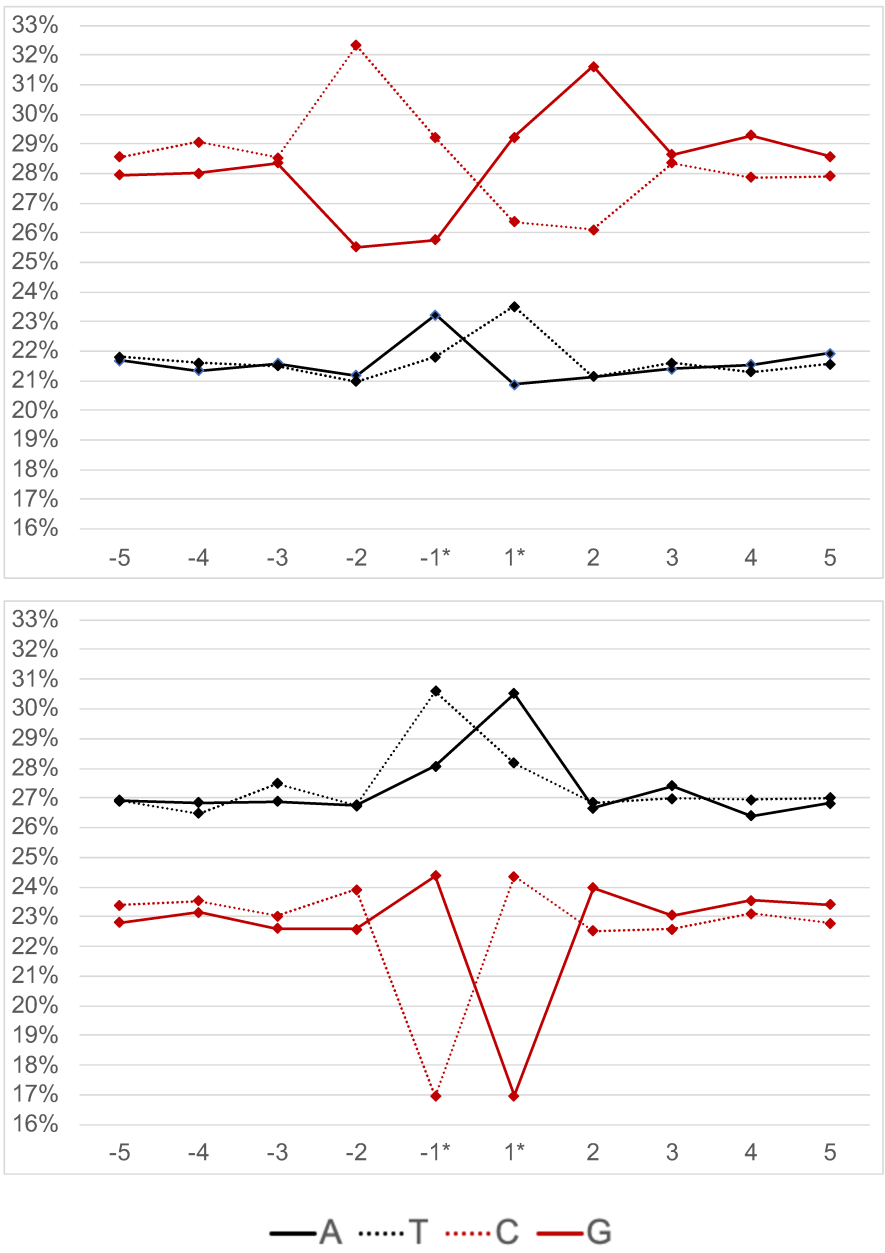


Figura Suplementar 98: Ocorrência de bases nas proximidades do SNP S (C/G) e a frequência basal esperada no genoma, nas regiões 5'UTR em *Sus scrofa*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição. Os asteriscos representam as posições significativamente distintas entre a região que contém o SNP e a ^baseline frequency^.

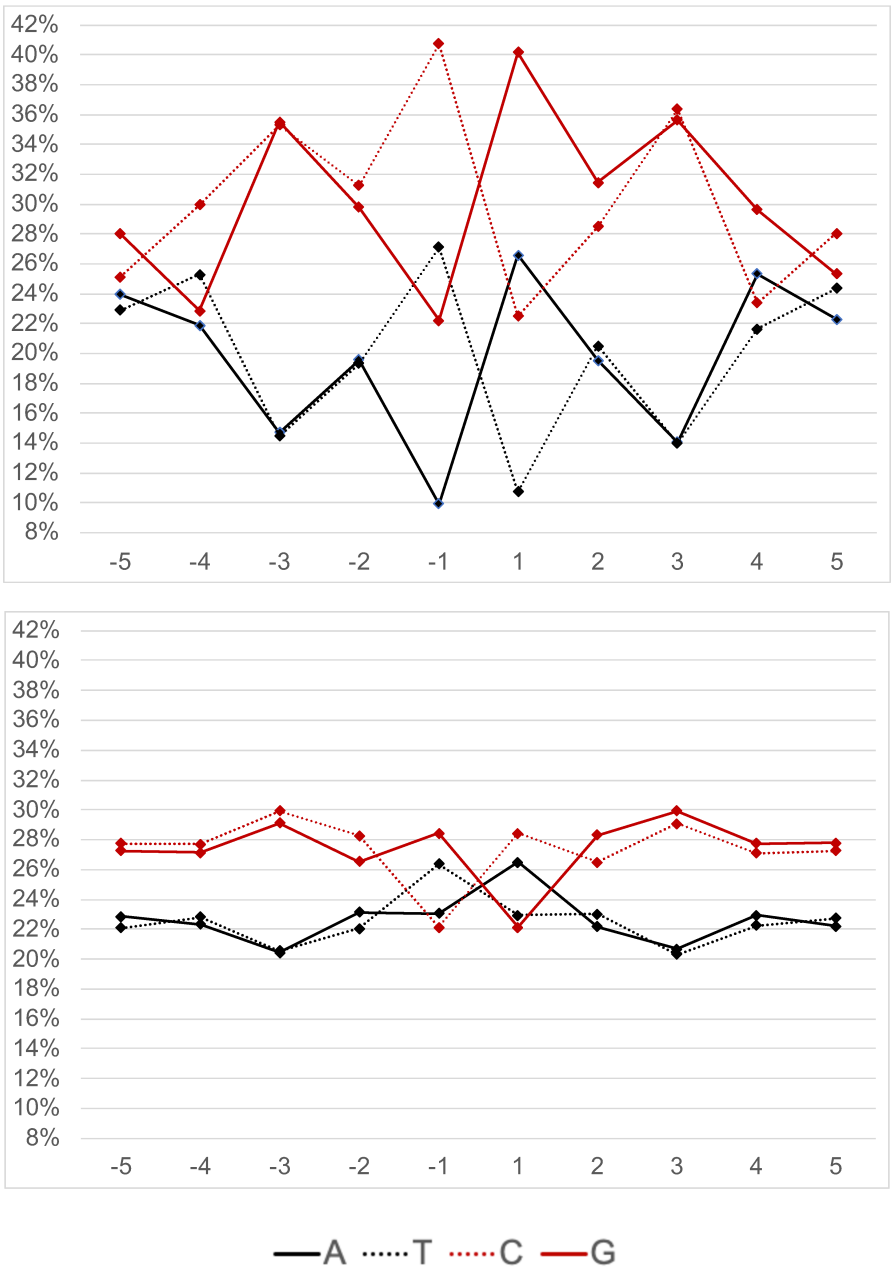


Figura Suplementar 99: Ocorrência de bases nas proximidades do SNP S (C/G) e a frequência basal esperada no genoma, nas regiões sinônimas em *Sus scrofa*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

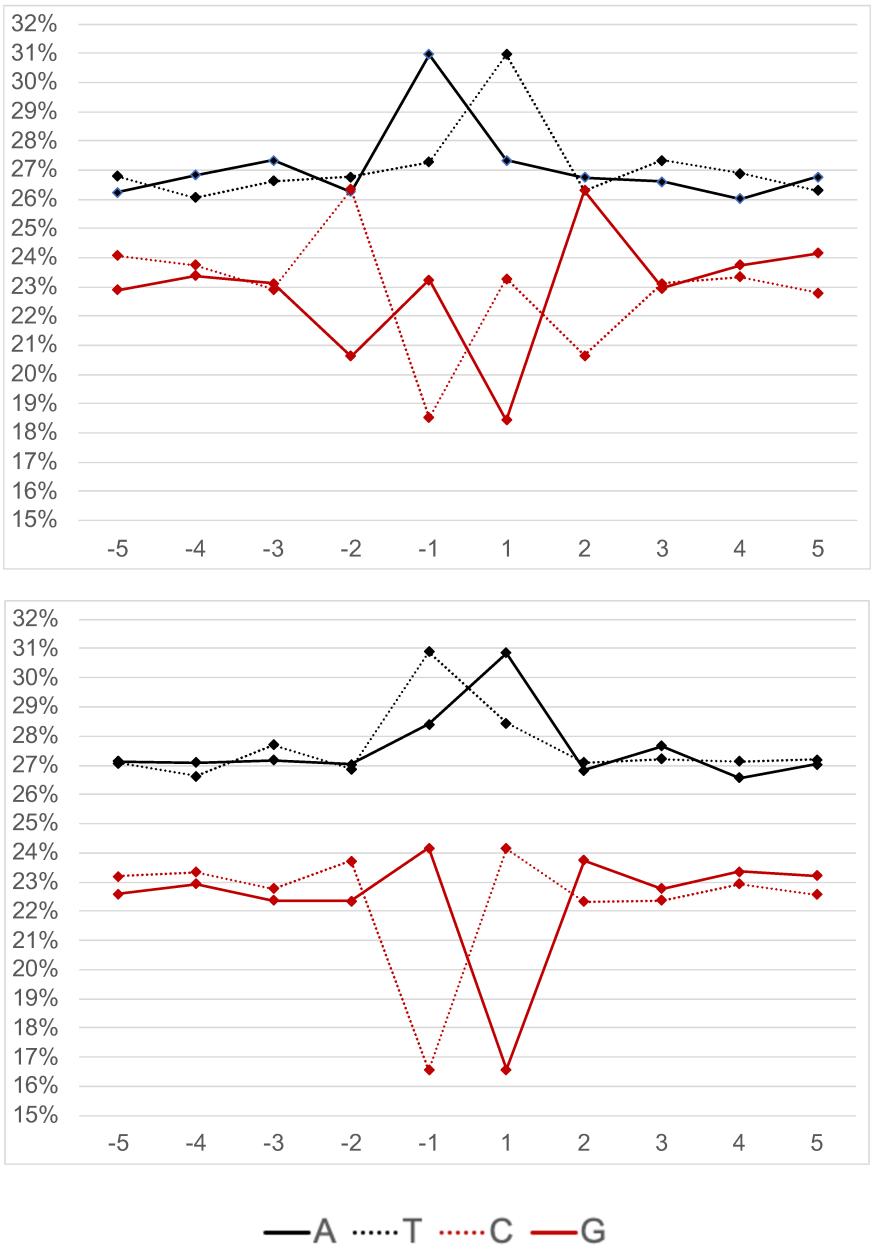


Figura Suplementar 100: Ocorrência de bases nas proximidades do SNP S (C/G) e a frequência basal esperada no genoma, nas regiões intrônicas em *Sus scrofa*. São mostrados resultados para os SNPs observados no gráfico superior e a frequência basal que se esperaria no genoma "baseline frequency" no gráfico inferior. As linhas para pirimidinas (T e C) são tracejadas e contínuas para purinas (A e G). O eixo X mostra as posições a 5' do SNP (-5 a -1) e as posições a 3' (1 a 5). O eixo Y indica as porcentagens de ocorrência de cada base naquela posição.

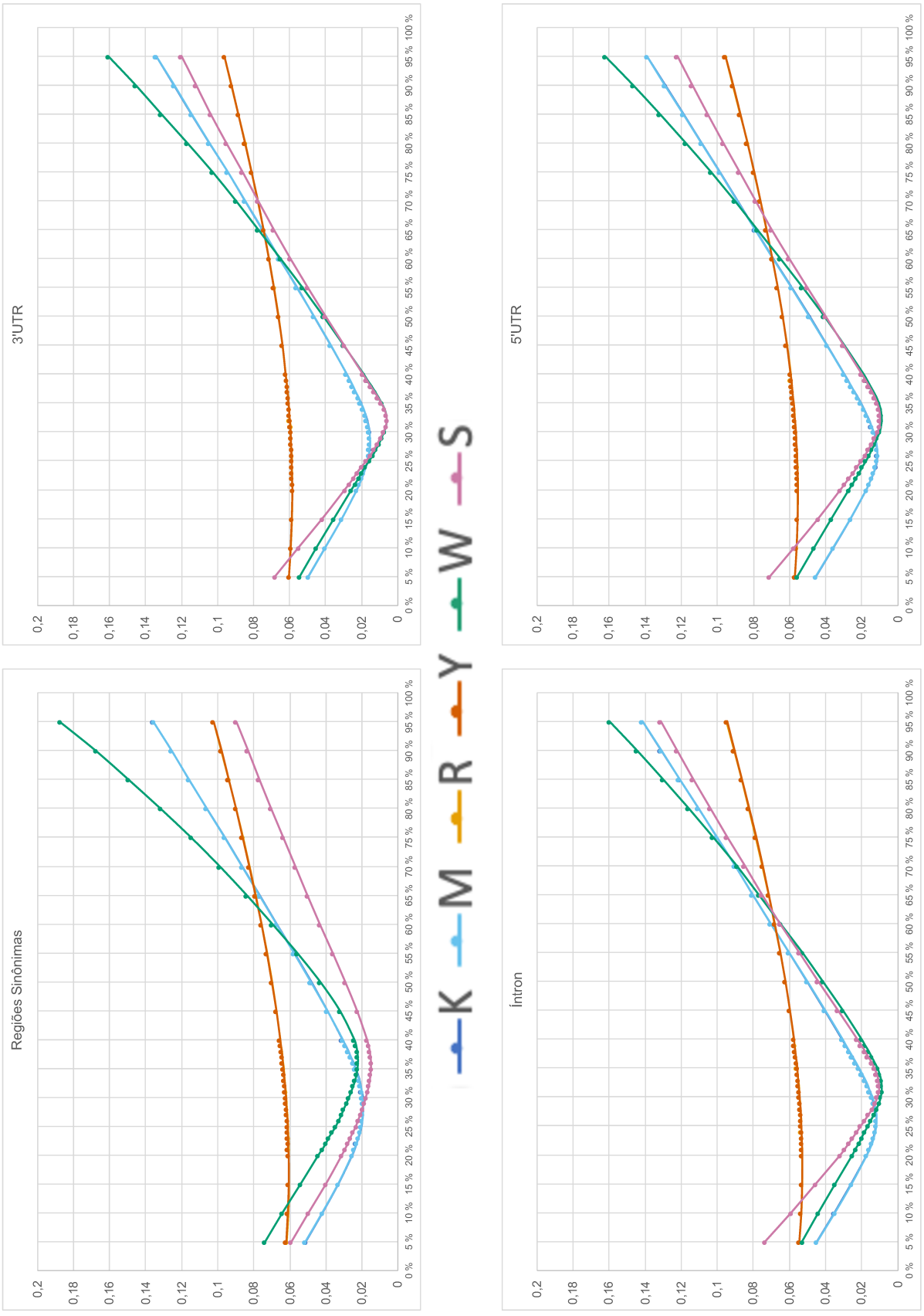


Figura Suplementar 101: Distância Euclidiana entre as médias das posições 5 e 1 em diferentes regiões em *Bos taurus*.
No eixo X temos os diferentes percentuais de aminação que da respectiva região do genoma e no eixo Y os valores da distância euclidiana entre a média das posições +1 e -1 e a média das posições +5 e -5.

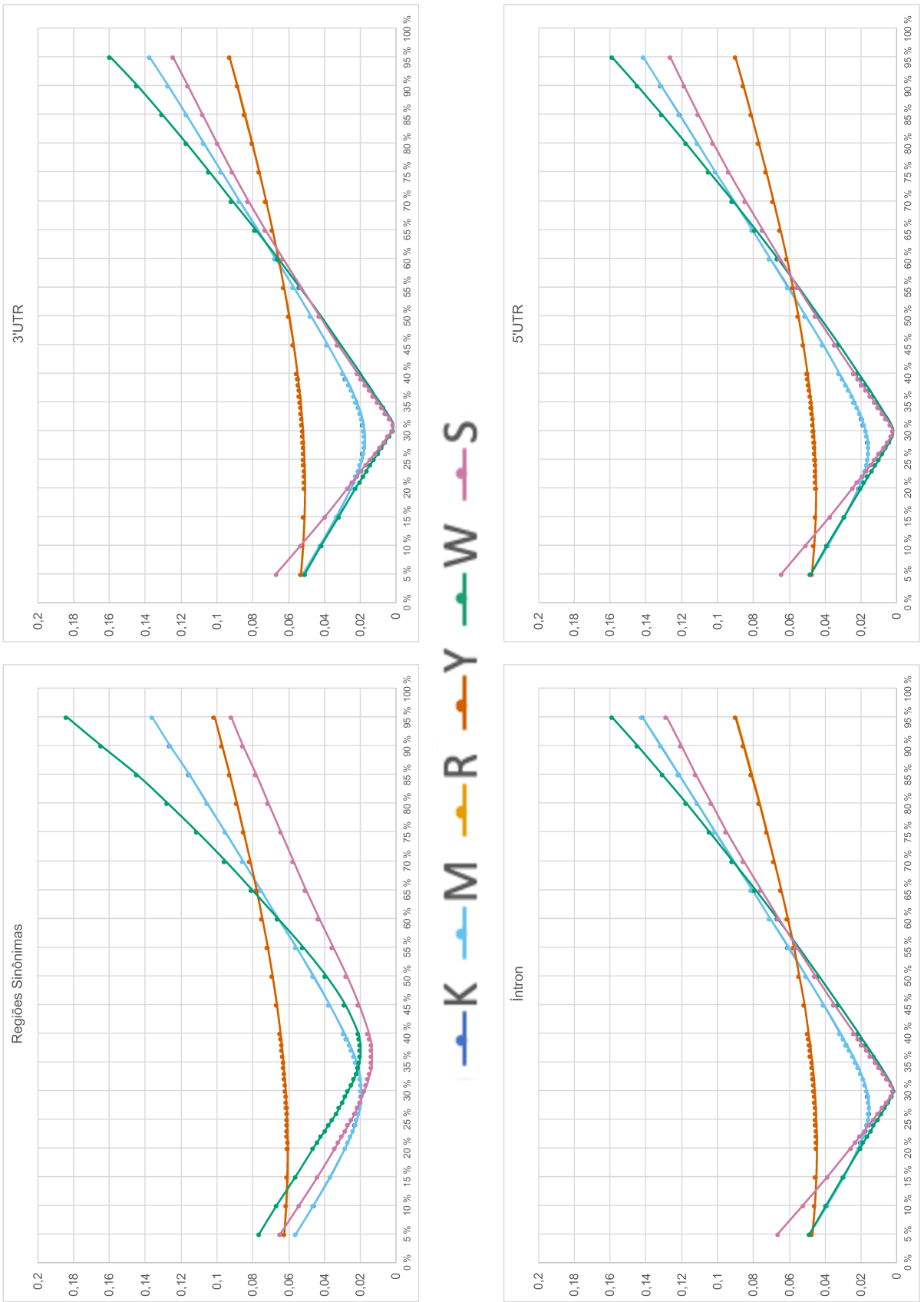


Figura Suplementar 102: Distância Euclidiana entre as médias das posições 5 e 1 em diferentes regiões em *Homo sapiens*.

No eixo X temos os diferentes percentuais de aminação que da respectiva região do genoma e no eixo Y os valores da distância euclidiana entre a média das posições +1 e -1 e a média das posições +5 e -5.

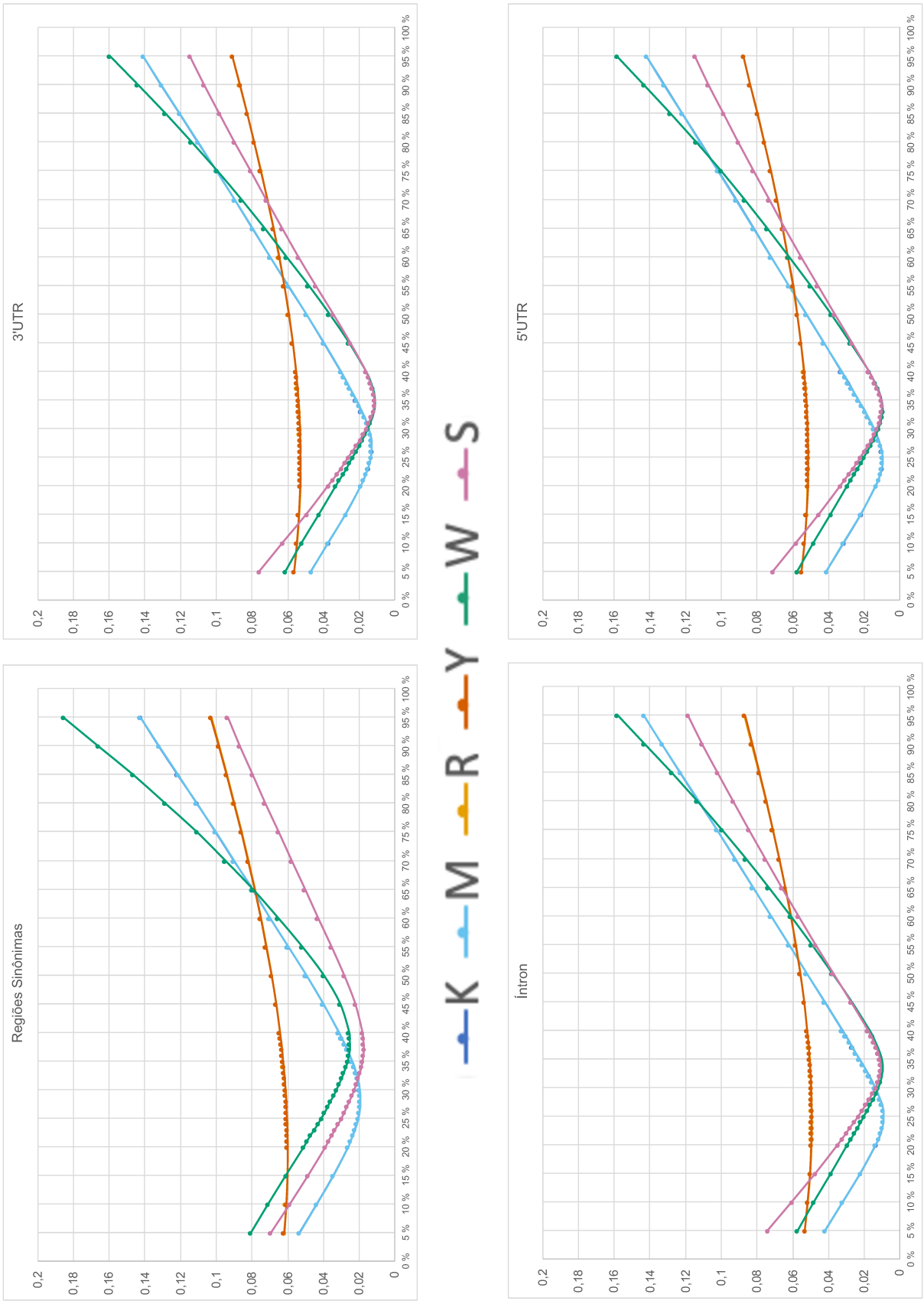


Figura Suplementar 103: Distância Euclidiana entre as médias das posições 5 e 1 em diferentes regiões em *Rattus norvegicus.*

No eixo X temos os diferentes percentuais de aminação que da respectiva região do genoma e no eixo Y os valores da distância euclidiana entre a média das posições +1 e -1 e a média das posições +5 e -5.

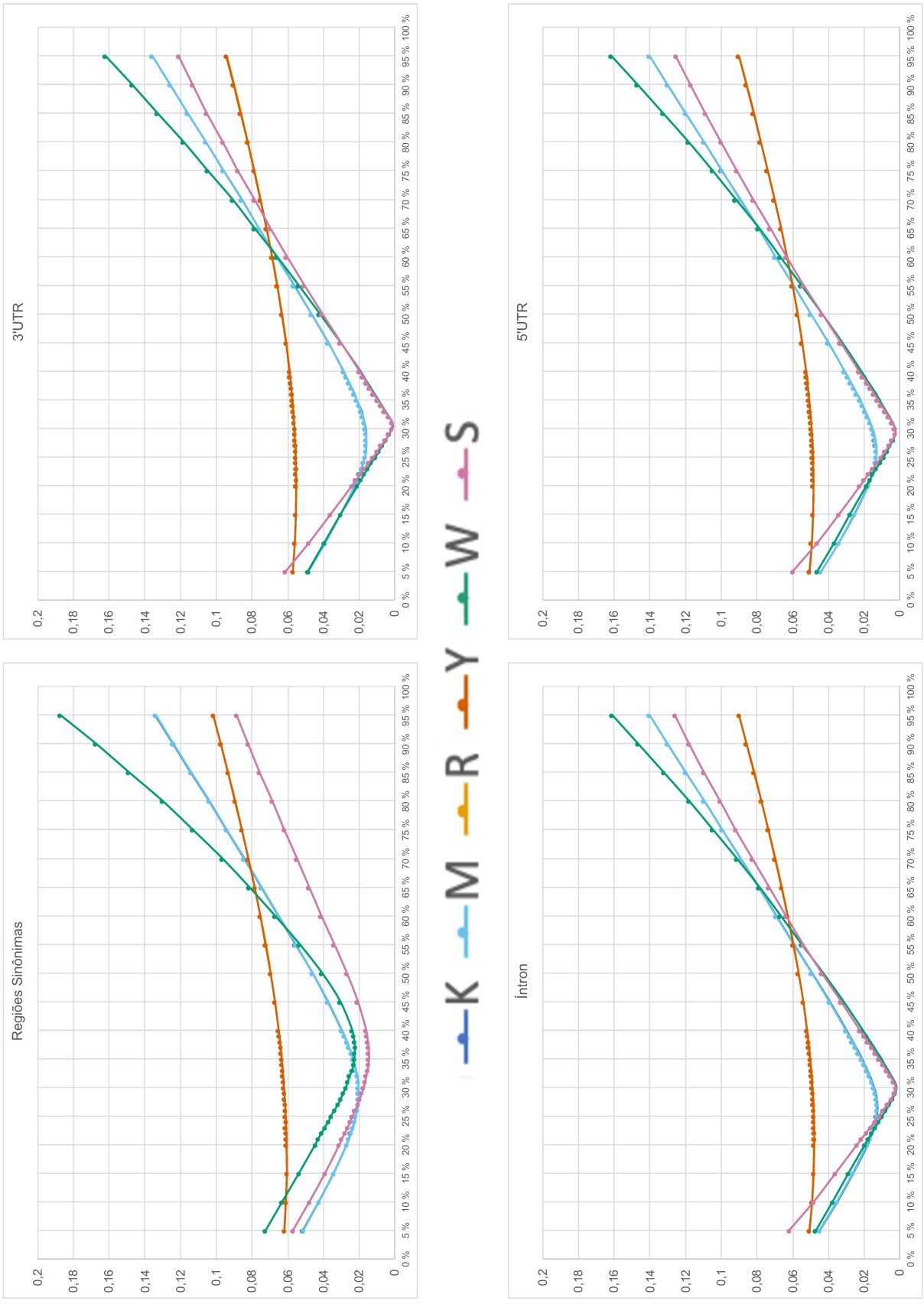


Figura Suplementar 104: Distância Euclidiana entre as médias das posições 5 e 1 em diferentes regiões em *Sus scrofa.*

No eixo X temos os diferentes percentuais de aminação que da respectiva região do genoma e no eixo Y os valores da distância euclidiana entre a média das posições +1 e -1 e a média das posições +5 e -5.