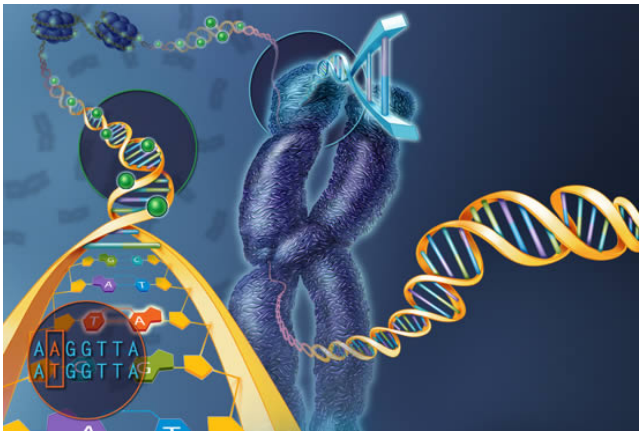
Universidade de Aveiro Departamento de Matemática,
2016

Vera Alexandra do Amparo R. Enes

Análise das variações genéticas: Caraterização do contexto em que ocorrem as variações de nucleótido único



Universidade de Aveiro Departamento de Matemática,
2016

Vera Enes

Análise das variações genéticas: Caraterização do contexto em que ocorrem as variações de nucleótido único

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requesitos necessários à obtenção do grau de Mestre em Matemática e Aplicações: Ramo de Estatística e Investigação Operacional, realizada sob a orientação científica da Professora Doutora Vera Afreixo, Professora do Departamento de Matemática da Universidade de Aveiro e do Professor Doutor João Rodrigues, Professor do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro

Júri

Presidente	**Professora Doutora Andreia Oliveira Hall** Professora Associada, Universidade de Aveiro
Vogal - Arguente Principal	**Professor Doutor Miguel Francisco Almeida Pereira Rocha** Professor Associado, Universidade do Minho
Vogal - Orientador	**Professora Doutora Vera Mónica Almeida Afreixo** Professora Auxiliar, Universidade de Aveiro

Agradecimentos

Aos meus orientadores, Professora Doutora Vera Mónica Almeida Afreixo e Professor Doutor João Rodrigues por toda a ajuda, compreensão, disponibilidade e incentivo para conseguir levar a bom porto esta dissertação.

Ao Antero e à minha família, pelo apoio e carinho oferecidos nos momentos mais difíceis. A eles dedico este trabalho.

Palavras-chave

Variação genética, SNV, Projeto 1000 Genomas, Contexto na vizinhança, HWE

Resumo

A identificação e caracterização das variações genéticas podem ajudar os médicos no diagnóstico e prevenção de doenças com origem genética. Com a sequenciação completa do genoma de referência foi possível, posteriormente, encontrar um grande número de variações genéticas entre a população, das quais as mais comuns são a variação de um único nucleótido (SNV). No entanto, para ser possível capturar a diversidade humana, torna-se necessário sequenciar o genoma de muitas pessoas. O projeto 1000 Genomas surge com objetivo de caracterizar a variação genética humana e construir o maior catálogo público de dados sobre variações, com especial foco nas variações raras. Neste trabalho, pretende-se caracterizar os diferentes tipos de SNVs, recorrendo aos dados disponibilizados na primeira fase do projeto do 1000 Genomas. Esses dados contêm 38 milhões de SNVs identificadas no genoma de 1092 indivíduos, provenientes de 14 populações. Tem-se como objetivo, avaliar a forma como ocorrem os diferentes tipos de SNVs ao longo do genoma, fazendo uma análise por cromossoma ou de acordo com a prevalência, assim como, verificar se os genótipos associados a cada SNV satisfazem o equilíbrio de Hardy-Weinberg (HWE). Para além disso, acreditando que as variações genéticas não ocorrem por acaso e que numa vizinhança da posição onde ocorre a SNV existirá informação que seja indicadora do fenómeno, um dos objetivos principais deste trabalho é caracterizar o contexto na vizinhança de cada SNV, através da contagem de palavras de diferentes tamanhos. Observou-se que os diferentes tipos de SNVs ocorrem de forma homogénea ao longo do genoma, mas que as frequências de palavras na vizinhança do local de ocorrência da SNV, estão associadas a cada tipo de variação, principalmente nas posições imediatamente adjacentes a esse local. Verificou-se ainda que para a maioria das SNVs, os respetivos genótipos estão de acordo com o HWE.

Keywords Genetic variation, SNV, 1000 Genomes Project, Context in the neighborhood, HWE

Abstract The identification and characterization of genetic variations can help physicians in the diagnosis and prevention of diseases with genetic origin. With the complete sequencing of the reference genome, was possible to find a large number of genetic variations among the population, of which the most common are single nucleotide variation (SNV). However, to be able to capture human diversity, it becomes necessary to sequence the genome of many people. The 1000 Genomes project is designed to characterize human genetic variation and build the largest public catalog of variation data, with a special focus on rare variations. In this work, we intend to characterize the different types of SNVs, using the data available in the first phase of the 1000 Genomes project. These data contain 38 million SNVs identified in the genome of 1092 individuals from 14 populations. The objective of this study was to evaluate the way different types of SNVs occur along the genome, by a chromosome analysis or according to the prevalence, as well as to verify if the genotypes associated to each SNV satisfy the Hardy- Weinberg equilibrium (HWE). In addition, believing that genetic variations do not occur by chance and that in a neighborhood of the position where the SNV occurs there will be information that is indicative of the phenomenon, one of the main objectives of this work is to characterize the context in the vicinity of each SNV, by counting words of different sizes. It was observed that the different types of SNVs occur homogeneously throughout the genome, but that the word frequencies in the vicinity of the SNV occurrence site are associated with each type of variation, especially at the positions immediately adjacent to that site. It was also verified that for most SNVs, the respective genotypes are in agreement with the HWE.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

P1000G Projeto 1000 Genomas

A Adenina

ADN Ácido Desoxirribonucleico

C Citosina

CRAN *Comprehensive R Archive Network*

FDR *False Discovery Rate*

FWER *Family-Wise Error Rate*

FWEC *Family-Wise Error under the Complete null*

G Guanina

GWAS *Genomic Wide Association Studies*

HWE *Hardy-Weinberg Equilibrium*

K-S Kolmogorov-Smirnov

MAF *Minor Allele Frequency*

SNV *Single Nucleotide Variation*

SNP *Single Nucleotide Polymorphism*

SV *Structural Variant*

T Timina

VCF *Variant Call Format*

Capítulo 1

Introdução

Devido ao potencial da genómica na melhoria da saúde humana, esta tem sido uma das áreas mais estudadas na pesquisa biomédica principalmente desde que foi publicada a sequência de referência do genoma humano [NHGRI et al., 2011]. Assim, considerando uma sequência de referência é possível identificar, numa determinada população, variações em relação a essa sequência, numa determinada posição da cadeia de ADN. O tipo mais comum de variação genética, que está presente em todo o genoma humano, é a variação de nucleótido único [Stram, 2014].

O local onde ocorre a variação pode implicar alterações fenotípicas ou alterar a estrutura e a função de certas proteínas levando ao desenvolvimento de doenças [Kim and Misra, 2007]. A compreensão da variação genética humana contribuirá para o melhoramento do que atualmente se chama de medicina personalizada, como por exemplo, uma agregação da informação genética, fenotípica e de factores ambientais poderá melhorar a previsão da resposta individual às terapêuticas [Squassina et al., 2010].

Para capturar a diversidade humana é necessário sequenciar o genoma de muitas pessoas. Deste modo, surgiu o projeto 1000 Genomas, que teve como objetivo caracterizar a variação genética humana, sequenciando numa primeira fase o genoma completo de 1092 indivíduos provenientes de 14 populações, construindo uma base de dados que possa ajudar a entender o contributo da genética para as diversas doenças [1000Genomes, 2012].

Acredita-se que as variações genéticas não ocorrem por acaso e que, numa vizinhança da posição onde ocorre a variação, existirá informação que seja indicadora do fenómeno. No trabalho de [Zhang and Zhao, 2004], os autores analisaram 433 192 variações de nucleótido único, provenientes do genoma de ratos, com o propósito de estudar a influência dos nucleótidos presentes na vizinhança dos locais de variação considerados. Observaram que, comparativamente aos valores globais no genoma, existia um viés significativo no número de nucleótidos presentes nas posições imediatamente adjacentes ao local onde houve a substituição e que esse mesmo viés ia diminuindo à medida que eram consideradas posições mais afastadas. Para além disso, verificaram que um certo tipo de variação, as transições, era influenciado pelos nucleótidos C e G nas posições próximas do local onde ocorreu a variação e que os nucleótidos A e T surgiam mais frequentemente noutro tipo de variação, as transversões. Em [Jiang et al., 2008], considerando o genoma bovino, foram examinados padrões de frequência de nucleótidos na vizinhança de cerca de 15 mil variações de nucleótido único tendo observado que certos tipos de variações apresentavam padrões de frequências complementares. Verificaram também que ao considerar todas as combinações possíveis de nucleótidos, nas duas posições imediata-

mente adjacentes ao local de variação, nas combinações em que era possível formar estruturas CpG, existia um aumento significativo da taxa de ocorrência de certos tipos de variações. Na análise realizada no trabalho [Plyler et al., 2015], consideraram-se quartetos de dinucleótidos, definidos como os dois pares de nucleótidos a montante e a jusante do local de variação. Os autores pretenderam estudar a influência da vizinhança na formação de variações em exões, intrões e genes de ratinhos, assim como, em genes humanos provenientes da primeira fase do projeto 1000 Genomas e mutações específicas para o cancro da mama humano. Em [Voight and Aggarwala, 2016], os autores partem da hipótese de que o contexto na vizinhança do local de variação, isto é, os nucleótidos presentes nas sequências à esquerda e à direita desse local, podem explicar a variabilidade observada nas probabilidades de substituição de nucleótidos, numa dada região do genoma. Para testar essa hipótese, foram definidos e comparados vários modelos estatísticos baseados na estimação das probabilidades de substituição de nucleótidos, de acordo com o contexto considerado. Esses contextos englobaram o caso trinucleótido (quando se considera a posição de variação e os dois nucleótidos imediatamente adjacentes a essa posição), pentanucleótido (o local de variação e os dinucleótidos imediatamente adjacentes) e por fim heptanucleótido (considerando os trinucleótidos mais próximos da posição de variação incluindo a mesma). Para realizar a análise, os autores recorreram ao dados disponibilizados na fase 1 do projeto 1000 Genomas, tendo concluído que é o contexto heptanucleótido que melhor explica a variabilidade nas probabilidades de substituição.

Com este trabalho, em semelhança ao que foi abordado nos trabalhos citados anteriormente, pretende-se estudar se o contexto na vizinhança do local da variação é indicador ou não da ocorrência dessa variação, principalmente nas posições imediatamente adjacentes a esse local. Considerando vizinhanças de diversas amplitudes, vão analisar-se padrões de frequência que envolvem a contagem de nucleótidos, dinucleótidos e trinucleótidos tendo com o objetivo, caracterizar o contexto em que ocorrem variações genéticas. Espera-se confirmar, no estudo dos trinucleótidos, que estes são os que têm maior influência na ocorrência da variação. Pretende-se realizar uma análise mais global, uma vez que se consideram variações de nuclótido único, das cerca de 38 milhões disponibilizadas na fase 1 do projeto 1000 Genomas, não restringindo o estudo apenas às variações identificadas em genes. Para além disso, pretende-se caracterizar as variações estratificando-as de acordo com a sua prevalência e verificar se as mesmas, satisfazem um princípio fundamental da genética moderna, que é o equilíbrio de Hardy-Weinberg.

Neste capítulo apresenta-se uma introdução sobre alguns conceitos de genómica e sobre o projeto 1000 Genomas. Uma vez que para realizar a análise das variações de nucleótido único, foi necessário desenvolver várias ferramentas estatísticas computacionais, apresenta-se também uma breve abordagem ao *software* R e *packages* do Bioconductor. Finaliza-se o capítulo com a descrição dos dados provenientes da fase 1 do projeto 1000 Genomas, seguindo-se os objetivos gerais deste trabalho e a estrutura da dissertação.

1.1 Um pouco de genómica

Na biologia molecular, o ácido desoxirribonucleico (ADN) é considerado como a base fundamental da hereditariedade de qualquer organismo. Nas células humanas, o ADN está organizado em pares de cromossomas homólogos, os quais se encontram no núcleo da célula. Trata-se de uma molécula que é constituída por duas cadeias de subunidades chamadas de nucleótidos [Stram, 2014]. A figura 1.1 representa uma célula, destacando um cromossoma e respetiva molécula de ADN. Cada nucleótido possui uma base, um açúcar de cinco atómos de carbono (desoxirribose) e um grupo fosfato. As bases podem ser classificadas como

- Purinas: Adenina (A) e Guanina (G);
- Pirimidinas: Citosina (C) e Timina (T).

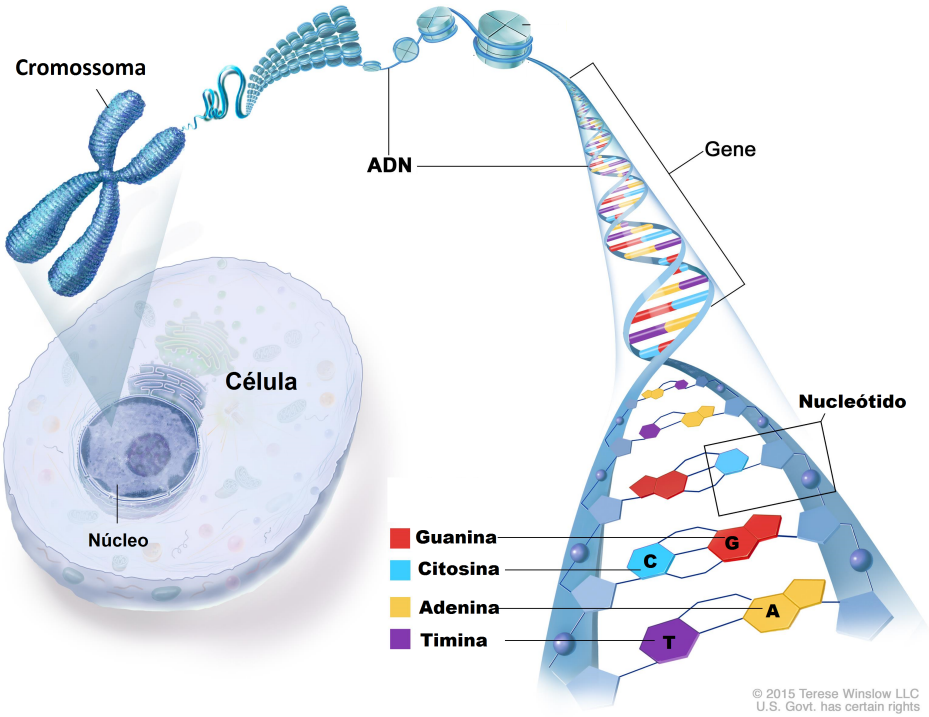


Figura 1.1: Representação da estrutura do ADN, como parte integrante dos cromossomas presentes no núcleo de uma célula. Adaptado de [Winslow, 2016]

No início dos anos 50, Erwin Chargaff [Chargaff, 1950] verificou certas regras empíricas sobre as quantidades de cada componente do ADN.

1ª Regra de Chargaff: Para a cadeia dupla verifica-se $A\% = T\%$ e $C\% = G\%$.

2ª Regra de Chargaff:

(a) Para cada cadeia $A\% \approx T\%$ e $C\% \approx G\%$;

(b) O total de pirimidinas $(T + C)$ é igual ao total de purinas $(A + G)$;

(c) $(C + G)\%$ não é necessariamente igual a $(A + T)\%$.

Em 1953, James Watson e Francis Crick [Watson and Crick, 1953] apresentaram o modelo da dupla hélice para a estrutura da molécula do ADN. Tal como foi possível observar na figura 1.1, o ADN é formado por duas cadeias helicoidais constituídas pelos nucleótidos, que se enrolam à volta de um eixo comum. O "esqueleto"de cada hélice é constituído pelas ligações entre o açúcar de um nucleótido e o fosfato do nucleótido seguinte. Uma das hélices é orientada de 5' para 3' e a outra de 3' para 5', devido ao sistema de numeração dos carbonos da desoxirribose [Reece, 2004]. As bases dos nucleótidos emparelham de modo complementar, formando pares de bases (pb), de tal forma que em frente a uma adenina fica sempre uma timina e em frente a uma citosina fica sempre uma guanina, ligadas por pontes de hidrogénio, sendo duas entre A e T e três entre C e G [Regateiro, 2007].

A sequência de nucleótidos encontrada numa das hélices do ADN pode ser interpretada como um conjunto de instruções que definem o funcionamento da célula. Ao conjunto completo de instruções genéticas encontrado numa célula chama-se genoma [Feero et al., 2010]. O genoma humano é composto por 23 pares de cromossomas, dos quais 22 pares autossomas e um par de cromossomas sexuais (XX ou XY). Normalmente, a maioria das células contêm duas cópias do genoma, as quais se designam como diplóides [Stram, 2014]. O genoma diplóide humano é constituído por cerca de $6 \times 10^9 pb$, sendo que uma parte corresponde a cerca de 30 000 a 40 000 genes. Cada gene é constituído, em média, por $1 \times 10^4 pb$ [Regateiro, 2007]. Define-se gene como uma sequência ordenada de nucleótidos, localizada numa determinada posição da cadeia de ADN, que tem a capacidade de codificar para a célula, um produto funcional específico, como por exemplo, uma proteína [Feero et al., 2010]. Os exões são as regiões do gene que codificam uma proteína e os intrões correspondem às partes não codificantes do gene. Às sequências de três nucleótidos consecutivos chamam-se codões, sendo estes responsáveis pela codificação de determinados aminoácidos [Draghici, 2011].

No que diz respeito à sequência de ADN, pode dizer-se que os humanos são bastante semelhantes entre si, cerca de 99,6%. No entanto, ao longo do genoma, há localizações específicas onde as diferenças entre indivíduos são normalmente referidas como variações. Numa determinada população, há variações que são mais comuns e outras que são mais raras. As formas alternativas de uma sequência genética, associadas a uma variação, num local específico do genoma, designam-se de alelos. Quando a frequência do alelo menor (MAF, *Minor Allele Frequency*) é superior a 1%, as variações são chamadas de polimorfismos [Feero et al., 2010]. Ao conjunto de todos os alelos de um indivíduo chama-se genótipo. Por vezes, o termo genótipo também é usado para referir um subconjunto de características genéticas individuais (como por exemplo, genes). O fenótipo corresponde às características observáveis de um indivíduo e normalmente é determinado pelo genótipo e pelas condições ambientais e de evolução. Por vezes, uma combinação de alelos podem ser transmitida de uma geração para outra. A essa combinação de alelos dá-se o nome de haplótipo [Draghici, 2011].

O tipo mais comum de variação genética no genoma humano é a variação de nucleótido único (SNV, *Single Nucleotide Variation*) que consiste na variação, entre indivíduos, de um único nucleótido num local específico do genoma. A cada SNV podem estar associados até 4 alelos, sendo que as SNVs bi-alélicas são as mais frequentes. Na figura 1.2 exemplifica-se, numa dada região do genoma, as sequências genéticas de diferentes indivíduos que podem diferir em apenas um nucleótido.

Uma SNV para ser considerada um polimorfismo de nucleótido único (SNP, *Single Nucleotide Polymorphism*) deve ter uma frequência de pelo menos 1%, em toda a população [Foulkes, 2009]. Neste trabalho, optou-se por usar sempre o termo geral SNV, uma vez que a distinção entre SNV e SNP é relativa pois a frequência das SNVs, geralmente dependem da população.

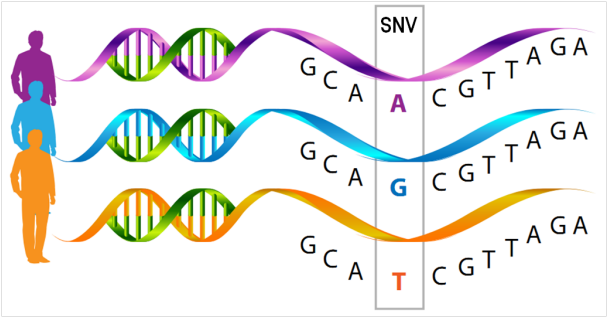


Figura 1.2: Esquema representativo da variação de nucleótido único (SNV) entre diferentes indivíduos. Adaptado de [NEI, 2016]

Tem-se ainda que as SNVs podem dividir-se em dois grupos: as transições (T_s), quando uma purina (pirimidina) é substituída por outra purina (pirimidina); e as transversões (T_v), quando uma purina (pirimidina) é substituída por uma pirimidina (purina). No genoma, a ocorrência de transições é mais frequente do que a de transversões. Por exemplo, quando uma citosina é substituída por uma timina, a variação ocorre no sentido $C \rightarrow T$, quando a variação corresponde à substituição de uma timina por uma citosina, o sentido é $T \rightarrow C$. Note-se que neste trabalho não se consideraram os sentidos de cada SNV. Assim, por exemplo, a variação $C \leftrightarrow T$ representará os casos C→T e T→C. Deste modo, globalmente, definem-se seis tipos de SNVs:

- T_s: $A \leftrightarrow G$ e $C \leftrightarrow T$;
- T_v: $A \leftrightarrow C$, $A \leftrightarrow T$, $C \leftrightarrow G$ e $G \leftrightarrow T$;

Note-se que muitas vezes, em estudos de associação genética (GWAS, *Genome-Wide Association Studies*), há interesse em investigar SNVs que estão presentes num gene, associado a uma determinada doença. No entanto, como normalmente se desconhece o local específico da SNV que pode desencadear a doença genética, é comum considerar um conjunto de SNVs próximos desse local, aos quais se chama marcadores genéticos (ver figura 1.3). O genótipo observado nesses locais tende a estar associado ao verdadeiro genótipo causador da doença [Foulkes, 2009].

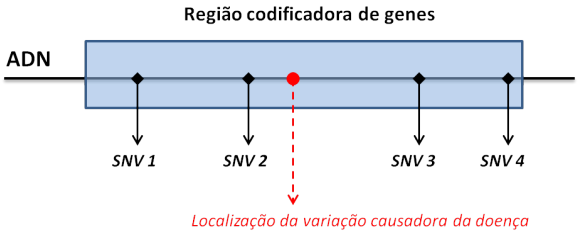


Figura 1.3: Marcadores SNVs numa região codificadora de genes.Adaptado de [Foulkes, 2009].

1.2 Projeto 1000 Genomas

O projeto 1000 Genomas (P1000G) [1000-Genomes-Project, 2016], decorrido entre 2008 e 2015, consistiu num esforço de colaboração internacional para desenvolver um catálogo público de dados sobre a variação genética humana. Este projeto foi o primeiro a sequenciar os genomas de um grande número de pessoas, algo possível, devido à redução dos custos do sequenciamento decorrente dos avanços nas técnicas de sequenciação. Os dados resultaram da combinação do sequenciamento de várias amostras usando várias "coberturas" de modo a permitir uma deteção mais eficaz das variações.

O P1000G realizou-se em quatro fases, uma fase piloto e três fases do projeto principal.

Na fase piloto foram sequenciadas, amostras do genoma de 179 indivíduos provenientes de 4 populações, amostras de trios familiares (pai-mãe-filho) e exões de 697 indivíduos pertencentes a 7 populações. Identificaram-se as localizações, frequências bi-alélicas e haplótipos de aproximadamente 15 milhões de SNVs, 1 milhão de *indels* (inserções ou deleções no genoma) e 20 000 variações estruturais (SV, *Structural Variation*). Pelo menos 95% das SNVs mais comuns (MAF $> 5\%$) foram identificadas na fase piloto do P1000G [1000Genomes, 2010].

Relativamente à fase 1 do P1000G, cujos resultados foram publicados em [1000Genomes, 2012], pretendeu-se caracterizar o espectro geográfico e funcional da variação genética humana, considerando amostras do genoma de 1092 indivíduos provenientes de 14 populações da Europa, leste da Ásia, África sub-Shariana e Américas. Nesta fase foram construidos os haplótipos de cerca de 38 milhões de SNVs, 1.4 milhões de *indels* e mais de 14000 grandes deleções, com especial foco nas variações raras ou de baixa frequência (MAF $< 5\%$), uma vez que permaneceram pouco caracterizadas na fase piloto. As amostras foram sequenciadas através da combinação de várias coberturas, sendo alta ($50-100\times$) em exões e baixa ($2-6\times$) no restante genoma [1000Genomes, 2012]. A figura 1.4 mostra um esquema exemplificativo do procedimento usado na construção dos haplótipos das variações bi-alélicas, em que cada uma das amostras dos 1092 indivíduos foi comparada com o genoma de referência. Para cada indivíduo e em cada posição, registaram-se quantas cópias $(0, 1, 2)$ existiam do alelo alternativo à referência. Na fase 1, identificaram-se até 98% de SNVs com uma frequência de 1% em populações relacionadas. Concluiu-se que as variações comuns são compartilhadas por todas as populações (com ancestrais africanos) e que as variações raras são específicas de determinadas populações. Provavelmente, as variações raras correspondem a mutações recentes que ainda não tiveram tempo suficiente para se espalhar por toda a população mundial, refletindo adaptação local [Zeggini and Morris, 2015].

Na fase 2 do projeto, não foram produzidos novos dados, uma vez que esta fase concentrou-se apenas, no desenvolvimento das técnicas utilizadas nas fases anteriores.

Mais recentemente, já após o início deste trabalho, em [1000Genomes, 2015a] e [1000Genomes, 2015b], foram divulgados os resultados da fase 3 do P1000G, na qual se recorreu a amostras de 2504 indivíduos de 26 populações. Nesta última fase, foram caracterizadas um total de 88 milhões de variações (84.7 milhões de SNVs, 3.6 milhões de pequenas indels e 60 000 SVs), dando especial foco às SVs. Para além disso, ao contrário das fases anteriores, estendeu-se a análise a variações multi-alélicas. Foram ainda validadas 80 milhões das variações catalogadas em [dbSNP, 2016].

Tendo em conta os objetivos principais desta dissertação que são, caracterizar o contexto nas vizinhanças dos locais de SNV e verificar se cada SNV bi-alélica satisfaz o equilíbrio de Hardy-Weinberg, recorreu-se a SNVs das cerca 38 milhões disponibilizadas pela fase 1 do P1000G, pois a verificação do equilíbrio em SNVs multi-alélicas seria muito mais complexa.

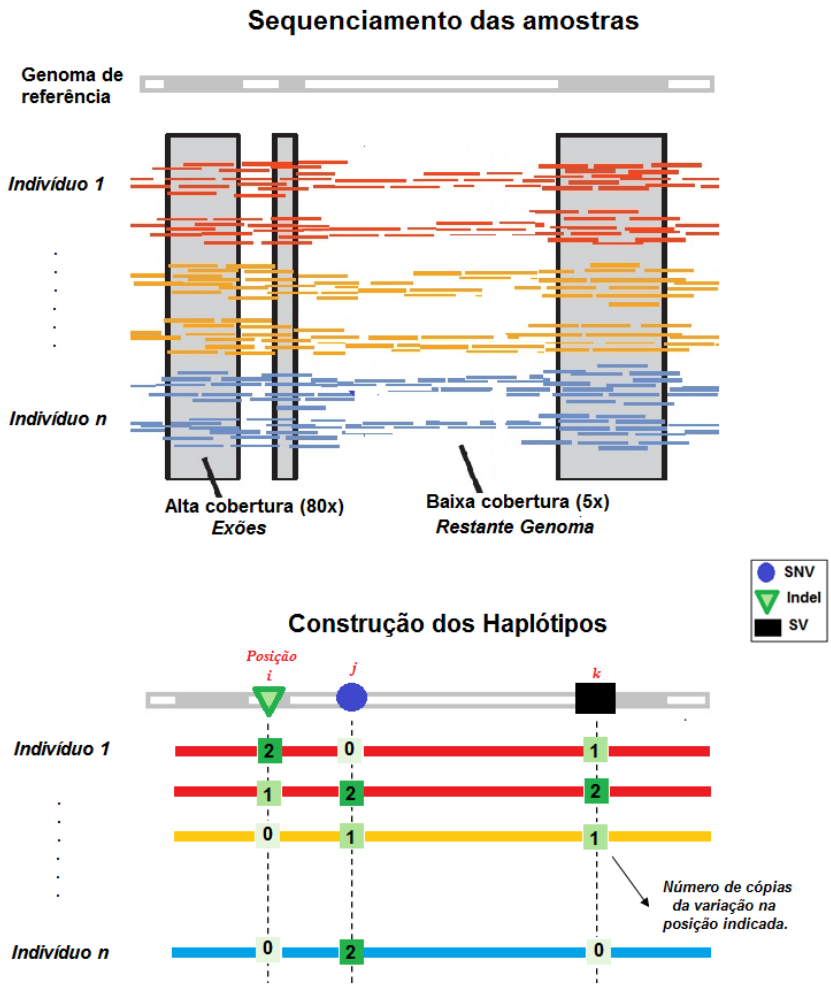


Figura 1.4: Procedimento usado na construção dos haplótipos das variações encontradas na fase 1 do P1000G. Adaptado de [1000Genomes, 2012] e [Zeggini and Morris, 2015].

1.3 R e Bioconductor

O R é uma linguagem de programação estatística de código aberto e que permite a manipulação eficiente de dados, fazer cálculos e gráficos [Draghici, 2011]. Trata-se de uma linguagem básica, que tem a sua própria sintaxe e que é enriquecida com pacotes (*packages*), os quais são desenvolvidos pelos diversos contribuidores. A versão atualizada e respetivos *packages* do R estão disponíveis a partir do CRAN (*Comprehensive R Archive Network*) [R-Project, 2016]. A interface do R é muito simples e funciona com comandos executados numa consola. O conjunto desses comandos podem ser guardados em ficheiros de texto simples designados de *scripts*, para serem usados posteriormente. Sendo uma linguagem de programação flexível permite que os utilizadores possam usar funções já existentes, quando querem realizar tarefas avançadas de análise ou, por outro lado, possam implementar as suas próprias funções. A versão utilizada neste trabalho para manipular os dados, efetuar as análises estatísticas e produzir gráficos foi a `3.3.1 (2016-06-21)`.

O Bioconductor é um projeto aberto, iniciado há cerca de dez anos, que desenvolve *software* para a análise e compreensão dos dados genómicos. Funciona principalmente com base na linguagem R, no entanto são aceites as contribuições feitas em qualquer tipo de linguagem de programação [Draghici, 2011]. Este projeto foi ganhando credibilidade ao longo do tempo devido à abordagem estatística rigorosa em diversas áreas da bioinformática. Presentemente existem mais de 600 *packages* para o Bioconductor, os quais podem ser instalados a partir de [Bioconductor, 2016]. Neste repositório pode ainda encontrar-se uma grande variedade de documentação, nomeadamente as vinhetas que servem de auxílio aos *packages* desenvolvidos.

A utilização do R e do Bioconductor tem particularmente interesse quando se pretende analisar dados genómicos cujas dimensões dos mesmos são muito grandes. Outra vantagem associada é ser possível partilhar informação e desenvolver ferramentas que posteriormente podem ser usadas e melhoradas para trabalhos e investigações futuras.

Ao longo deste trabalho recorreu-se a diversas funções provenientes dos *packages* do R e do Bioconductor. Para além disso, foram também criadas novas funções de acordo com o interesse da análise em questão. Dos *packages* utilizados neste trabalho destacam-se o HardyWeinberg, específico do *software* R, o Biotrings e o Bsgenome, específicos do Bioconductor.

- O *package* HardyWeinberg consiste num conjunto de ferramentas para analisar dados referentes a marcadores genéticos bi-alélicos. Sob determinadas condições, esses marcadores devem, em princípio, estar no equilíbrio de Hardy-Weinberg (HWE), que é descrito na secção 2.5. As SNVs são marcadores bi-alélicos que dão origem a 3 genótipos. Por exemplo, para a variação $A \leftrightarrow G$, os genótipos são AA, AG e GG. Os dados devem conter as frequências dos alelos e dos genótipos de cada SNV [Graffelman and Camarena, 2016]. Este *package* permite testar o HWE recorrendo à função `HWChisq`, desde que seja introduzido um vector com as contagens de cada um dos três genótipos. Esta função, por definição, inclui a correção à continuidade de Yates. Para não aplicar a correção à continuidade, basta acrescentar o parâmetro `cc=0`. Quando é necessário calcular as estatísticas χ^2 (ou valores-p), para um grande conjunto de SNVs (grandes matrizes com as contagens dos genótipos de cada SNV), usa-se a função `HWChisqStats`. Para além disso, este *package* fornece também várias ferramentas gráficas, que ajudam a tirar conclusões sobre a verificação do HWE num dado conjunto de SNVs. Dessas ferramentas destacam-se o `HWGenotypePlot` e o `HWTernaryPlot` [Graffelman, 2015].

- O *package* Biostrings fornece ferramentas para trabalhar com sequências genómicas. Neste *package* os objetos pertencem às classes `DNAString` e `DNAStringSet`, uma vez que é possível trabalhar com uma ou mais sequências de ADN. Existem ainda classes adicionais para representar aminoácidos e outras cadeias biológicas [Carlson et al., 2015]. A função `alphabetFrequency` permite contar a frequência de cada letra do alfabeto que compõe a sequência de ADN como por exemplo, o número de nucleótidos. Quando se pretende contabilizar palavras com outros tamanhos, como por exemplo dinucleótidos e trinucleótidos, pode recorrer-se à função `oligonucleotideFrequency`. Neste caso é possível especificar um parâmetro, k, correspondente ao comprimento da palavra, sendo $k = 1$ para contar nucleótidos, $k = 2$ para os dinucleótidos e $k = 3$ para os trinucleótidos;

- O *package* BSgenome armazena sequências do genoma completo de um dado organismo. Cada sequência é armazenada em objetos `DNAString` e normalmente tem origem num arquivo `FASTA`. Um ficheiro `FASTA` consiste em uma ou mais sequências, onde cada uma é precedida por uma única linha que fornece um identificador único para a sequência

e outras informações. Depois da linha de cabeçalho e comentários, a sequência é representada numa linha [Gentleman, 2009]. O BSgenome.Hsapiens.UCSC.hg19 fornece as sequências completas do genoma de referência para o Homo sapiens, versão GRCh37, disponível em [UCSC, 2016]. Também em [NCBI, 2016] é possível aceder ao genoma de referência. A função `str` utiliza-se para mostrar um resumo dos ficheiros e a função `getSeq` permite visualizar uma dada sequência num dado local do genoma, especificando qual o cromossoma e respetiva posição.

A figura 1.5 mostra a estrutura do ficheiro em R, da versão GRCh37 do genoma de referência humano. Apresenta-se ainda um exemplo de uma sequência de ADN, entre as posições 16050408 e 16050520, extraída do cromossoma 1 e respetiva contagem dos dinucleótidos.

```
> str(Hsapiens)
Formal class 'BSgenome' [package "BSgenome"] with 17 slots
  ..@ pkgname           : chr "BSgenome.Hsapiens.UCSC.hg19"
  ..@ single_sequences  :Formal class 'TwobitNamedSequences' [package "BSgenome"] with 1 slot

    .........

  ..@ organism          : chr "Homo sapiens"
  ..@ common_name       : chr "Human"
  ..@ provider          : chr "UCSC"
  ..@ provider_version  : chr "hg19"
  ..@ release_date      : chr "Feb. 2009"
  ..@ release_name      : chr "Genome Reference Consortium GRCh37"
  ..@ seqinfo           :Formal class 'Seqinfo' [package "GenomeInfoDb"] with 4 slots
  .. .. ..@ seqnames   : chr [1:93] "chr1" "chr2" "chr3" "chr4" ...
  .. .. ..@ seqlengths : int [1:93] 249250621 243199373 198022430 191154276 180915260  ...
  .. .. ..@ is_circular: logi [1:93] FALSE FALSE FALSE FALSE FALSE FALSE ...
  .. .. ..@ genome     : chr [1:93] "hg19" "hg19" "hg19" "hg19" ...

> getSeq(Hsapiens, "chr1", start=16050408, end=16050460)
  53-letter "DNAString" instance
seq: GGCCGGCAAATAGAGCACGACCAGCCATGGCCCTTGCCATTGACCCTCACCCA

> oligonucleotideFrequency(getSeq(Hsapiens, "chr1", start=16050408, end=16050460),2)
AA AC AG AT CA CC CG CT GA GC GG GT TA TC TG TT
 2  4  3  3  7 10  2  2  3  6  3  0  1  1  3  2
```

Figura 1.5: Estrutura do ficheiro em R da versão GRCh37 do genoma de referência humano. Exemplo de sequência de ADN, entre as posições 16050408 e 16050520, do cromossoma 1 e respetiva contagem dos dinucleótidos.

1.4 Descrição dos dados

Tal como já foi referido anteriormente, os dados utilizados neste trabalho são os da fase 1 do P1000G, em que foram consideradas as amostras do genoma de 1092 indivíduos e nos quais foram reportados cerca de 38 milhões de SNVs.

Os dados correspondentes a essas variações são armazenados num formato de arquivo de texto `VCF` (*Variant Call Format*), que corresponde a um ficheiro que contém linhas de meta-informação, uma linha de cabeçalho e as linhas de dados [VCF.file, 2016].

A figura 1.6 mostra um exemplo de um ficheiro `VCF` extraído dos dados do P1000G, correspondente ao cromossoma 1. Os detalhes omitidos foram substituidas por "...".

Cada registo contém vários campos de informação para uma única variação local. Esses campos informativos são:

- `CHROM` e `POS` identificam o cromossoma e o local da variação relativo ao genoma de referência;
- `ID` é uma lista de identificadores únicos. Caso seja uma variação da *dbSNP* (repositório de dados público sobre SNPs disponível em [dbSNP, 2016]) utiliza-se a designação `rs`;
- `REF` e `ALT` são os campos onde é codificado o tipo de variação, pois especificam, respetivamente, o nucleótido no genoma de referência e o nucleótido alternativo observado em pelo menos uma das cópias dos cromossomas homólogos das amostras individuais;
- `FORMAT` é o campo que contém a codificação utilizada nas anotações das amostras individuais. Por exemplo, na primeira linha de dados, tanto a amostra `HG00096` como a `HG00097` têm o genótipo (`GT`) igual a `0|0`, o que significa que no local `10583` (posição no cromossoma 1) os indivíduos são homozigóticos com o alelo da referência ou seja, `G|G`. Na terceira linha, para a posição `13302`, a amostra `HG00097` tem o genótipo igual a `0|1`, o que significa que o indivíduo é heterozigótico, isto é, `C|T`. Pode também existir outros indivíduos que registem o genótipo do tipo `1|0`, que corresponde ao caso `T|C`, ou ainda indivíduos com o genótipo `1|1` (homozigóticos com o alelo alternativo), correspondente ao caso `T|T`.

```
##fileformat=VCFv4.1
##INFO=<ID=LDAF,Number=1,Type=Float,Description="MLE Allele Frequency Accounting for LD">
##INFO=<ID=AVGPOST,Number=1,Type=Float,Description="Average posterior probability from MaCH/Thunder">
...
##reference=GRCh37
#CHROM  POS    ID           REF  ALT  QUAL  FILTER  INFO  FORMAT    HG00096  HG00097
1       10583  rs58108140   G    A    100   PASS    …     GT:DS:GL  0|0:…    0|0:…
1       10611  rs189107123  C    G    100   PASS    …     GT:DS:GL  0|0:…    0|1:…
1       13302  rs180734498  C    T    100   PASS    …     GT:DS:GL  0|0:…    0|1:…
1       13327  rs144762171  G    C    100   PASS    …     GT:DS:GL  0|0:…    0|1:…
...
```

Figura 1.6: `VCF` extraído dos dados do P1000G, correspondente ao cromossoma 1.

Para pré-processar a informação contida nos arquivos `VCF` do P1000G, foi escrito um pequeno programa em `C`, que consistiu em:

1. Descartar campos não desejados;
2. Selecionar apenas variações SNV (rejeitando *indels* e SV);
3. Contabilizar quantas amostras têm o genótipo igual a `0|0`, `0|1`, `1|0` ou `1|1`.

Assim, foram produzidos ficheiros mais pequenos, com apenas algumas colunas, tal como é exemplificado na figura 1.7. As colunas `C0|0:`, `C0|1:`, `C1|0:` e `C1|1:` mostram o número de amostras individuais de cada genótipo, isto é, `C0|0:` corresponde ao número de indivíduos, que comparativamente ao genoma de referência e na posição considerada, não apresentaram nenhuma alteração; `C0|1:` e `C1|0:` os que apresentaram alteração numa das cópias; e `C1|1:` os que verificaram alteração nas duas cópias.

#CHROM	POS	ID	REF	ALT	C0|0:	C0|1:	C1|0:	C1|1:
1	10583	rs58108140	G	A	783	304	0	5
1	10611	rs189107123	C	G	1051	37	4	0
1	13302	rs180734498	C	T	849	192	45	6
1	13327	rs144762171	G	C	1033	44	15	0
1	13980	rs151276478	T	C	1047	33	12	0
1	30923	rs140337953	G	T	146	177	131	638
1	51476	rs187298206	T	C	1074	12	6	0
1	51479	rs116400033	T	A	905	69	70	48

...

Figura 1.7: *Output* do pré-processamento do `VCF` extraído dos dados do P1000G, correspondente ao cromossoma 1.

1.5 Objectivos gerais

Com o trabalho descrito nesta dissertação, pretende-se investigar e caracterizar as variações genéticas de nucleótido único, recorrendo aos dados da fase 1 disponibilizados pelo P1000G. Assim, os objetivos principais são:

- Desenvolver ferramentas de análise estatística no *software* `R` para analisar os diferentes tipos de SNVs;
- Averiguar se as ocorrências das SNVs são distribuídas de forma homogénea por cromossoma e por grupo de prevalência;
- Testar, para cada tipo de SNV numa localização específica no genoma, se as frequências dos alelos estão de acordo com o equilíbrio de Hardy-Weinberg;
- Caracterizar o contexto na vizinhança da posição onde ocorre a SNV, através de perfis de frequência que consistem na contagem de nucleótidos, dinucleótidos e trinucleótidos, de modo a encontrar um padrão que seja indicador do fenómeno.

1.6 Estrutura da dissertação

Esta dissertação é constituída por cinco capítulos e dois apêndices.

Neste Capítulo 1, introduzem-se alguns conceitos de genómica, faz-se uma breve descrição do projeto 1000 Genomas, do *software* R e dos *packages* do Bioconductor. Apresentam-se ainda os dados analisados e os objetivos gerais da dissertação.

No Capítulo 2, abordam-se os métodos estatísticos utilizados na análise dos dados da fase 1 do projeto 1000 Genomas. Inicia-se o capítulo com a notação geral utilizada nas tabelas de contingência, seguindo-se o teste de independência/homogeneidade, as medidas de associação e a análise de resíduos, para avaliar a força de associação entre as variáveis. Como também foi necessário, estudar a distribuição de uma dada variável e testar se duas variáveis contínuas apresentam a mesma distribuição, aborda-se ainda, o teste de ajustamento do χ^2, o teste de Kolmogorov-Smirnov para duas amostras independentes e correções para corrigir o erro do tipo I, devido às comparações múltiplas. Em seguida, apresenta-se a metodologia estatística que suporta o equilíbrio de Hardy-Weinberg. Termina-se o capítulo com a classificação hierárquica, uma vez que foi necessário agrupar diversas variáveis de forma a identificar padrões.

No Capítulo 3 e 4, apresentam-se os resultados da aplicação dos métodos estatísticos às diversas variáveis de interesse. Essas variáveis são, por exemplo, o tipo de variação, os cromossomas, a prevalência de cada variação, o grupo de prevalência e o contexto na vizinhança do local onde ocorre a variação. Assim, o Capítulo 3, apresenta uma análise global dos diferentes tipos de SNV, por cromossoma e por grupo de prevalência, assim como os resultados da avaliação do equilíbrio de Hardy-Weinberg nas SNVs consideradas. O Capítulo 4, mostra a caracterização do contexto das vizinhanças nos locais onde ocorrem as variações, recorrendo à contagem de nucleótidos, dinucleótidos e trinucleótidos de acordo com várias amplitudes da vizinhança da posição onde se registou a variação.

Por fim, o Capítulo 5, apresenta as conclusões da dissertação e algumas ideias de trabalho futuro.

Nos Apêndices A e B, encontram-se alguns gráficos e tabelas auxiliares à análise estatística, assim como, as funções que foram desenvolvidas em R e o respetivo código.

Capítulo 2

Métodos estatísticos de análise de dados

Neste capítulo apresentam-se alguns dos métodos estatísticos de análise de dados utilizados neste trabalho. Devido à necessidade de estudar a relação entre duas variáveis categóricas através do uso de tabelas de contigência, o capítulo inicia-se com a notação geral utilizada nas tabelas de contingência, o teste de independência/homogeneidade, as medidas para estudar a força da associação entre variáveis e a respetiva análise de resíduos. Em seguida, aborda-se o teste de ajustamento do χ^2 e o teste de Kolmogorov-Smirnov para a comparação de duas amostras, uma vez que foi necessário estudar a distribuição de algumas variáveis e testar se duas variáveis contínuas apresentam a mesma distribuição. Por outro lado, para corrigir os valores-p obtidos nas comparações múltiplas de certas variáveis, faz-se uma breve referência às correções de Bonferroni e Šidák. Apresenta-se ainda a metodologia estatística inerente ao equilíbrio de Hardy-Weinberg. O capítulo termina com a classificação hierárquica, uma vez que foi necessário agrupar variáveis, para, por exemplo, compreender padrões de ocorrência nas variações.

Note-se que na realização de qualquer teste estatístico é usual definir um nível de significância α. Neste trabalho, considerar-se-á um dos níveis de significância mais usuais, que é $\alpha = 0.05$.

2.1 Tabelas de Contingência

As tabelas de contingência são tabelas que servem para organizar os dados de variáveis qualitativas, em que o objetivo principal é avaliar se existe associação entre as variáveis.

Na análise de dados genómicos é frequente recorrer à análise de tabelas de contingência, uma vez que a maioria dos dados tratados são qualitativos. Por exemplo, no trabalho de [Afreixo et al., 2006] construiram-se tabelas de contingência 64×64 para estudar a associação entre pares de codões nas sequências genomicas de *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Candida albicans* e *Escherichia coli*, de forma a descrever o contexto de ocorrência dos codões.

Neste trabalho, foi necessário construir diversas tabelas de contigência para estudar a existência de independência/homogeneidade entre variáveis categóricas como por exemplo, tipo de variação, cromossoma, grupo de prevalência e contexto na vizinhança do local onde ocorreu a variação.

De acordo com a nomenclatura usada em [Everitt, 1977], suponha-se que uma amostra de N observações é classificada em relação a duas variáveis categóricas A e B, com r e c categorias respetivamente, mutuamente exclusivas. A forma geral de uma tabela de contingência bidimensional, $r \times c$, formada por A (linha) e B (coluna), subdivididas nas categorias $A_1, ..., A_r$ e $B_1, ..., B_c$ é representada pela tabela 2.1.

	B_1	...	B_j	...	B_c	Total
A_1	n_{11}	...	n_{1j}	...	n_{1c}	$n_{1\cdot}$
...	...	...	...	...	...	...
A_i	n_{i1}	...	n_{ij}	...	n_{ic}	$n_{i\cdot}$
...	...	...	...	...	...	...
A_r	n_{r1}	...	n_{rj}	...	n_{rc}	$n_{r\cdot}$
Total	$n_{\cdot 1}$	...	$n_{\cdot j}$	...	$n_{\cdot c}$	N

Tabela 2.1: Forma geral de um tabela de contingência $r \times c$.

Cada célula da tabela contém a frequência observada de um certo par de categorias. O número de observações que são simultaneamente da categoria A_i da variável A e da categoria B_j da variável B, isto é, da célula (i, j), é dado por n_{ij}. Representa-se por $n_{i\cdot}$ o número total (marginal) de observações na categoria A_i e por $n_{\cdot j}$ o número total (marginal) de observações na categoria B_j, com $i = 1, ..., r$ e $j = 1, ..., c$, definidos por:

$$n_{i\cdot} = \sum_{j=1}^{c} n_{ij} \quad e \quad n_{\cdot j} = \sum_{i=1}^{r} n_{ij} \tag{2.1}$$

Assim, resulta que:

$$\sum_{j=1}^{c} \sum_{i=1}^{r} n_{ij} = \sum_{i=1}^{r} n_{i\cdot} = \sum_{j=1}^{c} n_{\cdot j} = N \tag{2.2}$$

Para $i = 1, ..., r$ e $j = 1, ..., c$ sejam:

- $p_{ij} = P(A_i \cap B_j)$ a probabilidade de uma observação pertencer simultaneamente à categoria A_i e à categoria B_j;
- $p_{i\cdot} = P(A_i)$ a probabilidade marginal de uma observação pertencer à categoria A_i;
- $p_{\cdot j} = P(B_j)$ a probabilidade marginal de uma observação pertencer à categoria B_j.

Neste caso, tem-se que:

$$p_{i\cdot} = P(A_i) = \sum_{j=1}^{c} p_{ij} \quad e \quad p_{\cdot j} = P(B_j) = \sum_{i=1}^{r} p_{ij} \tag{2.3}$$

$$\sum_{j=1}^{c} \sum_{i=1}^{r} p_{ij} = \sum_{i=1}^{r} p_{i\cdot} = \sum_{j=1}^{c} p_{\cdot j} = 1. \tag{2.4}$$

2.1.1 Teste de independência/homegeneidade

Segundo [Kateri, 2014], na tabela de contingência 2.1 podem surgir duas situações:

(1) fixa-se o total N;

(2) fixam-se, por exemplo, os totais marginais das colunas $n_{\cdot j}$.

Na situação (1) a amostra aleatória é recolhida de acordo com a dimensão N pré-fixada e as respetivas observações são cruzadas de acordo com as categorias de A e de B, em que o objetivo é testar a independência ou inexistência de associação entre as variáveis A e B.

Em (2), quando os totais das colunas da variável B são prefixadas (não aleatórias), a hipótese que se pretende ensaiar é a da homogeneidade.

Independência

Fixando o total N, o vetor aleatório $(N_{11}, ..., N_{ij}, ..., N_{rc})$ segue uma distribuição multinomial, sendo N_{ij} a variável aleatória que representa o número de observações na célula (i, j) para $i = 1, ..., r$ e $j = 1, ..., c$. A respetiva função de verosimilhança é dada por:

$$L(p_{11}, ..., p_{rc}) = \frac{N!}{\prod_{i=1}^{r} \prod_{j=1}^{c} n_{ij}!} \prod_{i=1}^{r} \prod_{j=1}^{c} p_{ij}^{n_{ij}}. \tag{2.5}$$

Caso as variáveis A e B sejam independentes, verifica-se a relação:

$$P(A_i \cap B_j) = P(A_i)P(B_j) \tag{2.6}$$

que exprime a igualdade entre a probabilidade conjunta e o produto das probabilidades marginais. Portanto, no teste de independência, a hipótese nula corresponde a:

$$H_0: \quad p_{ij} = p_{i\cdot} p_{\cdot j} \quad para \quad i = 1, ..., r; j = 1, ..., c \tag{2.7}$$

Na equação 2.7 as probabilidades marginais, $p_{i\cdot}$ e $p_{\cdot j}$, são parâmetros desconhecidos. No entanto, as respetivas estimativas de verosimilhança máxima, quando H_0 é verdadeira, correspondem às frequências relativas marginais $\hat{p}_{i\cdot}$ e $\hat{p}_{\cdot j}$ [Agresti, 2010].

Teorema 2.1. *Sejam A e B variáveis categóricas cujas observações, n_{ij}, de cada célula (i, j) com $i = 1, ..., r$ e $j = 1, ..., c$, correspondem a uma amostra aleatória de dimensão N. Considere-se $(N_{11}, ..., N_{ij}, ..., N_{rc})$ o vetor aleatório com distribuição multinomial de parâmetros (N, p_{ij}). Então, assumindo que as variáveis A e B são independentes, as estimativas de verosimilhança máxima, das probabilidades marginais $p_{i\cdot}$ e $p_{\cdot j}$ são dadas por:*

$$\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{N} \quad e \quad \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{N}. \tag{2.8}$$

Demonstração. Seja L a função de verosimilhança do vetor aleatório $(N_{11}, ..., N_{ij}, ..., N_{rc})$ dada pela equação 2.5. Logaritmizando a função L, obtém-se:

$$ln(L) = ln(\frac{N!}{\prod_{i=1}^{r} \prod_{j=1}^{c} n_{ij}!}) + ln(\prod_{i=1}^{r} \prod_{j=1}^{c} p_{ij}^{n_{ij}}) \tag{2.9}$$

Como a primeira parcela do segundo membro não depende de p_{ij}, para estudar os maximizantes da função de verosimilhança, considerar-se-á apenas a segunda parcela. Assim, assumindo a independência e tendo em conta as equações 2.3 e 2.4 tem-se que:

$$ln(\prod_{i=1}^{r}\prod_{j=1}^{c} p_{ij}^{n_{ij}}) = ln(\prod_{i=1}^{r} p_{i\cdot}^{n_{i\cdot}} \prod_{j=1}^{c} p_{\cdot j}^{n_{\cdot j}}) = ln(p_{r\cdot}^{n_{r\cdot}} p_{\cdot c}^{n_{\cdot c}} \prod_{i=1}^{r-1} p_{i\cdot}^{n_{i\cdot}} \prod_{j=1}^{c-1} p_{\cdot j}^{n_{\cdot j}})$$

$$= ln((1-\sum_{i=1}^{r-1} p_{i\cdot})^{n_{r\cdot}} (1-\sum_{j=1}^{c-1} p_{\cdot j})^{n_{\cdot c}} \prod_{i=1}^{r-1} p_{i\cdot}^{n_{i\cdot}} \prod_{j=1}^{c-1} p_{\cdot j}^{n_{\cdot j}})$$

$$= \sum_{i=1}^{r-1}[\frac{n_{r\cdot}}{r-1} ln(1-\sum_{i=1}^{r-1} p_{i\cdot}) + n_{i\cdot} ln(p_{i\cdot})] + \sum_{j=1}^{c-1}[\frac{n_{\cdot c}}{c-1} ln(1-\sum_{j=1}^{c-1} p_{\cdot j}) + n_{\cdot j} ln(p_{\cdot j})]$$

Portanto, as derivadas parciais de $ln(L)$ são dadas por:

$$\frac{\partial ln(L)}{\partial p_{i\cdot}} = -\frac{n_{r\cdot}}{r-1} \frac{r-1}{1-\sum_{i=1}^{r-1} p_{i\cdot}} + \frac{n_{i\cdot}}{p_{i\cdot}}$$

$$\frac{\partial ln(L)}{\partial p_{\cdot j}} = -\frac{n_{\cdot c}}{c-1} \frac{c-1}{1-\sum_{j=1}^{c-1} p_{\cdot j}} + \frac{n_{\cdot j}}{p_{\cdot j}}$$

Resolvendo a equação $\frac{\partial ln(L)}{\partial p_{i\cdot}} = 0$, obtém-se:

$$p_{i\cdot} = \frac{n_{i\cdot}}{\frac{n_{r\cdot}}{p_{r\cdot}}} \quad . \tag{2.10}$$

Uma vez que $\sum_{i=1}^{r} n_{i\cdot} = N$, resulta que:

$$1 = \sum_{i=1}^{r} p_{i\cdot} = \sum_{i=1}^{r-1} \frac{n_{i\cdot}}{\frac{n_{r\cdot}}{p_{r\cdot}}} + p_{r\cdot} = \frac{N-n_{r\cdot}}{\frac{n_{r\cdot}}{p_{r\cdot}}} + p_{r\cdot} = p_{r\cdot}\frac{N}{n_{r\cdot}} \Leftrightarrow p_{r\cdot} = \frac{n_{r\cdot}}{N} \quad .$$

Substituindo na equação 2.10, tem-se:

$$p_{i\cdot} = \frac{n_{i\cdot}}{N} \quad .$$

De forma análoga, a partir de $\frac{\partial ln(L)}{\partial p_{\cdot j}} = 0$, obter-se-ia:

$$p_{\cdot j} = \frac{n_{\cdot j}}{N} \quad .$$

Verificando que a segunda derivada é negativa, conclui-se a demonstração. □

Assim, verificadas as condições do teorema, as estimativas apresentadas na equação 2.8, permitem estimar a frequência esperada de uma observação pertencer à célula (i, j), isto é:

$$\hat{e}_{ij} = N\hat{p}_{i\cdot}\hat{p}_{\cdot j} = \frac{n_{i\cdot}n_{\cdot j}}{N}. \tag{2.11}$$

A estatística do teste, sugerida por Karl Pearson (1904), é dada por:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \quad . \tag{2.12}$$

Para um tamanho fixo da amostra, grandes diferenças $n_{ij} - \hat{e}_{ij}$ produzem grandes valores para χ^2 e mais forte é a evidência contra H_0. Como maiores valores de χ^2 são mais contraditórios para H_0, o valor-p é a probabilidade de χ^2 ser pelo menos tão grande quanto o valor observado. A estatística χ^2 tem aproximadamente uma distribuição de qui-quadrado, para N grande. A aproximação qui-quadrado melhora à medida que $\hat{e}_{ij}$ aumenta e $\hat{e}_{ij} \geq 5$ é geralmente suficiente para uma aproximação razoável [Agresti, 2007].

Note-se que, o número de parâmetros independentes é $(r-1)+(c-1)$ pois o número total de parâmetros é $r + c$ $(p_{1\cdot}, ..., p_{r\cdot}, p_{\cdot 1}, ..., p_{\cdot c})$ e a soma das probabilidades marginais por linha (coluna) são iguais a 1, na tabela de contingência $r \times c$ [Kateri, 2014]. Assim, o número de parâmetros a estimar sob H_0 é

$$(r-1) + (c-1)$$

donde o número de graus de liberdade $(g.l.)$ da estatística do teste é igual a

$$g.l. = rc - 1 - (r-1) - (c-1) = (r-1)(c-1)$$

ou seja, a distribuição assintótica para 2.12 sob H_0 é

$$\chi^2_{(r-1)(c-1)} \ .$$

Note-se que é frequente aplicar a chamada correção à continuidade proposta por Yates (1934) e substituir a estatística de teste, dada pela equação 2.12, por

$$\chi^2_c = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(|n_{ij} - \hat{e}_{ij}| - c)^2}{\hat{e}_{ij}} \tag{2.13}$$

onde c representa a correção à continuidade e cujo valor usual é $c = 0.5$. A estatística de teste corrigida 2.13, normalmente produz, valores-p mais próximos dos exatos do que a estatística sem a correção [Hitchcock, 2016].

Homogeneidade

Fixando, por exemplo, os totais marginais das colunas da variável B, a hipótese testada é a da homogeneidade. Como cada B_j, $j \in \{1, ..., c\}$ estratifica uma subpopulação cujos elementos se distribuem pelas r categorias da variável A, a homogeneidade verifica-se quando de subpopulação para subpopulação, são iguais as proporções de observações em cada A_i, com $i \in \{1, ..., r\}$. De acordo com [Agresti, 2010], a probabilidade condicionada da categoria A_i na subpopulação B_j, é dada por

$$p_{i|j} = \frac{p_{ij}}{p_{\cdot j}}$$

e tal que

$$\sum_{i=1}^{r} p_{i|j} = 1 \, , \; j = 1, ..., c \; .$$

Neste caso, a hipótese testada é:

$$H_0 : p_{i|1} = p_{i|2} = ... = p_{i|c} \;\; \forall i \; . \tag{2.14}$$

A respetiva função de verosimilhança [Kateri, 2014] é dada por:

$$L(p_{11}, ..., p_{rc}) = \prod_{j=1}^{c} L(p_{1j}, ..., p_{rj}) = \prod_{j=1}^{c} (\frac{n_{\cdot j}!}{\prod_{i=1}^{r} n_{ij}!} \prod_{i=1}^{r} p_{i|j}^{n_{ij}}). \tag{2.15}$$

Quando a hipótese nula é verdadeira, considerando as subpopulações B_j e a frequência relativa da categoria A_i, no conjunto das N observações, a estimativa de verosimilhança máxima do valor comum das probabilidades condicionadas é,

$$\hat{p}_{i|1} = \hat{p}_{i|2} = ... = \hat{p}_{i|c} = \frac{n_{i\cdot}}{N} \; .$$

As frequências esperadas em cada célula (i, j) são obtidas multiplicando a dimensão de cada subamostra $(n_{\cdot j})$ pela estimativa da probabilidade da categoria A_i, comum a todas as subpopulações ou seja,

$$n_{\cdot j} \hat{p}_{i|j} = \frac{n_{i\cdot} n_{\cdot j}}{N} \; . \tag{2.16}$$

A estatística do teste da homogeneidade é igual à que foi apresentada na equação 2.12, mantendo-se os mesmos graus de liberdade, pois cada subpopulação contribui com $r-1$, sendo o total de $c(r-1)$. Como são estimados $r-1$ parâmetros (um por cada linha) [Murteira, 1990], tem-se

$$g.l. = c(r-1) - (r-1) = (r-1)(c-1) \; .$$

2.1.2 Medidas de Associação

Na maioria dos casos, quando se analisa tabelas de contingência há interesse em medir a força de associação entre as duas variáveis qualitativas envolvidas. O valor calculado da estatística de Pearson χ^2 tem a desvantagem de os respetivos graus de liberdade $(r-1)(c-1)$, dependerem da dimensão da tabela de contingência $(r \times c)$. Deste modo, são sugeridas várias medidas de associação que são baseadas na estatística χ^2, mas que não refletem o tamanho da amostra N [Everitt, 1977]. As medidas de associação utilizadas neste trabalho foram o coeficiente ϕ e o coeficiente V de Cramér, as quais se encontram descritas em seguida.

- **Coeficiente ϕ**

  O coeficiente ϕ é uma medida baseada na estatística χ^2 e é dado por

$$\phi = \sqrt{\frac{\chi^2}{N}} \tag{2.17}$$

  ou equivalentemente, $\phi^2 = \frac{\chi^2}{N}$. Trata-se de uma medida que não é completamente satisfatória uma vez que ϕ^2 não tem necessariamente um limite superior igual a 1. Existem várias propostas para valores de ϕ que podem servir como critérios para identificar a magnitude do tamanho de efeito da associação. Por exemplo: pequeno $(0,1 \leq \phi < 0,3)$, médio $(0,3 \leq \phi < 0,5)$ e grande $(\phi \geq 0,5)$ [Sheskin, 2003].

- **Coeficiente V de Cramér**

  Sugerido por Cramér (1946), o coeficiente é definido por

$$V = \sqrt{\frac{\chi^2/N}{\min(r-1, c-1)}} \ . \tag{2.18}$$

  Assumindo os valores entre 0 e 1, sendo igual a 0 quando não há nenhuma relação entre as duas variáveis categóricas e igual a 1 quando existe dependência completa. Note-se que quando a tabela de contigência é quadrada $(r = c)$ e $V = 1$, existe uma perfeita associação entre as duas variáveis e neste caso todas as observações estão distribuidas numa das diagonais da tabela. No entanto, quando $r \neq c$ e $V = 1$, a associação perfeita não é interpretada do mesmo modo de quando a tabela é quadrada. Neste caso a sua interpretação depende dos valores de r e c [Sheskin, 2003].

2.1.3 Análise de Resíduos

Tal como já foi referido, a estatística do teste χ^2 e o respetivo valor-p descrevem a evidência contra a hipótese nula. Uma comparação célula a célula das frequências observadas e esperadas ajuda a compreender melhor a natureza dessa evidência. As maiores diferenças entre n_{ij} e $\hat{e}_{ij}$ tendem a ocorrer em células que têm maiores frequências esperadas, portanto a diferença $n_{ij} - \hat{e}_{ij}$ é insuficiente [Agresti, 2007]. Assim, será útil considerar para cada célula (i, j) os resíduos de Pearson, dados por

$$r_{ij} = \frac{n_{ij} - \hat{e}_{ij}}{\sqrt{\hat{e}_{ij}}} \tag{2.19}$$

onde

$$\sum_{i=1}^{r}\sum_{j=1}^{c} r_{ij}^2 = \chi^2 \ .$$

Na condição de independência, a variância de $\hat{e}_{ij}$ é estimada por v_{ij} onde v_{ij} corresponde à variância assintótica da frequência esperada da célula (i, j). Neste caso, os resíduos de Pearson seguem assintoticamente uma distribuição normal $r_{ij} \sim N(0, v_{ij})$ com $v_{ij} \neq 1$ [Kateri, 2014].

Haberman (1973) provou que sob a hipótese de independência e para uma amostra multinomial,

$$v_{ij} = (1 - p_{i\cdot})(1 - p_{\cdot j})$$

quando $N \to \infty$. Substituindo pelos estimadores de verosimilhança máxima definidos nas equações 2.8, com $i = 1, ..., r$ e $j = 1, ..., c$, as variâncias estimadas são

$$\hat{v}_{ij} = \left(1 - \frac{n_{i\cdot}}{N}\right)\left(1 - \frac{n_{\cdot j}}{N}\right) . \tag{2.20}$$

Assim, os resíduos ajustados (*standardized residuals*) podem definir-se como

$$d_{ij} = \frac{r_{ij}}{\sqrt{\hat{v}_{ij}}} . \tag{2.21}$$

O seguinte teorema, sobre resíduos ajustados, corresponde a um resultado de Haberman, que pode ser encontrado em [Agresti, 2010], [Everitt, 1977] e [Santner and Duffy, 1989].

Teorema 2.2. *Seja N_{ij} o número de observações da célula (i, j) numa tabela de contingência $r \times c$, correspondente a uma amostra casual de N observações das variáveis A e B. Se $(N_{11}, ..., N_{ij}, ..., N_{rc})$ é um vetor aleatório com distribuição multinomial então, sob a hipótese de independência de A e B, os resíduos ajustados D_{ij}, com os valores observados d_{ij} dados pela equação 2.21, têm assintoticamente $(N \to +\infty)$ uma distribuição normal N(0,1).*

Note-se que caso os resíduos ajustados tenham uma distribuição muito diferente da $N(0, 1)$ dever-se-á rejeitar a hipótese de independência, na tabela de contingência $r \times c$.

Assumindo H_0 verdadeira, considerando um nível de confiança de 95%, por exemplo, tem-se que $P(-2 < d_{ij} < 2) = 0.95$. Espera-se que cerca de 5% dos resíduos ajustados fiquem mais distantes de 0 do que ± 2, isto é, pode identificar-se a célula (i, j) como responsável pela rejeição de independência, se o respetivo resíduo ajustado é tal que $|d_{ij}| \geq 2$ [Agresti, 2007].

Assim, face à rejeição de H_0, qualquer análise de uma tabela de contingência $r \times c$, deve incluir uma análise de resíduos que consista em:

i) Procurar células (i, j) cujos resíduos ajustados sejam $|d_{ij}| \geq 2$. Os valores negativos de d_{ij} estão associados às células preteridas relativamente a H_0 e os valores positivos associados às células preferidas;

ii) Construção de gráficos de resíduos de acordo com os respetivos índices de linha/coluna (por exemplo, *heatmaps*);

iii) Se N é grande, construção de histogramas dos resíduos ajustados ordenados ou gráficos de quantil-quantil de uma distribuição normal padrão (*QQ-plots*).

2.2 Teste de ajustamento do χ^2

Para avaliar se uma determinada amostra aleatória foi extraída de uma população com distribuição especificada, pode utilizar-se o teste de ajustamento do χ^2.

Considere-se uma amostra casual de N elementos, extraída de uma população com distribuição desconhecida, sobre os quais se observa uma característica (qualitativa ou quantitativa). As observações da característica em estudo são repartidas por r classes, mutuamente exclusivas, A_1, ..., A_r. Sejam:

- n_i o número de observações da classe A_i com $i = 1, ..., r$ e tal que $\sum_{i=1}^{r} n_i = N$;
- p_i a probabilidade (desconhecida) de obter uma observação da classe A_i, $\sum_{i=1}^{r} p_i = 1$;
- p_{0i} a probabilidade de obter uma observação da classe A_i, assumindo que a observação foi extraída de uma população com a distribuição especificada.

Neste caso, pretende-se ensaiar as r hipóteses,

$$H_0 : p_i = p_{0i} \quad i = 1, ..., r \ . \tag{2.22}$$

Quando H_0 é verdadeira, a frequência esperada de uma observação estar na classe A_i é dada por $\hat{e}_i = Np_{0i}$. A estatística de teste, obtida por Karl Pearson, baseia-se numa medida de ajustamento entre as frequências observadas, n_i, e as frequências esperadas, $\hat{e}_i$, sendo definida por

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - \hat{e}_i)^2}{\hat{e}_i} \tag{2.23}$$

Sob a validade de H_0, devem registar-se pequenas diferenças entre cada valor observado e o respetivo valor esperado. Um valor elevado para χ^2 é indicador de que há um desajuste entre a distribuição de frequências amostrais e teóricas, levando à rejeição da hipótese nula.

Seja ainda N_i, a variável aleatória que representa o número de observações na categoria A_i, com $i \in \{1, ..., r\}$. Neste caso, o vetor aleatório $(N_1, ..., N_{r-1})$ tem distribuição multinomial, tal que $\sum_{i=1}^{r} N_i = N$, com função de probabilidade dada por

$$P(N_1 = n_1, ..., N_{r-1} = n_{r-1}) = \frac{N!}{n_1! \dots n_r!} p_1^{n_1} \dots p_r^{n_r}, \tag{2.24}$$

onde $n_r = N - n_1 - n_2 - ... - n_{r-1}$ e $p_r = 1 - p_1 - p_2 - ... - p_{r-1}$ [Agresti, 2010].

Teorema 2.3. *Seja $(N_1, ..., N_{r-1})$ um vetor aleatório com distribuição multinomial de parâmetros $N, p_{01}, ..., p_{0(r-1)}$. Então, a variável aleatória*

$$\chi^2 = \sum_{i=1}^{r} \frac{(N_i - Np_{0i})^2}{Np_{0i}} \tag{2.25}$$

tem assintoticamente uma distribuição do qui-quadrado, com $r - 1$ graus de liberdade.

A demonstração deste teorema encontra-se em [Cramér, 1946]. O teorema 2.3 é deduzido para $N \to +\infty$ no entanto, para que a aproximação continue válida no caso finito, a frequência esperada, Np_{0i}, da variável aleatória N_i com $i \in \{1, ..., r\}$, nunca deve ser inferior a cinco [Murteira, 1990].

Caso H_0 não se encontre completamente especificada, a estatística do teste, dada pela equação 2.25, tem distribuição assintótica do qui-quadrado com $r - k - 1$ graus de liberdade, onde k representa o número de parâmetros desconhecidos estimados a partir da amostra.

2.3 Teste de Kolmogorov-Smirnov para duas amostras independentes

O teste de Kolmogorov-Smirnov (K-S) para comparação de duas amostras independentes, desenvolvido por Smirnov (1939), utiliza-se para testar se duas amostras aleatórias e independentes são provenientes de populações contínuas com a mesma distribuição.

Enquanto que o teste de ajustamento de K-S para uma amostra avalia o ajustamento a uma distribuição conhecida e compara-se a função de distribuição empírica da amostra com a função de distribuição da população (teórica), no teste de K-S para duas amostras independentes, comparam-se as funções de distribuição empíricas das duas amostras [Gibbons and Chakraborti, 2010].

De facto, se as duas amostras forem derivadas da mesma população, espera-se que as funções de distribuição empíricas sejam idênticas ou razoavelmente semelhantes. Caso exista uma diferença significativa em qualquer ponto, pode concluir-se que há uma grande possibilidade de as amostras serem provenientes de diferentes populações [Sheskin, 2003].

Sejam $(X_1, ..., X_m)$ e $(Y_1, ..., Y_n)$ duas amostras aleatórias independentes que provêm de populações contínuas com funções de distribuição desconhecidas, F_X e F_Y, respetivamente.

Considerando as estatísticas de ordem $X_{(1)}, ..., X_{(m)}$ e $Y_{(1)}, ..., Y_{(n)}$ as respetivas funções de distribuição empíricas, designadas de $S_m(x)$ e $S_n(x)$, são definidas por:

$$S_m(x) = \begin{cases} 0 & \text{se} \quad x < X_{(1)} \\ t/m & \text{se} \quad X_{(t)} \leq x < X_{(t+1)} \quad \text{para} \quad t = 1, 2, ..., m-1 \\ 1 & \text{se} \quad x \geq X_{(m)} \end{cases} \tag{2.26}$$

$$S_n(x) = \begin{cases} 0 & \text{se} \quad x < Y_{(1)} \\ t/n & \text{se} \quad Y_{(t)} \leq x < Y_{(t+1)} \quad \text{para} \quad t = 1, 2, ..., n-1 \\ 1 & \text{se} \quad x \geq Y_{(n)} \end{cases} \tag{2.27}$$

A hipótese nula que se pretende ensaiar é

$$H_0 : \; F_Y(x) = F_X(x) \;\; \textit{para todo} \;\; x \tag{2.28}$$

contra a alternativa,

$$H_1 : \; F_Y(x) \neq F_X(x) \;\; \textit{para algum} \;\; x \, . \tag{2.29}$$

Quando H_0 é verdadeira, as distribuições das populações são idênticas e conclui-se que as duas amostras pertencem à mesma população. Note-se que as funções de distribuição populacionais, $F_X(x)$ e $F_Y(x)$, podem ser estimadas a partir das respetivas funções de distribuição empíricas, $S_m(x)$ e $S_n(x)$. Assim, quando a hipótese nula é válida, espera-se que $S_m(x)$ e $S_n(x)$ estejam muito próximas para todos os valores de x. Para medir a proximidade ou o afastamento entre as funções de distribuição empíricas, define-se a estatística do teste (bilateral) de Kolomogorov-Smirnov para duas amostras, $D_{m,n}$. Esta estatística baseia-se na diferença absoluta máxima entre as duas distribuições empíricas, sendo dada por

$$D_{m,n} = \max_x |S_m(x) - S_n(x)| \, . \tag{2.30}$$

Uma vez especificado o nível de significância α, a região de rejeição é na cauda superior, definida por $D_{m,n} \geq c_\alpha$. Neste caso, considerando d o valor observado da estatística, o valor-p corresponde à probabilidade

$$P(D_{m,n} \geq d \mid H_0) \;\; . \tag{2.31}$$

Tal como acontece com a estatística do teste de K-S para uma amostra, $D_{m,n}$ não tem restrições relativamente à especificidade da distribuição comum das duas populações contínuas, apenas se exige que as $m+n$ observações estejam ordenadas.

Note-se que para implementar o teste é necessário determinar a distribuição cumulativa nula de $mnD_{m,n}$. Tal como é referido em [Gibbons and Chakraborti, 2010], existem vários métodos para calcular os valores exatos da distribuição nula. Quando a dimensão das amostras é relativamente pequena, os quantis da distribuição exacta encontram-se tabelados. No entanto, para para dimensões de amostra maiores, os valores críticos para a estatística do teste são calculados de forma aproximada. No teorema seguinte, apresenta-se o resultado obtido por Smirnov(1939) [Bogacka, 2016].

Teorema 2.4. *Sejam $X_1, ..., X_m$ e $Y_1, ..., Y_n$ duas amostras aleatórias independentes, extraídas da mesma população com função de distribuição contínua. Além disso, sejam S_m e S_n as respetivas funções de distribuição empíricas, tais que $D_{m,n} = \max_x |S_m(x) - S_n(x)|$. Então,*

$$\lim_{m,n\to\infty} P\Big(\sqrt{\frac{mn}{m+n}} D_{m,n} \leq d\Big) = L(d) \tag{2.32}$$

onde

$$L(d) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2d^2} \ . \tag{2.33}$$

A demonstração deste toerema pode ser encontrada em [Gibbons and Pratt, 1981].

Quando o teste é unilateral, por exemplo, $H_0 : F_Y(x) \leq F_X(x)$ para todo x, a estatística é dada por

$$D^+_{m,n} = \max_x |S_m(x) - S_n(x)| \ . \tag{2.34}$$

Neste caso, no teorema 2.4, tem-se

$$\lim_{m,n\to\infty} P\Big(\sqrt{\frac{mn}{m+n}} D^+_{m,n} \leq d\Big) = 1 - e^{-2d^2} \ . \tag{2.35}$$

2.4 Comparações múltiplas

Por vezes, é necessário proceder a comparações múltiplas, isto é, testar várias hipóteses nulas em simultâneo. A probabilidade de rejeitar incorretamente H_0, ou seja, obter um falso positivo, chama-se erro do tipo I e é limitada pelo nível de significância α. Por outro lado, a probabilidade de não rejeitar H_0 quando é falsa, corresponde a um erro do tipo II, denotado por β. A potência do teste é definida como $1-\beta$, que significa a probabilidade de rejeitar correctamente H_0.

Ao realizar vários testes de hipóteses é importante analisar o efeito da multiplicidade das inferências sobre o erro do tipo I e a respetiva correção. Torna-se necessário controlar o erro do tipo I a um nível de significância α, ou seja

$$Erro\ tipo\ I = P(rejeitar\ \ H_0 \mid H_0\ \ verdadeira) \leq \alpha \ . \tag{2.36}$$

Grande parte da literatura sobre métodos de correção para comparações múltiplas enfatiza e controla uma de duas taxas de erro: a taxa da família de erros dos testes (FWER, *Family-Wise Error Rate*) e a taxa de falsos positivos (FDR, *False Discovery Rate*) [Foulkes, 2009].

Neste trabalho, apenas é definida a FWER, uma vez que só foram aplicadas correções que controlam esta taxa de erro.

Considerarem-se m múltiplas hipóteses nulas dadas por $H_0^1, ..., H_0^m$. Suponha-se que das m hipóteses nulas, m_0 são verdadeiras e que R é o número total de hipóteses rejeitadas (significativas). A soma de todos os testes produz os resultados indicados na tabela 2.2 onde V é o número de erros do tipo I e T é o número de erros do tipo II.

	Teste		
	Não significativo	Significativo	Total
H_0 verdadeira	U	V	m_0
H_0 falsa	T	S	$m - m_0$
	$m - R$	R	m

Tabela 2.2: Erros do tipo I e II em testes de hipóteses múltiplos.

A FWER é definida como a probabilidade de fazer pelo menos um erro de tipo I isto é,

$$FWER = P(V \geq 1) \ . \tag{2.37}$$

Quando todas as hipóteses nulas são verdadeiras, a FWER é designada de FWEC (*Family-Wise Error under the Complete null*), correspondendo à probabilidade de rejeitar pelo menos uma dessas hipóteses nulas quando todas são verdadeiras. Ou seja,

$$FWEC = P(V \geq 1 \mid H_0^C \ verdadeira) \tag{2.38}$$

onde $H_0^C = [H_0^1, ..., H_0^m]$ é conjunto completo de todas as hipóteses nulas [Foulkes, 2009].

Suponha-se que cada teste de hipóteses é controlado ao nível de significância α. Neste caso, a probabilidade de cometer um erro do tipo I, isto é, rejeitar incorretamente a hipótese nula H_0^i, com $i = 1 \ldots m$, é dada por

$$P(rejeitar \ H_0^i \mid H_0^i \ verdadeira) \leq \alpha \ . \tag{2.39}$$

Assumindo que os testes são independentes tem-se

$$\begin{aligned} FWEC &= P(V \geq 1 \mid H_0^C \ verdadeira) \\ &= 1 - P(V = 0 \mid H_0^C \ verdadeira) \\ &= 1 - \prod_{i=1}^{m} P(nao \ rejeitar \ H_0^i \mid H_0^i \ verdadeira) \\ &= 1 - \prod_{i=1}^{m} \left[1 - P(rejeitar \ H_0^i \mid H_0^i \ verdadeira)\right] \\ &\leq 1 - \prod_{i=1}^{m} (1 - \alpha) = 1 - (1 - \alpha)^m \ . \end{aligned} \tag{2.40}$$

Quando só se realiza um único teste ($m = 1$) resulta que $FWEC \leq \alpha$. Por exemplo, para $\alpha = 0.05$, ao realizarem-se 2 testes de hipóteses independentes ($m = 2$), sabe-se que a probabilidade de cometer um erro do tipo I é menor ou igual a $1 - 0.95^2 = 0.0975$. Quando o número de testes aumenta para $m = 10$, tem-se $FWEC \leq 1 - 0.95^{10} = 0.4012$. Isso significa

que, embora se controle cada um dos 10 testes a um nível de significância $\alpha = 0.05$, no geral, a possibilidade de cometer um erro do tipo I pode ser tão grande quanto 40%. Desta forma, quantos mais testes de hipóteses forem realizados, mais provável será encontrar falsos positivos. No entanto, ao controlar de forma excessiva o erro do tipo I, aumenta-se o erro do tipo II e diminui-se o poder do teste, sendo o teste conservador.

Existem vários métodos para aplicar no caso das comparações múltiplas, dos quais se destacam neste trabalho, as correções de Bonferroni e Šidák. Estas duas correções fazem parte dos métodos de etapa única ($single - step\ adjustment$), que consiste na aplicação de um único critério para avaliar a significância de todos os valores- p. Em seguida, apresenta-se uma breve abordagem às correções mencionadas anteriormente.

2.4.1 Correção de Bonferroni

A correção de Bonferroni é provavelmente o método mais conhecido quando se realizam comparações múltiplas. É um procedimento de etapa única e tem como objetivo controlar o erro numa família de testes (FWER) e evitar a acumulação de falsos positivos que advém da realização de vários testes de hipóteses, contudo é conservativo.

Tal como é referido em [Goemana and Solari, 2012], na correção de Bonferroni, para cada teste individual da família de m testes, considera-se um nível de significância

$$\alpha' = \frac{\alpha}{m} \ . \tag{2.41}$$

Isto é, para o valor-p de cada teste (p_i) com $i = 1, ..., m$), a FWER é controlada quando se rejeita cada hipótese nula tal que

$$p_i \leq \frac{\alpha}{m} \ .$$

Quando todos os m_0(hipóteses nulas verdadeiras) testes são independentes, a probabilidade de fazer uma falsa rejeição é dada por

$$1 - \Big(1 - \frac{\alpha}{m}\Big)^{m_0} \ . \tag{2.42}$$

De facto, o método de Bonferroni é um corolário da desigualdade de Boole [Hsu, 1996]. Sejam $E_1, ..., E_k$ uma qualquer sequência de eventos. Tem-se que a fórmula de Boole é

$$P\Big(\bigcup_{i=1}^{k} E_i\Big) = \sum_{i=1}^{k} P(E_i) - \sum_{i<j} P\Big(E_i \bigcap E_j\Big) + \ldots + (-1)^{k-1} P\Big(\bigcap_{i=1}^{k} E_i\Big) \ . \tag{2.43}$$

Da equação 2.43 resulta o limite superior

$$P\Big(\bigcup_{i=1}^{k} E_i\Big) \leq \sum_{i=1}^{k} P(E_i) \ . \tag{2.44}$$

Assim, de acordo com [Goemana and Solari, 2012], considerando $q_1, ..., q_{m_0}$, com $m_0 \leq m$, os valores-p das hipóteses nulas verdadeiras, a probabilidade de existir um i tal que $q_i \leq \alpha/m$ é dada por

$$P\Big(\min_i q_i \leq \frac{\alpha}{m}\Big) = P\Big(\bigcup_{i=1}^{m_0} \Big\{q_i \leq \frac{\alpha}{m}\Big\}\Big) \leq \sum_{i=1}^{m_0} P\Big(q_i \leq \frac{\alpha}{m}\Big) \leq m_0 \frac{\alpha}{m} \leq \alpha \ . \tag{2.45}$$

2.4.2 Correção de Šidák

A correção Šidák é uma forma alternativa ao método de Bonferroni, também para controlar a FWER.

Seja α, a probabilidade de cometer pelo menos um erro do tipo I numa família de m testes independentes. Neste caso, o nível de significância α da família de testes é dado por

$$\alpha = 1 - (1 - \alpha')^m \ , \tag{2.46}$$

onde α' corresponde à probabilidade de cometer um erro do tipo I em cada teste individual. Resolvendo a equação 2.46 em ordem a α', o nível de de significância de cada teste é

$$\alpha' = 1 - (1 - \alpha)^{\frac{1}{m}} \ . \tag{2.47}$$

Tendo em conta o corolário da desigualdade de Šidák, resultado que pode ser encontrado em [Hsu, 1996], tem-se que

$$\alpha \leq 1 - (1 - \alpha')^m \, . \tag{2.48}$$

Os dois métodos de correção apresentados são muito próximos um do outro, sendo que o de Šidák é menos conservativo. Por exemplo, suponha-se que ao realizar 4 testes independentes, pretende-se controlar o erro do tipo I, considerando um nível de significância global para a família de testes $\alpha = 0.05$. Neste caso, cada hipótese individual será rejeitada, caso o valor-p da cada teste seja inferior a um valor de α', tal que

$$\alpha' = 1 - (1 - 0.05)^{\frac{1}{4}} = 0.0127 \, .$$

No caso da correção de Bonferroni, cada teste individual é controlado a um nível de significância de

$$\alpha' = \frac{0.05}{4} = 0.0125 \, .$$

2.5 Equilíbrio de Hardy-Weinberg

Devido ao avanço das técnicas de sequenciação, em muitos dos estudos genéticos modernos já é possível determinar os genótipos de um grande número de marcadores genéticos (SNVs). Desta forma, com essas informações, podem construir-se bases de dados de grandes dimensões, como é o caso dos dados disponibilizados pelo P1000G. Para esses conjuntos de dados, calculam-se as frequências dos alelos e dos genótipos, correspondentes a cada SNV, de modo que essas frequências estão sujeitas a uma restrição de soma unitária [Graffelman and Camarena, 2016].

O equilíbrio de Hardy-Weinberg (HWE), formulado de forma independente por Hardy e por Weinberg (1908), é um princípio fundamental da genética moderna e desempenha um papel importante nos *GWAS*. Na ausência de forças perturbadoras (migração, mutação, seleção, etc.) a lei de Hardy-Weinberg prevê que as frequências dos genótipos e dos alelos permanecerão no seu estado de equilíbrio ao longo das gerações. O desequilíbrio pode resultar numa atribuição errada dos genótipos, isto é, confusão entre heterozigóticos e homozigóticos. Deste modo, testar o HWE pode ajudar a detetar esse "erro". Por outro lado, o desequilíbrio em estudos de caso-controlo pode ser indicativo de associação entre um marcador genético (SNV) a uma dada doença, pelo que a verificação do HWE pode fornecer pistas sobre essa situação [Graffelman, 2015].

Para cada SNV bi-alélica, de alelos x e y, considerem-se os três genótipos xx, xy e yy, cujas contagens observadas são, repetivamente, n_{xx}, n_{xy} e n_{yy}, com $n = n_{xx} + n_{xy} + n_{yy}$. Sejam p_x e $p_y = 1 - p_x$ as frequências dos alelos x e y, respetivamente. Para se verificar a lei de Hardy-Weinberg, as frequências dos genótipos, p_{xx}, p_{xy} e p_{yy}, devem verificar

$$p_{xx} = p_x^2 \ , \quad p_{xy} = 2p_x p_y \quad e \quad p_{yy} = p_y^2 \ , \tag{2.49}$$

tal que $p_{xx} + p_{xy} + p_{yy} = 1$.

Para avaliar se uma SNV pode ser considerada em equilíbrio ou não, pode-se recorrer ao teste clássico de ajustamento do χ^2. Assim sendo, as frequências esperadas para cada genótipo, sob a hipótese de HWE, são $e_{xx} = np_x^2$, $e_{xy} = 2np_x p_y$ e $e_{yy} = np_y^2$.

A estatística do teste do χ^2 para o HWE é dada por

$$\chi^2 = \frac{(n_{xx} - e_{xx})^2}{e_{xx}} + \frac{(n_{xy} - e_{xy})^2}{e_{xy}} + \frac{(n_{yy} - e_{yy})^2}{e_{yy}} \ , \tag{2.50}$$

que tem distribuição qui-quadrado com um grau de liberdade, χ_1^2, sob a hipótese nula (HWE).

Note-se que o teste do χ^2 para a independência, aplica-se tipicamente a tabelas ou matrizes. No entanto, os dois testes são equivalentes, se o vetor com as três contagens de cada genótipo for reorganizado sob a forma da tabela 2.3.

	x	y	
x	n_{xx}	$\frac{1}{2}n_{xy}$	$\frac{1}{2}n_x$
y	$\frac{1}{2}n_{xy}$	n_{yy}	$\frac{1}{2}n_y$
	$\frac{1}{2}n_x$	$\frac{1}{2}n_y$	n

Tabela 2.3: Representação das contagens dos genótipos, n_{xx}, n_{xy} e n_{yy}, através de uma tabela de 2×2.

Multiplicando por 2, a tabela 2.3, obtém-se a tabela 2.4, na qual aparece o número total de alelos $2n$ e as contagens para cada um dos alelos, $n_x = 2n_{xx} + n_{xy}$ e $n_y = 2n_{yy} + n_{xy}$.

	x	y	
x	$2n_{xx}$	n_{xy}	n_x
y	n_{xy}	$2n_{yy}$	n_y
	n_x	n_y	$2n$

Tabela 2.4: Total de alelos, $2n$, e contagens de cada um dos alelos, n_x e n_y, representados numa tabela 2×2.

Neste caso, as frequências dos alelos x e y, são dadas por

$$p_x = \frac{n_x}{2n} = \frac{2n_{xx} + n_{xy}}{2n} \tag{2.51}$$

$$p_y = \frac{n_y}{2n} = \frac{2n_{yy} + n_{xy}}{2n} \tag{2.52}$$

Note-se que a estatística χ^2 não é suficientemente informativa sobre a natureza do desequilíbrio de Hardy-Weinberg. Assim, torna-se importante calcular o chamado coeficiente de desequilíbrio, D. Este coeficiente é definido, segundo [Weir, 1996], pelas seguintes igualdades:

$$p_{xx} = p_x^2 + D \qquad p_{xy} = 2p_xp_y - 2D \qquad e \quad p_{yy} = p_y^2 + D\,. \tag{2.53}$$

O coeficiente D indica um desvio relativamente às contagens heterozigóticas. Portanto, desequilíbrios causados por excesso ou défice de heterozigóticos, conduzem a valores de D positivos ou negativos, respetivamente. Quando o HWE válido, o coeficiente de desequilíbrio é zero. Logo, uma formulação alternativa para testar o HWE, é testar a hipótese H_0 : *Coeficiente de desequilíbrio nulo.* Assim, a estatística que foi apresentada na equação 2.50 pode ser escrita como

$$\chi^2 = \frac{nD^2}{p_x^2 p_y^2}\,. \tag{2.54}$$

Ora, substituindo D, na equação 2.50 resulta que

$$\begin{aligned}
\chi^2 &= \frac{(nD)^2}{np_x^2} + \frac{(-2nD)^2}{2np_xp_y} + \frac{(nD)^2}{np_y^2} \\
&= \frac{2p_y^2(nD)^2 + (-2nD)^2 p_xp_y + 2p_x^2(nD)^2}{2np_x^2p_y^2} \\
&= \frac{2(1-p_x)^2(nD)^2 + (-2nD)^2 p_x(1-p_x) + 2p_x^2(nD)^2}{2np_x^2p_y^2} \\
&= \frac{2(1-2p_x+p_x^2)n^2D^2 + 4n^2D^2(p_x - p_x^2) + 2p_x^2n^2D^2}{2np_x^2p_y^2} \\
&= \frac{n^2D^2 - 2p_xn^2D^2 + p_x^2n^2D^2 + 2n^2D^2p_x - 2n^2D^2p_x^2 + p_x^2n^2D^2}{np_x^2p_y^2} \\
&= \frac{nD^2}{p_x^2p_y^2}
\end{aligned}$$

Considerem-se, por exemplo, duas SNVs, $A \leftrightarrow G$ e $C \leftrightarrow G$, presentes no cromossoma 1 dos dados do P1000G. Para a variação $A \leftrightarrow G$ o vetor das contagens dos genótipos, AA, AG e GG, é dado por $(5; 304; 783)$. Ao aplicar o teste do ajustamento do χ^2 para HWE, os resultados obtidos são $\chi^2 \approx 18.660$, $g.l. = 1$, valor-$p \approx 1.562 \times 10^{-05}$ e $D \approx 17.572$, donde se rejeita a hipótese de HWE. Para a variação $C \leftrightarrow G$, as contagens dos genótipos CC, CG e GG são $(1051; 41; 0)$. Nesta SNV, aceita-se a hipótese nula de HWE pois aplicando o teste de ajustamento do χ^2, obtêm-se os resultados $\chi^2 \approx 0.400$, $g.l. = 1$, valor-$p \approx 0.527$ e $D \approx 0.385$. Assim, conclui-se que para $A \leftrightarrow G$, existe um excesso de heterozigóticos ($D > 0$) e que para $C \leftrightarrow G$ existe um desequilíbrio negligenciável entre as contagens homozigóticas e heterozigóticas ($D \approx 0$).

A relação entre as frequências dos genótipos de cada SNV, pode ser explorada a partir de diagramas de dispersão, nos quais é representada uma curva que traduz o HWE. Uma formulação alternativa à lei de Hardy-Weinberg, definida pelas igualdades da equação 2.49, pode ser obtida ao fazer o quadrado da frequência heterozigótica, ou seja,

$$p_{xy}^2 = 4p_{xx}p_{yy}. \tag{2.55}$$

Da equação 2.55, resulta a equação da curva que representa a relação entre as frequências heterozigóticas (p_{xy}) *versus* homozigóticas (p_{xx}), dada por

$$p_{xy} = 2(\sqrt{p_{xx}} - p_{xx}) \ . \tag{2.56}$$

No caso da relação entre as frequências homozigóticas (p_{yy}) *versus* homozigóticas (p_{xx}), a equação da curva correspondente é

$$p_{yy} = (1 - \sqrt{p_{xx}})^2 \, . \tag{2.57}$$

A figura 2.1 exemplifica os diagramas de dispersão das frequências heterozigóticas/ homozigóticas e homozigóticas/homozigóticas, para as seis primeiras SNVs no cromossoma 1.

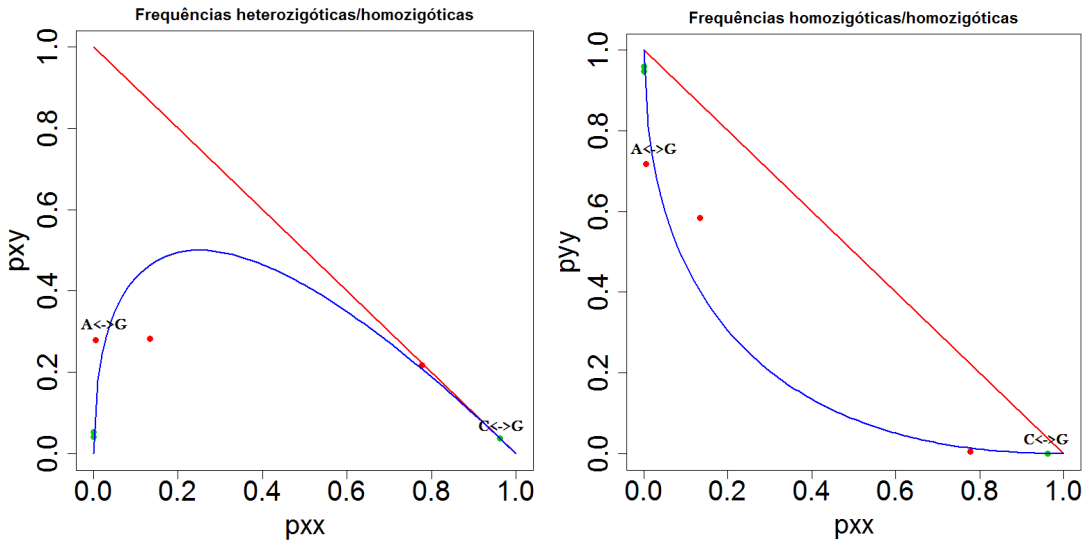


Figura 2.1: Diagrama de dispersão das frequências p_{xy}/p_{xx} e p_{yy}/p_{xx} correspondentes às seis primeiras SNVs do cromossoma 1 dos dados do P1000G.

Considerando o valor crítico do teste pré-especificado, $\chi_1^2(\alpha)$, para a estatística χ^2, é possível expressar-se a frequência heterozigótica, p_{xy}, em função das frequências alélicas p_x e $p_y = 1 - p_x$, obtendo-se a equação de duas parábolas. Ou seja,

$$p_{xy} = 2p_xp_y \pm 2p_xp_y\sqrt{\chi_1^2(\alpha)/n} \ . \tag{2.58}$$

De facto, por 2.53, como $-2D = p_{xy} - 2p_xp_y \Leftrightarrow D^2 = \frac{(p_{xy}-2p_xp_y)^2}{4}$, substituindo D^2 na equação 2.54, resulta que

$$4\chi_1^2(\alpha)p_x^2p_y^2 = n(p_{xy} - 2p_xp_y)^2 \Leftrightarrow 4\chi_1^2(\alpha)p_x^2p_y^2 = np_{xy}^2 - 4np_xp_yp_{xy} + 4np_x^2p_y^2 \Leftrightarrow$$

$$\Leftrightarrow np_{xy}^2 - 4np_xp_yp_{xy} + 4np_x^2p_y^2 - 4\chi_1^2(\alpha)p_x^2p_y^2 = 0\,.$$

Resolvendo a equação quadrática em ordem a p_{xy}, tem-se que

$$\begin{aligned}
p_{xy} &= \frac{4np_xp_y \pm \sqrt{(4np_xp_y)^2 - 4n(4np_x^2p_y^2 - 4\chi_1^2(\alpha)p_x^2p_y^2)}}{2n} \\
&= \frac{4np_xp_y \pm \sqrt{16n^2p_x^2p_y^2 - 16n^2p_x^2p_y^2 + 16n\chi_1^2(\alpha)p_x^2p_y^2}}{2n} \\
&= \frac{2np_xp_y \pm 2p_xp_y\sqrt{n\chi_1^2(\alpha)}}{n} \\
&= 2p_xp_y \pm 2p_xp_y\sqrt{\chi_1^2(\alpha)/n}\ .
\end{aligned}$$

Quando $\chi^2 = 0$ a frequência heterozigótica é igual a $p_{xy} = 2p_xp_y$, ou seja, verifica-se o HWE. A partir da equação 2.58 conclui-se que a hipótese de HWE será rejeitada, sempre que a frequência de heterozigóticos for muito grande ou muito pequena. Assim, a região de aceitação para o HWE é dada por

$$2p_xp_y - 2p_xp_y\sqrt{\chi_1^2(\alpha)/n} \leq p_{xy} \leq 2p_xp_y + 2p_xp_y\sqrt{\chi_1^2(\alpha)/n} \ . \tag{2.59}$$

Tem-se que o limite superior e o limite inferior da região de aceitação para o HWE, apresentados na equação 2.59, são equações quadráticas em p_y, por exemplo, que podem ser representadas num diagrama ternário (*ternary plot*).

Este tipo de gráfico, em que cada vértice do triângulo representa um dos genótipos possíveis para uma dada SNV, é útil para inferir se o HWE é verificado ou não. Normalmente, o vértice superior está associado ao genótipo heterozigótico e os dois vértices inferiores aos genótipos homozigóticos. Ao representar várias amostras para vários tipos de SNV, apesar de os vértices não ficarem especificados para os genótipos de cada SNV, o gráfico permanece informativo pois é possível visualizar quais as SNVs cujas frequências alélicas e genotípicas estão de acordo com o equilíbrio. A figura 2.2 exemplifica o *ternary plot*, no qual estão representas as seis primeiras SNVs do cromossoma 1. Os pontos verdes correspondem às SNVs que ficaram dentro da região de aceitação para o HWE e os pontos vermelhos correspondem aos casos significativos, isto é, as variações que não verificaram o HWE. Note-se que as SNVs com $D > 0$ (excesso heterozigótico) situam-se acima da região de aceitação e as que têm $D < 0$ (escassez heterozigótica) abaixo. Na situação representada, observam-se 3 SNVs significativas

(rejeição do HWE), $A \leftrightarrow G$, $C \leftrightarrow T$ e $G \leftrightarrow T$. Tem-se ainda que para a variação $A \leftrightarrow G$, as frequências dos alelos A e G são $p_A \approx 0.14$ e $p_G \approx 0.86$, respetivamente e que as frequências dos genótipos são $p_{AA} \approx 0.02$, $p_{AG} = 0.24$ e $p_{GG} \approx 0.74$. No *ternary plot*, está destacada a marcação da frequência do alelo G, no eixo p_y (no caso particular desta SNV corresponde a p_G). Observa-se também que, quanto menor for o valor da frequência do genótipo xx, mais próximo do lado de vértices xy e yy, será marcado o ponto correspondente à SNV.

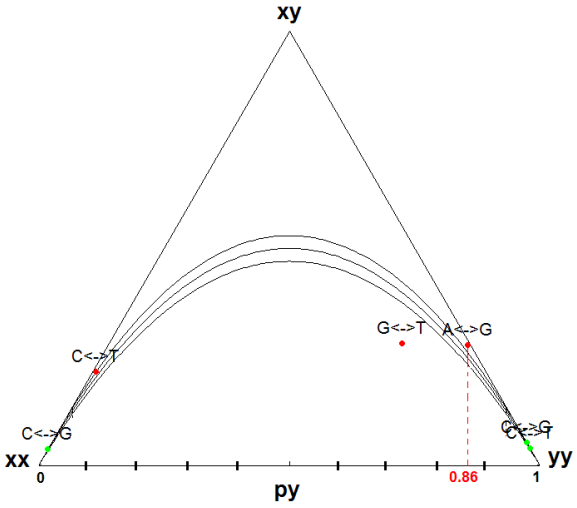


Figura 2.2: *Ternary plot* das 6 primeiras SNVs do cromossoma 1 e respetiva região de aceitação para o HWE.

Note-se que para investigar se uma SNV verifica ou não o HWE, também se pode recorrer a procedimentos exatos ou ao teste de ajustamento do χ^2 com correção à continuidade de Yates. Contudo, os procedimentos exatos são computacionalmente intensivos, especialmente para grandes amostras [Graffelman, 2015]. Por outro lado, quando se aplica a correção à continuidade de Yates, para frequências muito baixas do alelo menor, a correção pode levar a taxas excessivas de erro tipo I [Graffelman, 2016]. Portanto, tendo em conta a grande dimensão dos dados do P1000G, neste trabalho optou-se por testar o HWE recorrendo ao teste clássico de ajustamento do χ^2, sem aplicar a correção à continuidade.

2.6 Classificação Hierárquica

Neste trabalho, há interesse em averiguar se, por exemplo, as ocorrências dos diferentes tipos de variação (SNV) manifestam semelhanças entre si em função dos cromossomas, grupos de prevalência ou contexto na vizinhança. Pretende-se agrupar as várias variáveis, de modo a fazer uma classificação. Deste modo, será pertinente proceder a uma análise classificatória.

Em [Reis, 2001], a classificação hierárquica (ou análise de *clusters*) é descrita da seguinte forma: dado um conjunto de n indivíduos para os quais existe informação sobre a forma de p variáveis, o método procede ao agrupamento das variáveis em função da informação existente, de tal modo que os indivíduos pertencentes a um mesmo grupo sejam tão semelhantes quanto possível e sempre mais semelhantes aos elementos do mesmo grupo do que a elementos dos restantes grupos. Assim, a classificação hierárquica é um termo genérico para uma ampla gama de métodos com o objetivo comum de descobrir classes de observações que são

homogéneas entre si e diferentes das outras classes, produzindo uma classificação hierárquica dos dados [Everitt and Hothorn, 2011].

O problema surge pelo facto de não existir um único critério de partição e/ou agrupamento das observações com base numa única medida de (dis)semelhança. Das várias etapas que envolvem a classificação hierárquica é fundamental fazer duas escolhas para classificar hierarquicamente os dados: a medida de semelhança entre as observações e o critério de agregação entre grupos. A utilização de diferentes medidas de semelhança e/ou diferentes critérios de agregação podem levar a resultados distintos [Reis, 2001].

Numa classificação hierárquica, os dados não são divididos em grupos numa única etapa. Em vez disso, a classificação consiste numa série de partições que podem ser executadas a partir de um único grupo (*cluster*) contendo todas as observações, para n grupos contendo cada, uma única observação. As técnicas de agrupamento hierárquico podem ser subdivididas em métodos aglomerativos e divisívos. Os métodos aglomerativos são provavelmente os mais utilizados dos métodos hierárquicos e produzem partições por uma série de fusões sucessivas das n observações nos grupos. Essas fusões, uma vez feitas, são irreversíveis, de modo que quando são colocadas duas observações no mesmo grupo estas já não podem aparecer posteriormente em grupos diferentes [Everitt and et al., 2011].

Assim, sejam $P_n, P_{n-1}, ..., P_1$, uma série de partições dos dados em que a primeira, P_n, consiste em n agrupamentos de uma só observação e a última, P_1, consiste num único grupo contendo todas as n observações. Considerem-se também os grupos $C_1, C_2, ..., C_n$, cada um contendo uma única observação. O algoritmo dos métodos hierárquicos aglomerativos, descrito em [Everitt and Hothorn, 2011], é dado por:

(1) Encontrar o par mais próximo de grupos distintos, isto é, C_i e C_j, fundindo C_i e eliminando C_j, diminuindo o número de grupos numa unidade.

(2) Se o número de grupos é igual a um então parar o processo; caso contrário voltar a (1).

No entanto, para aplicar o algoritmo é necessário calcular uma matriz de distâncias entre as observações (inter-individual) ou matriz de similaridade. Como existem várias formas de calcular distâncias ou semelhanças entre pares de observações, neste trabalho, apresentar-se-á apenas a medida de distância mais utilizada, que é a distância Euclidiana. Sendo d_{ij} a distância entre a observação i, para as q variáveis $x_{i1}, x_{i2}, ..., x_{iq}$, e a observação j, para as mesmas variáveis $x_{j1}, x_{j2}, ..., x_{jq}$, pode calcular-se a distância Euclideana através de

$$d_{ij} = \sqrt{\sum_{k=1}^{q} (x_{ik} - x_{jk})^2} \ . \tag{2.60}$$

As distâncias euclidianas entre cada par de observações (i, j) podem ser dispostas numa matriz com zeros na diagonal principal e simétrica, uma vez que $d_{ij} = d_{ji}$. Note-se que esta medida de semelhança pode não ser adequada quando as variáveis estão em escalas muito diferentes. Dada a matriz de similaridade, o agrupamento hierárquico começa e em cada etapa do processo são fundidos os grupos de observações, formados anteriormente, que estão mais próximos (ou mais semelhantes). Assim, é necessário calcular a distância entre uma observação e um grupo (contendo várias observações) e a distância entre dois grupos de observações [Everitt and Hothorn, 2011].

Para calcular a distância entre grupos podem usar-se vários critérios de agregação de acordo com a distância considerada, como por exemplo: ligação simples, ligação completa,

ligação da média, ligação do centróide e método de Ward. No critério de ligação completa ou critério do vizinho mais afastado, a distância entre dois grupos é definida como a distância entre os seus elementos mais afastados ou menos semelhantes. Neste caso, dados dois grupos A e B, a distância entre eles é a maior das distâncias entre os seus elementos ou seja,

$$d_{AB} = \max_{i \in A,\ j \in B} (d_{ij}) \ . \tag{2.61}$$

As classificações hierárquicas podem ser representadas por um diagrama bidimensional conhecido como dendrograma ou árvore de agrupamento, permitindo visualizar as fusões feitas em cada etapa da análise. Na figura 2.3 pode observa-se um exemplo de um dendrograma. Os nós ou nodos representam os grupos e o comprimentos dos "ramos" representam as distâncias.

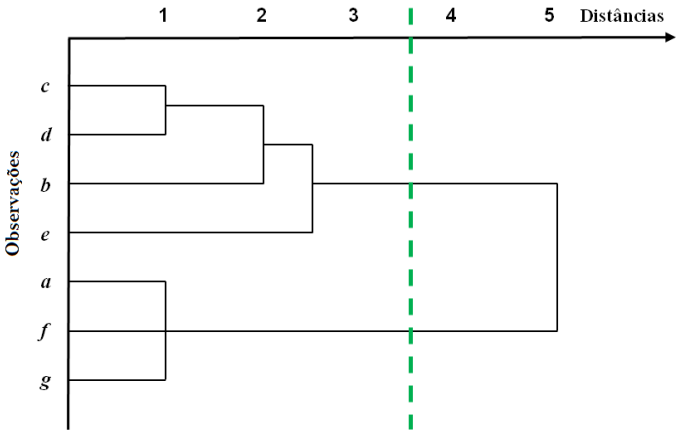


Figura 2.3: Exemplificação de um dendrograma. Adaptado de [Reis, 2001].

Uma vez que a classificação hierárquica tem como objetivo a formação de grupos homogéneos, surge o problema relativo à escolha do número adequado de grupos. A observação do dendrograma pode sugerir uma estimativa para o número de grupos. A determinação do número de grupos é feita, por exemplo, cortando verticalmente o dendrograma, tal como exemplifica a figura 2.3. O corte do dendrograma a uma distância de aproximadamente 4 sugere a existência de dois grupos $\{c, d, b, e\}$ e $\{a, f, g\}$. Este método de escolha do número de grupos não é considerado satisfatório devido à sua subjetividade e ser enviesado pela necessidade de conhecimento prévio quanto à correta estrutura dos dados. O procedimento mais usual para contornar esta situação é a utilização de vários critérios de agrupamento e a comparação posterior dos resultados obtidos, para averiguar se são semelhantes entre si [Reis, 2001].

Nesta dissertação, para a construção dos dendrogramas utilizaram-se as definições padrão do *software* R, em que o método de agregação escolhido é o de ligação completa e a medida de semelhança é a distância Euclideana.

Capítulo 3

Análise das variações de nucleótido único

Neste capítulo, apresentam-se os resultados de uma análise global das SNVs por cromossoma e por grupo de prevalência e a verificação do equilíbrio de Hardy-Weinberg nos dados.

Para realizar a análise dos dados foi necessário desenvolver algumas funções em linguagem R, as quais se encontram no apêndice A com a respetiva descrição.

Após o pré-processamento dos ficheiros `VCF`, exemplificado pela figura 1.7, os respetivos *outputs* foram importados para o ambiente do *software* R. A leitura e normalização foi feita através das funções `get.file` (função B.1) e `norm.file`(função B.2), obtendo-se os dados que são apresentados na figura 3.1.

	CHROM	POS	var	Cxx	Cxy	Cyy	Px	Py
1	1	10583	A<->G	5	304	783	0.14377289	0.85622711
2	1	10611	C<->G	1051	41	0	0.98122711	0.01877289
3	1	13302	C<->T	849	237	6	0.88598901	0.11401099
4	1	13327	C<->G	0	59	1033	0.02701465	0.97298535
5	1	13980	C<->T	0	45	1047	0.02060440	0.97939560
6	1	30923	G<->T	146	308	638	0.27472527	0.72527473

Figura 3.1: Dados normalizados correspondentes ao cromossoma 1.

A normalização dos dados consistiu em classificar o par (`REF`, `ALT`) num dos seis tipos de SNV ($A \leftrightarrow C, A \leftrightarrow G, A \leftrightarrow T, C \leftrightarrow G, C \leftrightarrow T, G \leftrightarrow T$). Ou seja, para $x, y \in \{A, C, G, T\}$ onde $x < y$ (na ordem lexicográfica), os casos em que `REF`$= x$ e `ALT`$= y$ ou `REF`$= y$ e `ALT`$= x$ são representados pelo mesmo tipo de variação $x \leftrightarrow y$.

As colunas C_{xx} e C_{yy} contabilizam os indivíduos que na posição indicada são homozigóticos com o alelo x e homozigóticos com o alelo y, respetivamente. A coluna C_{xy} contabiliza o número de indivíduos heterozigóticos. A contagem de heterozigóticos foi obtida pela fusão das contagens dos dados do `VCF`, $C_{xy} =$ `C0|1:`$+$`C1|0:`. As contagens de homozigóticos são dadas por $(C_{xx}, C_{yy}) =$ (`C0|0:`, `C1|1:`), se `REF`$= x$ e `ALT`$= y$; ou $(C_{xx}, C_{yy}) =$ (`C1|1:`, `C0|0:`), se `REF`$= y$ e `ALT`$= x$.

Note-se que ao longo deste trabalho, optou-se por excluir da análise os cromossomas sexuais X e Y. Deste modo, consideraram-se aproximadamente 36.8 milhões das 38 milhões de SNVs, reportadas nos dados da fase 1 do P1000G.

3.1 Análise por cromossoma

Considerando os 22 autossomas, contabilizou-se o número de ocorrências para cada tipo de SNV, correspondendo a um total de 36 820 992 ($\approx$ 36.8 milhões) ocorrências. Os resultados das contagens por cromossoma encontram-se em apêndice na tabela A.1 e apresentam-se resumidos na tabela 3.1 e nas caixas de bigodes representadas na figura 3.2.

Verifica-se que as variações $A \leftrightarrow G$ e $C \leftrightarrow T$ são as que mais ocorrem no genoma e que cada uma corresponde, aproximadamente, a um terço dos dados. Cada um dos outros tipos de variação ($A \leftrightarrow C$, $A \leftrightarrow T$, $C \leftrightarrow G$, $G \leftrightarrow T$) ocorre menos de 10%. Confirma-se que o número de transições é superior ao de transversões sendo o rácio igual a $T_s/T_v = 2.16$. Tem-se ainda que a frequência de ocorrência de $A \leftrightarrow C$ é semelhante à de $G \leftrightarrow T$ e que a de $A \leftrightarrow G$ é semelhante à de $C \leftrightarrow T$, refletindo a simetria complementar do ADN.

SNV						
$A \leftrightarrow C$	$A \leftrightarrow G$	$A \leftrightarrow T$	$C \leftrightarrow G$	$C \leftrightarrow T$	$G \leftrightarrow T$	Total
3004495	12606476	2515150	3143318	12563696	2987857	36 820 992
8.16%	34.24%	6.83%	8.54%	34.12%	8.11%	100%

Tabela 3.1: Número total de ocorrências para cada tipo de SNV.

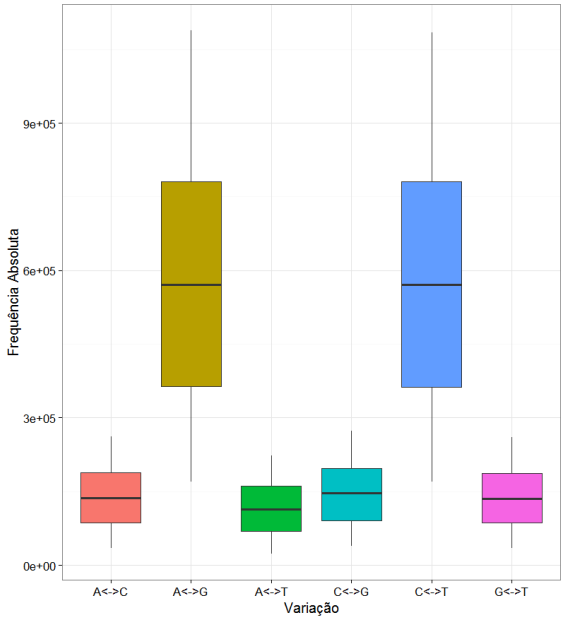


Figura 3.2: Caixas de bigodes das ocorrências por cromossoma, para cada tipo de SNV.

Na figura 3.3 apresenta-se um gráfico de barras com as frequências relativas de cada tipo de variação por cromossoma, a partir do qual se observa uma aparente homogeneidade na distribuição das SNVs por cromossoma.

Para avaliar se a ocorrência das SNVs ao longo dos cromossomas é homogénea, aplicou-se o teste de homogeneidade ($\chi^2 = 52109$, $g.l. = 105$, valor-$p \approx 0.000$).

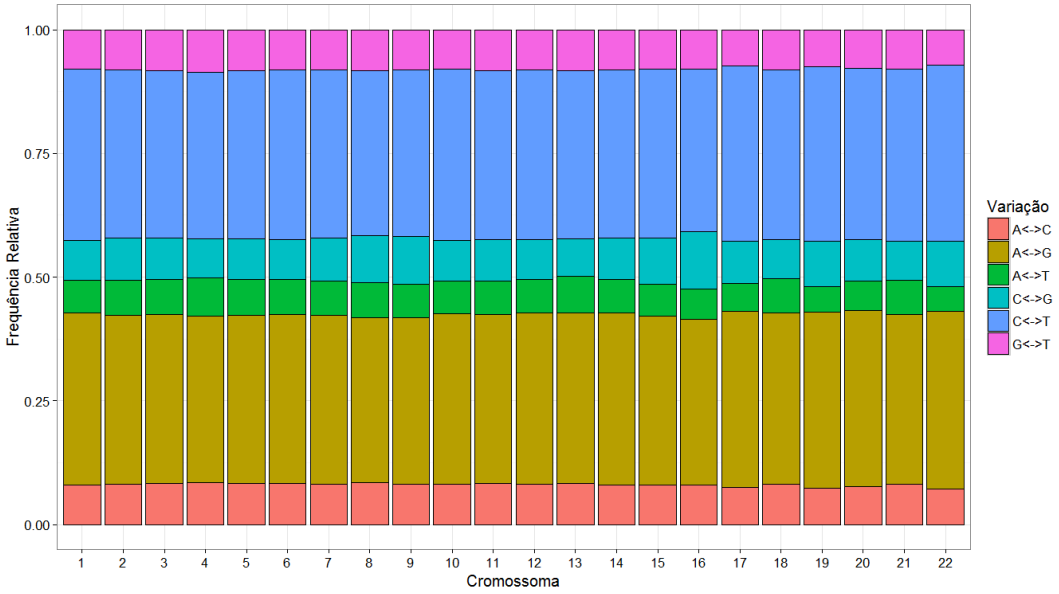


Figura 3.3: Gráfico de barras das frequências relativas de cada tipo de SNV por cromossoma.

Os valores obtidos levam à rejeição da hipótese nula o que pressupõe diferenças entre a forma como se distribuem as SNVs pelos cromossomas. Note-se que como o teste é sensível a dados de grandes dimensões é importante determinar medidas do tamanho do efeito (*effect size*). Ao calcular, por exemplo, o valor de ϕ (0.0376) conclui-se que o desajuste entre os cromossomas é negligenciável, indo de encontro à aparente homogeneidade da figura 3.3.

No entanto, como H_0 foi rejeitada, foi feita a respetiva análise de resíduos ajustados, cujos valores se encontram em apêndice na tabela A.2, uma vez que estes podem ser úteis para identificar quais as células $(cromossoma_i, SNV_j)$ que se opõem à homogeneidade. Por exemplo, os valores positivos dos resíduos informam as SNVs preferidas e os valores negativos informam as que são preteridas. Assim, foi construído um *heatmap* dos resíduos ajustados em que a escala de cor apresentada varia de vermelho a verde de acordo com a ordem de grandeza dos valores dos resíduos. Para valores negativos, nulos e positivos as cores correspondentes são vermelho, amarelo e verde, respetivamente. Pela observação da tabela A.2 e do *heatmap* 3.4 verifica-se que são as transversões $C \leftrightarrow G$ e $A \leftrightarrow T$ que têm os maiores resíduos ajustados em valor absoluto. Observa-se que os cromossomas 8, 9 e 16 apresentam preferência pela ocorrência da SNV $C \leftrightarrow G$, sendo mais evidente no cromossoma 16. No entanto esta variação é preterida pelos cromossomas 4, 6 e 13. No que diz respeito à SNV $A \leftrightarrow T$, constata-se que é a preferida do cromossoma 4 mas preterida pelos cromossomas 17, 19 e 22. Observa-se ainda que as transversões $A \leftrightarrow C$ e $G \leftrightarrow T$, apresentam um comportamento muito semelhante entre si, assim como as transições $C \leftrightarrow T$ e $A \leftrightarrow G$. Note-se que tanto as variações como os cromossomas foram agregados em 3 grupos através de dendrogramas. Relativamente às SNVs, os grupos são $\{A \leftrightarrow T, A \leftrightarrow C, G \leftrightarrow T\}$, $\{C \leftrightarrow T, A \leftrightarrow G\}$ e $\{C \leftrightarrow G\}$, ficando as transições agrupadas, algo que não aconteceu nas transversões, devido ao comportamento da variação $C \leftrightarrow G$. No histograma dos resíduos ajustados, 3.5, cuja média e o desvio padrão são -0.4713 e 21.705 respetivamente, verifica-se que não existe um ajustamento à distribuição *N(0,1)*, indo de encontro à rejeição da hipótese nula de homogeneidade entre os cromossomas relativamente à ocorrência das SNVs.

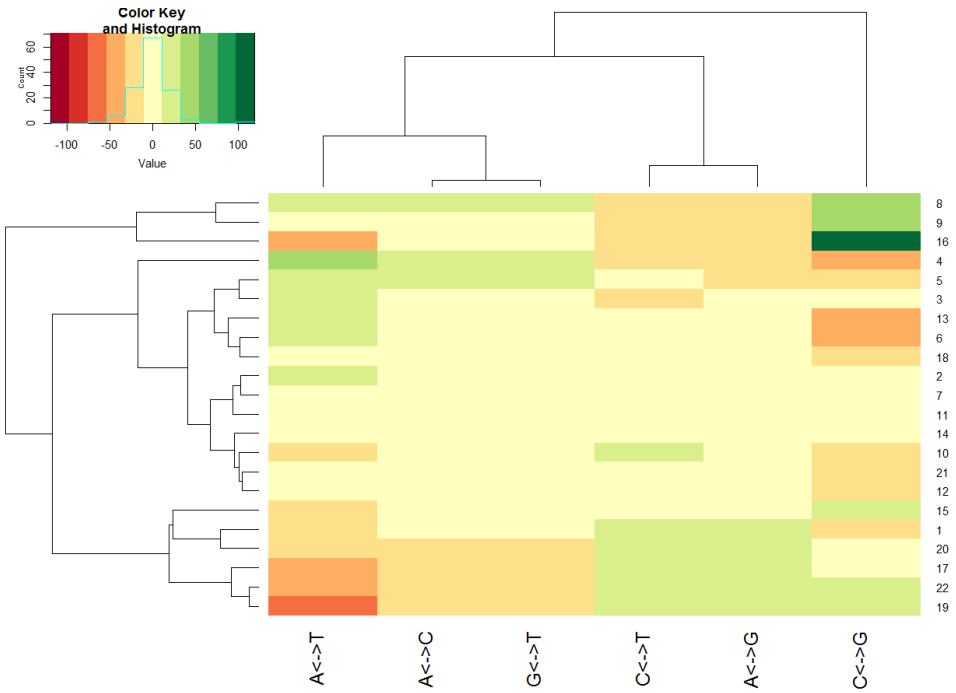


Figura 3.4: *Heatmap* dos resíduos ajustados do teste de homogeneidade entre cromossomas em relação à ocorrência das SNVs.

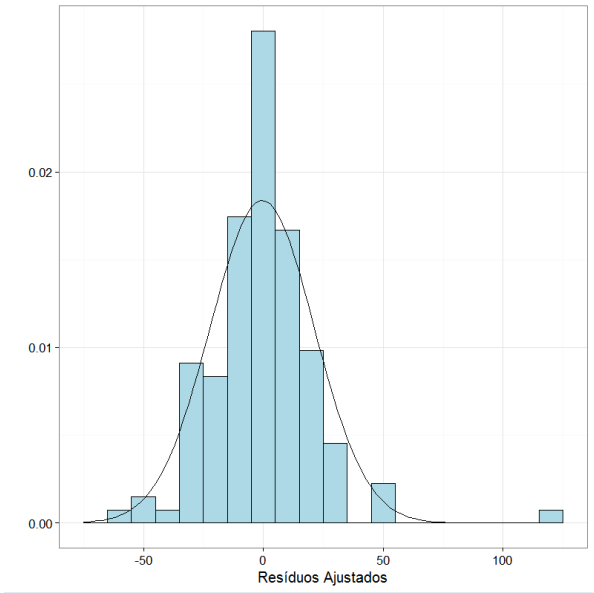


Figura 3.5: Histograma dos resíduos ajustados do teste de homogeneidade entre cromossomas em relação à ocorrência das SNVs.

3.2 Análise por grupo de prevalência

Para cada posição do genoma dos 1092 indivíduos, na qual se registou uma variação, $x \leftrightarrow y$, com $x, y \in \{A, C, G, T\}$, foi calculada a prevalência dos nucleótidos x e y, P_x e P_y, respetivamente, através das expressões

$$P_x = \frac{2C_{xx} + C_{xy}}{2 \times 1092} \quad e \quad P_y = \frac{2C_{yy} + C_{xy}}{2 \times 1092} \tag{3.1}$$

tal que $0 \leq P_x, P_y \leq 1$. Note-se que as prevalências calculadas para cada SNV bi-alélica, correspondentem às frequências dos alelos x e y, que foram apresentadas anteriormente na secção 2.5. Para além disso, em cada posição do genoma, também foi determinada a prevalência da variação $x \leftrightarrow y$, $P_{x \leftrightarrow y}$, definida como o mínimo das prevalências P_x e P_y, isto é,

$$P_{x \leftrightarrow y} = min\{P_x, P_y\} \tag{3.2}$$

onde $0 \leq P_{x \leftrightarrow y} \leq 0.5$. Para a variável prevalência da variação, $P_{x \leftrightarrow y}$, elaborou-se o histograma 3.6 e determinaram-se os respetivos quartis: $Q_1 = 0.0009$, $Q_2 = 0.0027$ e $Q_3 = 0.0188$. Verifica-se que a maioria das variações, aproximadamente 75.5%, têm uma prevalência muito baixa ($P_{x \leftrightarrow y} \leq 0.02$). Deste modo, as SNVs analisadas são, na sua grade maioria, variações raras ou de baixa frequência pois apresentam uma frequência do alelo menor $< 2\%$. Constatou-se também que, 172000 (0.47%) das SNVs têm prevalência igual a 0. Estes casos podem corresponder a posições do genoma onde se sabe que ocorre uma dada variação, mas a qual não foi observada na amostra do genoma dos 1092 indivíduos.

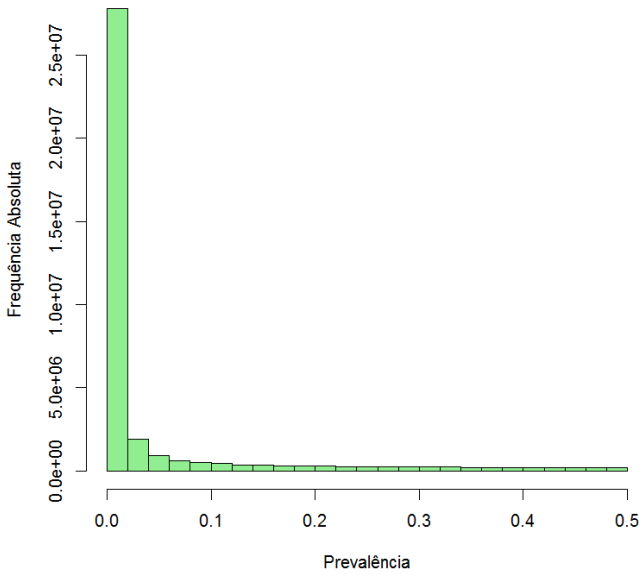


Figura 3.6: Histograma da variável prevalência $P_{x \leftrightarrow y}$.

Na figura 3.7 apresentam-se as caixas de bigodes das prevalências para cada tipo de SNV, usando uma escala semilogarítmica. Observa-se que não existem diferenças significativas na forma como são distribuídas as prevalências ao longo de cada uma das variações.

Em seguida, para averiguar se as prevalências de cada variação, provêm de populações com a mesma distribuição, aplicou-se o teste de Kolmogorov-Smirnov para duas amostras

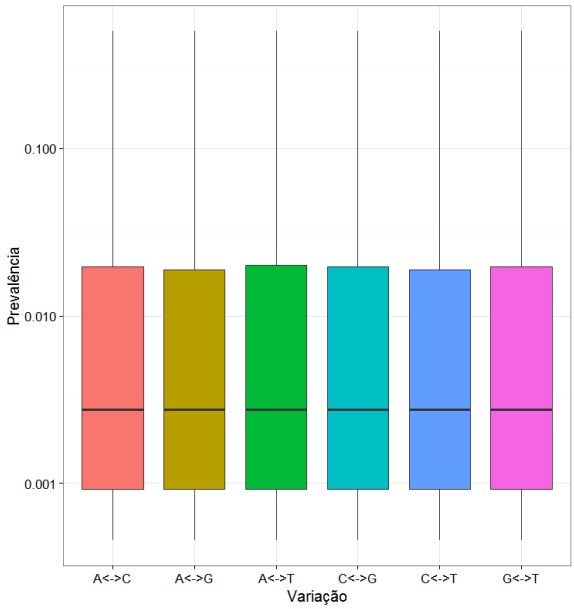


Figura 3.7: Caixas de bigodes das prevalências de cada tipo de variação.

independentes. Como existem seis tipos de variação, procedeu-se a comparações múltiplas duas a duas, donde o número de testes a efetuar é $C_2^6 = 15$. As estatísticas D_{KS} e respetivos valores-p corrigidos para cada teste, encontram-se na tabela 3.2. Como a realização de comparações múltiplas aumenta a probabilidade de rejeitar incorretamente a hipótese nula (cometer pelo menos um erro do tipo I), procedeu-se à correção dos valores-p através do método de Šidák. Para isso recorreu-se ao *package* multtest do Bioconductor e utilizou-se a função `mt.rawp2adjp`. Ao introduzir um vetor de valores-p não corrigidos, esta função corrige esses valores-p, para vários métodos com um forte controlo da FWER como por exemplo, o de Bonferroni e Šidák [Pollard et al., 2016]. Após a correção os valores-p continuam muito próximos de zero, donde se rejeita a hipótese nula em todos os casos. No entanto, verifica-se que as estatísticas D_{KS} são muito pequenas o que significa, que a distância entre as funções de distribuição das duas populações comparadas é pequena.

		valores-p					
		$P_{A\leftrightarrow C}$	$P_{A\leftrightarrow G}$	$P_{A\leftrightarrow T}$	$P_{C\leftrightarrow G}$	$P_{C\leftrightarrow T}$	$P_{G\leftrightarrow T}$
D_{KS}	$P_{A\leftrightarrow C}$		0.0000	0.0000	0.0000	0.0000	0.0000
	$P_{A\leftrightarrow G}$	0.0094		0.0000	0.0000	0.0000	0.0000
	$P_{A\leftrightarrow T}$	0.0195	0.0274		0.0000	0.0000	0.0000
	$P_{C\leftrightarrow G}$	0.0182	0.0261	0.0036		0.0000	0.0000
	$P_{C\leftrightarrow T}$	0.0354	0.0433	0.0159	0.0171		0.0000
	$P_{G\leftrightarrow T}$	0.0342	0.0421	0.0147	0.0159	0.0068	

Tabela 3.2: Estatísticas D_{KS} e valores-p corrigidos pelo método de Šidák, resultantes das comparações múltiplas entre as prevalências de cada tipo de SNV.

Tal como é referido em [1000Genomes, 2012], a maioria das variações comuns na população humana (prevalência $> 5\%$) foram descobertas na fase piloto do projeto 1000G. Contudo, as variações raras e de baixa frequência permanecem pouco caracterizadas. Estas variações têm particularmente interesse pois podem corresponder a mutações em locais potencialmente funcionais provocando, por exemplo, mudança na codificação de proteínas.

Assim, para cada uma das variações $x \leftrightarrow y$ com $x, y \in \{A, C, G, T\}$, estratificaram-se os dados em oito grupos de prevalência definidos, por exemplo, à custa da prevalência do nucleótido x, P_x. Os valores baixos da prevalência P_x (designados com "$-$") significam que para a variação $x \leftrightarrow y$ o nucleótido mais favorecido é o y e os valores altos da prevalência P_x (designados com "+") indicam que x é o nucleótido mais favorecido. Note-se que quando $P_{x \leftrightarrow y} = 0.5$ significa que os nucleótidos x e y são igualmente prevalentes. Deste modo, os oito grupos de prevalência são:

- Variação muito rara$-$ $P_x \leq 0.001$
- Variação rara$-$ $0.001 < P_x \leq 0.005$
- Variação de baixa frequência$-$ $0.005 < P_x \leq 0.05$
- Variação comum$-$ $0.05 < P_x \leq 0.5$
- Variação comum+ $0.5 < P_x \leq 0.95$
- Variação de baixa frequência+ $0.95 < P_x \leq 0.995$
- Variação rara+ $0.995 < P_x \leq 0.999$
- Variação muito rara+ $P_x > 0.999$

Para estratificar os dados de acordo com os oito grupos de prevalência recorreu-se à função `ALL.var.Gr.Px` (função B.3).

O número de ocorrências e as frequências relativas de cada SNV por grupo de prevalência são apresentados na tabela 3.3 e no gráfico de barras da figura 3.8. De acordo com os valores apresentados, verifica-se que a maioria ($\approx 82.06\%$) das SNVs analisadas correspondem a variações muito raras, raras ou de baixa frequência. As SNVs comuns (prevalência $> 5\%$) são apenas $\approx 17.94\%$. Observa-se que nos grupos das variações raras $-$ a SNV que mais ocorre é a $A \leftrightarrow G$ e nos grupos das variações raras $+$ a $C \leftrightarrow T$ é a mais frequente. Pela observação do gráfico, parece existir algum tipo de associação entre o tipo de variação e o grupo de prevalência.

Para averiguar se existe ou não associação entre o tipo de SNV e o grupo de prevalência, aplicou-se o teste de independência do χ^2. Verifica-se que a hipótese nula de independência é rejeitada ($\chi^2 = 1129352$, $g.l. = 35$, valor-$p \approx 0.000$), donde se confirma que existe associação significativa entre as duas variáveis categóricas. Foi ainda calculado o valor de ϕ para medir a força de associação, cujo valor é 0.1751, concluindo-se que a força do efeito é pequena, mas não negligenciável.

Como H_0 foi rejeitada, analisaram-se os resíduos ajustados do teste de independência, cujos valores se encontram na tabela A.3, sendo a média e o desvio padrão -0.0316 e 202.478 respetivamente. Construiu-se o respetivo *heatmap*, representado pela figura A.1, a partir do qual se confirma que os grupos das variações raras $-$ têm preferência pela variação $A \leftrightarrow G$ preterindo a variação $C \leftrightarrow T$, ao contrário do que acontece nos grupos das variações raras $+$.

Os dendrogramas fazem uma clara divisão em dois grupos de prevalência (+ e −). As SNVs foram agrupadas em três grupos, a variação $A \leftrightarrow G$, a variação $C \leftrightarrow T$ e as transversões todas.

	SNV						
Grupo	$A \leftrightarrow C$	$A \leftrightarrow G$	$A \leftrightarrow T$	$C \leftrightarrow G$	$C \leftrightarrow T$	$G \leftrightarrow T$	Total
Muito rara−	582140	2495638	427266	526778	1646614	423887	6102323
	9.54 %	40.90 %	7.00 %	8.63 %	26.98%	6.95%	16.57%
Rara−	442608	2010530	310892	391672	1254170	315222	4725094
	9.37 %	42.55 %	6.58 %	8.29%	26.54%	6.67%	12.83%
Baixa freq−	393119	1792156	289629	363042	1156502	299035	4293483
	9.16%	41.74%	6.75%	8.46%	26.94%	6.96%	11.66%
Comum−	290803	1238575	230472	285365	1005269	254597	3305081
	8.80%	37.47%	6.97%	8.63%	30.42%	7.70%	8.98%
Comum+	254983	1003525	230877	285834	1234853	289571	3299643
	7.73%	30.41%	7.00%	8.66%	37.42%	8.78%	8.96%
Baixa freq+	299039	1158125	287752	365341	1783244	388369	4281870
	6.98%	27.05%	6.72%	8.53%	41.65%	9.07%	11.63%
Rara+	317284	1257174	309872	394039	1995572	438507	4712448
	6.73%	26.68%	6.58 %	8.36%	42.35%	9.31%	12.80%
Muito Rara+	424519	1650753	428390	531247	2487472	578669	6101050
	6.96%	27.06%	7.02%	8.71%	40.77%	9.48%	16.57%
							36 820 992

Tabela 3.3: Número de ocorrências e frequências relativas das SNVs por grupo de prevalência.

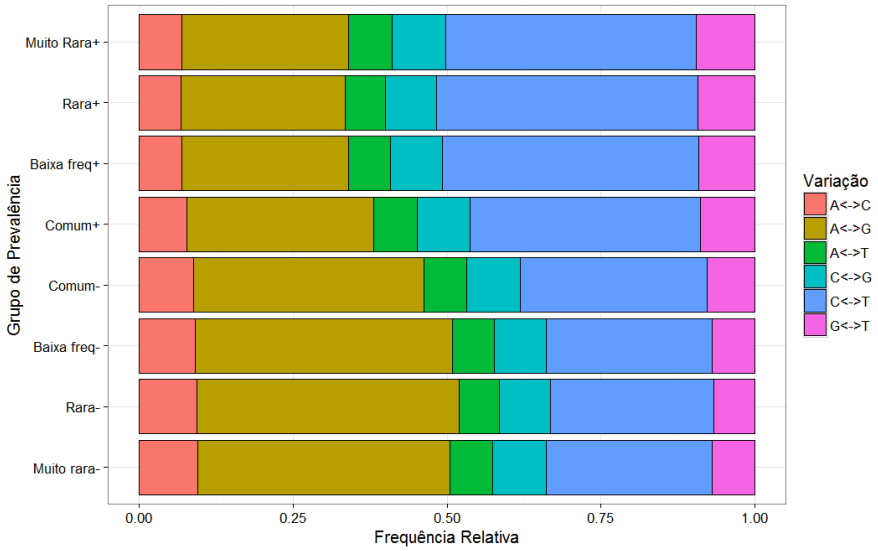


Figura 3.8: Gráfico de barras das frequências relativas de cada SNV por grupo de prevalência.

Verifica-se que não existe um ajustamento dos resíduos à distribuição normal padrão, confirmando a rejeição da hipótese de independência entre a ocorrência de cada SNV e o grupo de prevalência.

3.3 Equilíbrio de Hardy-Weinberg

Em seguida, verficou-se se as cerca de 36.8 milhões de SNVs, registadas no genoma dos 1092 indivíduos da fase 1 do P1000G, satisfazem ou não o HWE. Tendo em conta a grande dimensão dos dados, recorreu-se à função `HWChisqStats` para obter, de forma menos exaustiva, as estatísticas χ^2 e valores-p, correspondentes ao teste de χ^2 para o HWE. Esta função realiza, para cada posição do genoma onde ocorreu a variação, o teste de ajustamento do χ^2.

Para avaliar o HWE, foi realizada uma análise global (tabela 3.4), por cromossoma (tabela 3.5) e por SNV (tabela 3.6). Verificou-se que não é possível aplicar o HWE às SNVs cujos genótipos têm frequências, por exemplo, $(n_{xx},\ n_{xy},\ n_{yy}) = (1092, 0, 0)$. Estes casos correspondem às posições em que não houve variação (prevalência igual a zero). Nestas situações, as funções aplicadas do *software* R retornam um `Na`.

Verificam HWE	Não verificam HWE	`Na`
30 048 409 (81.99%)	6 600 583 (18.01%)	172 000

Tabela 3.4: Resultados globais do teste do χ^2 do HWE, aplicado às 36.8 milhões de SNVs.

	HWE		
Cromossoma	Verificam	Não verificam	`Na`
1	2365512 (82.07 %)	516850 (17.93 %)	14598
2	2615537 (82.37 %)	559763 (17.63 %)	14589
3	2184213 (82.36 %)	467664 (17.64 %)	12124
4	2140831 (81.62 %)	482131 (18.38 %)	11038
5	2006125 (82.61 %)	422253 (17.39 %)	10455
6	1897405 (81.76 %)	423382 (18.24 %)	9944
7	1738169 (81.87 %)	384872 (18.13 %)	9240
8	1728495 (82.24 %)	373386 (17.76 %)	9487
9	1301836 (82.01 %)	285642 (17.99 %)	7003
10	1482761 (82.10 %)	323331 (17.90 %)	8055
11	1498279 (82.41 %)	319871 (17.59 %)	9134
12	1435982 (82.00 %)	315249 (18.00 %)	8461
13	1077409 (82.00 %)	236571 (18.00 %)	5558
14	983968 (81.68 %)	220720 (18.32 %)	6324
15	885099 (81.67 %)	198696 (18.33 %)	5164
16	955811 (81.92 %)	211005 (18.08 %)	5790
17	819363 (81.83 %)	181980 (18.17 %)	5776
18	851469 (81.62 %)	191768 (18.38 %)	4734
19	624660 (80.40 %)	152299 (19.60 %)	5013
20	674504 (82.15 %)	146574 (17.85 %)	3875
21	402045 (81.26 %)	92739 (18.74 %)	3040
22	378936 (80.15 %)	93837 (19.85 %)	2598

Tabela 3.5: Resultados do teste do χ^2 para o HWE, aplicado às SNVs por cromossoma.

A partir da tabela 3.4 conclui-se que em 36 648 992 (99.5%) SNVs pode-se aplicar o teste de χ^2 para o HWE, e que a maioria das SNVs, cerca de 82%, estão de acordo com o HWE.

Pela tabela 3.5 conclui-se que a percentagem de SNVs que estão de acordo com o HWE é semelhante de cromossoma para cromossoma. Contudo, nos cromossomas 19 e 22, a percentagem de variações que violam o equilíbrio é maior do que nos outros cromossomas.

Para obter os diagramas de dispersão que permitam avaliar o HWE em cada um dos cromossomas, pode-se recorrer à função `HWGenotypePlot`, na qual se deve especificar o parâmetro `plottype` com `1` ou `2`, conforme se pretendam as frequências heterozigóticas/homozigóticas ou homozigóticas/homozigóticas, respetivamente. Na figura 3.9, estão exemplificados os dois tipos de diagramas de dispersão, para o cromossoma 16. Observa-se que no cromossoma 16, as SNVs seguem a tendência das parábolas em ambos os gráficos, indicando que a grande maioria verifica o HWE, tendência essa partilhada pelos restantes cromossomas.

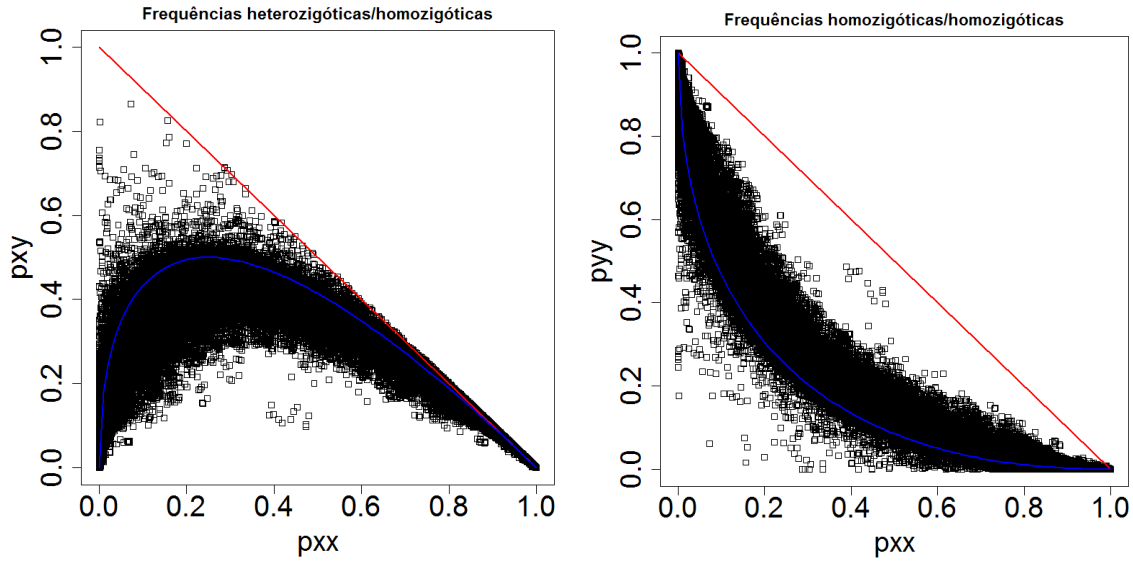


Figura 3.9: Diagramas de dispersão das frequências heterozigóticas/homozigóticas e homozigóticas/homozigóticas no cromossoma 16, com a respetiva curva do HWE.

Na figura 3.10 estão exemplifcados os *ternary plots* para os cromossomas 11 (à esquerda) e 22 (à direita). Este tipo de gráfico é elaborado através da função `HWTernaryPlot`, que por definição considera um nível de significância $\alpha = 0.05$. Os pontos verdes correspondem às SNVs que se encontram na região de aceitação para o HWE e os pontos vermelhos representam as variações que violam o HWE, estando mais dispersos e dando a sensação que se encontram em maior número, algo que não se verifica, tendo em conta a tabela 3.5. Pela observação da figura 3.10, não se encontram diferenças consideráveis entre os cromossomas 11 e 22 relativamente ao número de variações que verificam o HWE, padrão esse também partilhado pelos restantes cromossomas.

Foi ainda feita uma análise por tipo de SNV, no que diz respeito ao HWE. Através da tabela 3.6, verifica-se que as percentagem de SNVs que verificam o HWE são muito semelhantes, de variação para variação, concluindo-se o mesmo, entre as transições e transversões, a partir do gráfico 3.11.

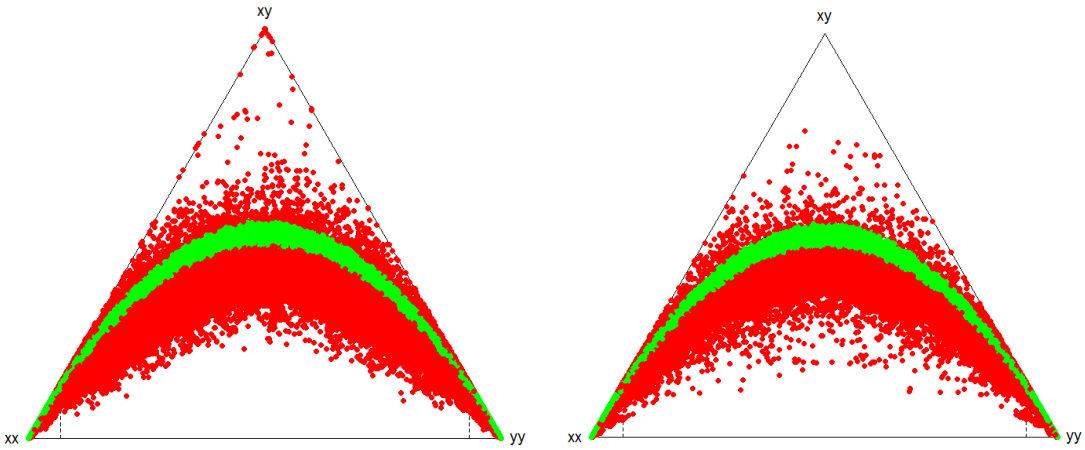


Figura 3.10: *Ternary plots* para os cromossomas 11 (à esquerda) e 22 (à direita).

SNV	Verificam HWE	Não verificam HWE	`Na`
$A \leftrightarrow C$	2447502 (81.86 %)	542300 (18.14 %)	14693
$A \leftrightarrow G$	10297877 (82.06 %)	2250987 (17.94 %)	57612
$A \leftrightarrow T$	2051385 (81.96 %)	451527 (18.04 %)	12238
$C \leftrightarrow G$	2554664 (81.65 %)	574323 (18.35 %)	14331
$C \leftrightarrow T$	10263224 (82.07 %)	2242065 (17.93 %)	58407
$G \leftrightarrow T$	2433757 (81.86 %)	539381 (18.14 %)	14719

Tabela 3.6: Resultados do teste do χ^2 para o HWE para cada SNV.

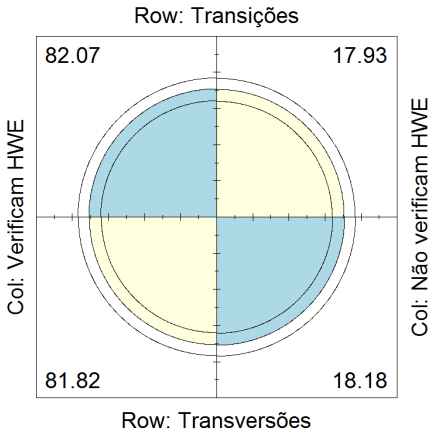


Figura 3.11: *Fourfold plot* dos resultados do HWE para as transições e transversões.

Capítulo 4

Análise do contexto onde ocorrem as variações de nucleótido único

Neste capítulo, apresentam-se os resultados que permitem caracterizar o contexto na vizinhança da posição onde ocorre a SNV, através da contagem de palavras de comprimento k com $k = 1, 2, 3$, isto é, nucleótidos, dinucleótidos e trinucleótidos, tendo como objetivo encontrar um padrão que seja indicador do fenómeno. Para isso começou-se por fazer um estudo global do contexto na vizinhança, para diversas amplitudes w da vizinhança, considerando $w = 5, 10, 20, 50, 100, 200$. Tal como se pode observar na tabela 4.1, por exemplo, para $w = 10$ e $k = 2$ vão contabilizar-se todos os dinucleótidos existentes na sequência do genoma de referência, quer à esquerda quer à direita, da posição onde se registou a ocorrência da SNV $A \leftrightarrow G$. Para além disso, procedeu-se à contagem de palavras de comprimento k, localizadas d nucleótidos à direita e à esquerda de cada local de variação, com $d \leq w$. Na tabela 4.1, as palavras de comprimento $k = 2$, localizadas $d = +2$ nucleótidos para a direita da posição onde ocorreu $A \leftrightarrow G$, estão realçadas. Fixando w, as contagens das palavras de comprimento $k = 1, 2, 3$ nas posições $d = \pm 1, \pm 2, \ldots, \pm w$, para cada tipo de SNV, foram organizadas em tabelas de contingência. Para visualizar a vizinhança, efetuar as contagens dos oligonucleótidos nas sequências correspondentes aos dois lados (esquerda e direita) e respetivas somas por variação, utilizaram-se as funções `snp.neighborhood` (função B.4), `oligonucleotideFrequency.snp.neighborhood` (função B.5), `sum.chr` (função B.6). Para contabilizar os oligonucleótidos existentes em cada posição d da vizinhança da SNV e encontrar um padrão que seja indicativo da ocorrência dessa variação, recorreu-se à função `neighborhood.pattern` (função B.7). Foi ainda realizado o mesmo tipo de análise por grupo de prevalência. Neste caso recorreu-se à função `sum.chr.grp` (função B.8) para efetuar a soma das contagens para cada variação e por grupo de prevalência.

CHR	POS	SNV	$A\|A$	$A\|G$	$G\|G$	Sequência à esquerda	Sequência à direita
1	10583	$A \leftrightarrow G$	5	304	783	$CCCTCGCGGT$	$CT\mathbf{CT}CCGGGT$
1	54421	$A \leftrightarrow G$	881	202	9	$TAATTGCTTT$	$TC\mathbf{AC}TCATAT$
1	54490	$A \leftrightarrow G$	13	149	930	$ATACTCTACC$	$GG\mathbf{CT}TCTGGA$
1	55330	$A \leftrightarrow G$	0	1	1091	$TACTATTTAC$	$CT\mathbf{TC}AGTAAA$

Tabela 4.1: Registos para a variação $A \leftrightarrow G$. Contagem de dinucleótidos ($k = 2$) numa vizinhança de amplitude $w = 10$.

4.1 Análise global do contexto em torno de cada variação

Na tabela 4.2 e na figura 4.1 apresentam-se as frequências relativas das contagens de nucleótidos, isto é, palavras de tamanho $k = 1$, na vizinhança de cada SNV considerando, por exemplo, uma amplitude de $w = 5$ e de $w = 100$. As respetivas contagens encontram-se em apêndice na tabela A.4. Verifica-se, que tanto para vizinhanças pequenas, $w = 5$, como para vizinhanças grandes, $w = 100$, os resultados são muito similares e próximos das frequências relativas dos nucleótidos no genoma global. Essas frequências relativas globais estimadas são aproximadamente $A(0.2953)$, $C(0.2045)$, $G(0.2046)$ e $T(0.2957)$. Constata-se, por exemplo, que os nucleótidos A e T são os que mais ocorrem em todas as variações, como já era esperado. No entanto, o nucleótido A apresenta uma maior frequência na vizinhança das variações $A \leftrightarrow C$ e $A \leftrightarrow T$ e o nucleótido T destaca-se nas variações $A \leftrightarrow T$ e $G \leftrightarrow T$. Quando a amplitude da vizinhança é $w = 5$, o nucleótido C é mais frequente na variação $A \leftrightarrow G$ enquanto que, quando $w = 100$, este ocorre com maior frequência na variação $C \leftrightarrow G$. Verifica-se também que para $w = 5$ a ocorrência do nucleótido G é maior na variação $C \leftrightarrow T$, ao contrário do que passa quando $w = 100$ em que a frequência é maior na SNV $C \leftrightarrow G$.

		SNV					
w	Nucleótidos	$A \leftrightarrow C$	$A \leftrightarrow G$	$A \leftrightarrow T$	$C \leftrightarrow G$	$C \leftrightarrow T$	$G \leftrightarrow T$
5	A	0.3089	0.2806	0.3141	0.2874	0.2807	0.2849
	C	0.2162	0.2219	0.1854	0.2119	0.2160	0.1889
	G	0.1895	0.2161	0.1858	0.2123	0.2220	0.2162
	T	0.2854	0.2814	0.3147	0.2885	0.2813	0.3099
100	A	0.2999	0.2931	0.3061	0.2872	0.2889	0.2928
	C	0.2056	0.2067	0.1935	0.2124	0.2107	0.2010
	G	0.2012	0.2109	0.1937	0.2126	0.2068	0.2056
	T	0.2933	0.2894	0.3066	0.2878	0.2937	0.3006

Tabela 4.2: Frequências relativas da contagem de nucleótidos ($k = 1$) na vizinhança de cada SNV, considerando as amplitudes $w = 5$ e $w = 100$.

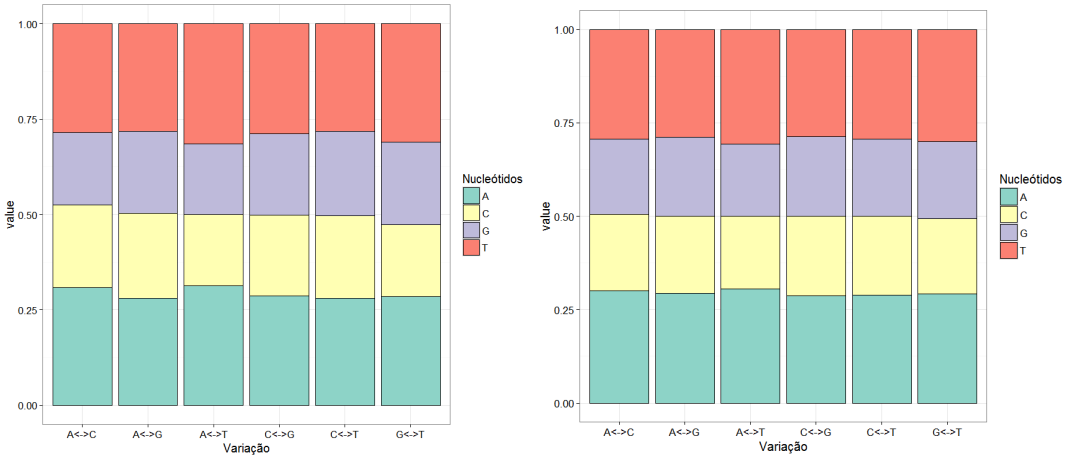


Figura 4.1: Gráficos de barras das frequências relativas da contagem de nucleótidos ($k = 1$) na vizinhança de cada SNV, considerando as amplitudes $w = 5$ e $w = 100$.

Em seguida, a tabela 4.3, mostra as frequências relativas das contagens dos dinucleótidos ($k = 2$) em vizinhanças com amplitudes $w = 10$ e $w = 200$. As contagens encontram-se em apêndice, nas tabelas A.5 e A.6 assim como, os respetivos gráficos de barras de frequências relativas, na figura A.2. Observa-se, que à medida que a amplitude da vizinhança aumenta, mais as frequências relativas dos 16 dinucleótidos se aproximam das do genoma, cujos valores são, aproximadamente, AA (0.0977), AC (0.0503), AG (0.0699), AT (0.0773), CA (0.0725), CC (0.0521), CG (0.0099), CT (0.06996), GA (0.0593), GC (0.0427), GG (0.0521), GT (0.0504), TA (0.0656), TC (0.0593) TG (0.0727) e TT (0.0980).

		SNV					
w	Dinucleótidos	$A \leftrightarrow C$	$A \leftrightarrow G$	$A \leftrightarrow T$	$C \leftrightarrow G$	$C \leftrightarrow T$	$G \leftrightarrow T$
10	AA	0.1063	0.0925	0.1169	0.0889	0.0865	0.0949
	AC	0.0534	0.0525	0.0480	0.0509	0.0511	0.0464
	AG	0.0682	0.0728	0.0639	0.0745	0.0667	0.0703
	AT	0.0778	0.0726	0.0852	0.0685	0.0726	0.0779
	CA	0.0743	0.0736	0.0680	0.0743	0.0742	0.0695
	CC	0.0562	0.0548	0.0443	0.0573	0.0578	0.0480
	CG	0.0091	0.0105	0.0071	0.0109	0.0104	0.0091
	CT	0.0704	0.0668	0.0638	0.0747	0.0729	0.0683
	GA	0.0572	0.0627	0.0563	0.0613	0.0593	0.0601
	GC	0.0404	0.0476	0.0356	0.0451	0.0476	0.0403
	GG	0.0482	0.0579	0.0443	0.0575	0.0549	0.0561
	GT	0.0466	0.0513	0.0482	0.0511	0.0528	0.0536
	TA	0.0667	0.0636	0.0766	0.0596	0.0637	0.0668
	TC	0.0601	0.0594	0.0563	0.0614	0.0628	0.0572
	TG	0.0697	0.0745	0.0682	0.0745	0.0738	0.0746
	TT	0.0952	0.0869	0.1172	0.0894	0.0929	0.1068
200	AA	0.1001	0.0962	0.1031	0.0932	0.0942	0.0961
	AC	0.0513	0.0505	0.0501	0.0505	0.0507	0.0493
	AG	0.0700	0.0715	0.0682	0.0717	0.0697	0.0699
	AT	0.0778	0.0753	0.0828	0.0728	0.0753	0.0778
	CA	0.0735	0.0730	0.0714	0.0734	0.0733	0.0716
	CC	0.0522	0.0530	0.0475	0.0554	0.0544	0.0509
	CG	0.0094	0.0103	0.0081	0.0109	0.0102	0.0094
	CT	0.0700	0.0698	0.0682	0.0717	0.0716	0.0701
	GA	0.0595	0.0606	0.0585	0.0603	0.0588	0.0594
	GC	0.0421	0.0440	0.0391	0.0451	0.0440	0.0421
	GG	0.0510	0.0545	0.0476	0.0555	0.0530	0.0521
	GT	0.0495	0.0508	0.0503	0.0507	0.0506	0.0514
	TA	0.0661	0.0638	0.0713	0.0613	0.0638	0.0661
	TC	0.0595	0.0589	0.0586	0.0603	0.0606	0.0595
	TG	0.0717	0.0735	0.0716	0.0736	0.0731	0.0737
	TT	0.0964	0.0945	0.1034	0.0936	0.0966	0.1005

Tabela 4.3: Frequências relativas de dinucleótidos ($k = 2$) na vizinhança de cada SNV, considerando as amplitudes $w = 10$ e $w = 200$.

Verifica-se, por exemplo, que para a variação $A \leftrightarrow T$, os dinucleótidos AA e TT têm maior frequência em vizinhança com pequena amplitude, $w = 10$, do que grande, $w = 200$. Para $w = 10$ tem-se que os dinucleótidos que ocorrem com maior frequência são: AC na variação $A \leftrightarrow C$; AG na variação $C \leftrightarrow G$; AT na variação $A \leftrightarrow T$; CA nas variações $A \leftrightarrow C$ e $C \leftrightarrow G$; CC na variação $C \leftrightarrow T$; CG na variação $C \leftrightarrow G$; CT na variação $C \leftrightarrow G$; GA na variação $A \leftrightarrow G$; GC nas variações $A \leftrightarrow G$ e $C \leftrightarrow T$; GG na variação $A \leftrightarrow G$; GT na variação $G \leftrightarrow T$; TA na variação $A \leftrightarrow T$; TC na variação $C \leftrightarrow T$ e TG na variação $G \leftrightarrow T$. Confirma-se que o dinucleótido CG é o menos frequente, para todas as SNVs, mas que este tem uma preferência pela transição $A \leftrightarrow G$ e pela transversão $C \leftrightarrow G$.

Também foram calculadas as frequências relativas das contagens dos trinucleótidos ($k = 3$) numa vizinhança de amplitude, por exemplo, $w = 10$. Os respetivos resultados encontram-se em apêndice nas tabelas A.7 e A.8 e no gráfico de barras da figura A.3. Destacam-se os trinucleótidos AAA e TTT como os mais frequentes e ACG, CCG, CGA, CGC, CGG, CGT, GCG, TCG como os menos frequentes. Os trinucleótidos mais raros são o GCG e o CGC, ambos contendo uma estrutura CpG. Os valores obtidos também vão de encontro às frequências relativas dos trinucleótidos, considerando todo o genoma.

Uma vez que, para cada amplitude ($w = 5, 10, 20, 50, 100, 200$), as contagens dos nucleótidos, dinucleótidos e trinucleótidos ($k = 1, 2, 3$) por SNV, foram organizadas em tabelas de contingência, foi avaliada a existência ou não de associação entre o contexto e as SNVs. Na tabela 4.4 apresentam-se os resultados dos testes de independência para cada uma das amplitudes. Os valores apresentados são a estatística do teste χ^2, o V de Cramér, os graus de liberdade ($g.l.$). Note-se que, por exemplo, para o caso $k = 1$, como são 4 nucleótidos e 6 tipos de SNV, tem-se $g.l. = (6 - 1)(4 - 1) = 15$. Para todas as vizinhanças w, rejeitou-se a hipótese de independência entre as contagens das palavras de comprimento k e as SNVs, pois obteve-se sempre o valor-$p \approx 0.000$. No entanto, ao proceder à medição da força da associação, constatou-se que quanto maior for a amplitude da vizinhança w, menor é a força da associação entre as contagens e o tipo de variação. Por outro lado também se observa que, fixando a amplitude da vizinhança w, a força da associação é maior no caso dos trinucleótidos ($k = 3$), indiciando que é este tipo de contexto que poderá ter uma maior influência na ocorrência de SNVs.

		$k = 1$		$k = 2$		$k = 3$	
		χ^2	V	χ^2	V	χ^2	V
w	5	744566	0.026	1632787	0.033	2426324	0.047
	10	1025739	0.022	2345548	0.027	3262358	0.033
	20	1368438	0.018	3078247	0.021	4366549	0.026
	50	2124700	0.014	4612100	0.016	6589400	0.019
	100	2828900	0.011	6082400	0.013	8788300	0.016
	200	3919400	0.009	8404600	0.011	12272000	0.013
$g.l.$		15		75		315	

Tabela 4.4: Resultados dos testes de independência entre as contagens de palavras de comprimento k ($k = 1, 2, 3$) e cada tipo de SNV, considerando diferentes amplitudes para a vizinhança ($w = 5, 10, 20, 50, 100, 200$).

Com a rejeição de H_0 pressupõe-se a realização de uma análise aos resíduos, resultantes dos testes de independência. Os valores dos resíduos ajustados correspondentes aos nucleótidos ($k = 1$), apresentam-se em apêndice na tabela A.9, para as amplitudes $w = 5$ e $w = 100$.

Os *heatmaps* para cada uma das amplitudes estão representados na figura 4.2. Para $w = 5$, por exemplo, a variação $A \leftrightarrow T$ prefere a ocorrência dos nucleótidos A e T e evita a ocorrência dos nucleótidos C e G. O nucleótido T é o preferido da variação $G \leftrightarrow T$ enquanto que, o nucleótido C é o preterido desta SNV. Na variação $A \leftrightarrow C$ é o nucleótido A o mais frequente e o C o menos frequente. Nas transições $A \leftrightarrow G$ e $C \leftrightarrow T$ os nucleótidos com maior frequência são o C e G, respetivamente, evitando a ocorrência dos nucleótidos A e T. Por fim, a variação $C \leftrightarrow G$ não manifesta nenhuma preferência relativamente a algum dos nucleótidos. Através dos dendrogramas do *heatmap* para $w = 5$, pode observar-se que tanto os nucleótidos como as SNVs foram agrupados em dois grupos: $\{A, T\}$ e $\{C, G\}$ para os nucleótidos e $\{A \leftrightarrow T, G \leftrightarrow T, A \leftrightarrow C\}$ e $\{A \leftrightarrow G, C \leftrightarrow T, C \leftrightarrow G\}$ para as variações. Note-se que as preferências das variações $A \leftrightarrow C/G \leftrightarrow T$ e $A \leftrightarrow G/C \leftrightarrow T$ parecem ser opostas, algo que poderá refletir a simetria complementar do ADN. Quando, por exemplo, a amplitude da vizinhança aumenta para $w = 100$, as preferências das variações $C \leftrightarrow T$, $C \leftrightarrow G$ e $A \leftrightarrow G$ não se mantêm, mas a forma como os nucleótidos e as SNVs se agrupam continua igual.

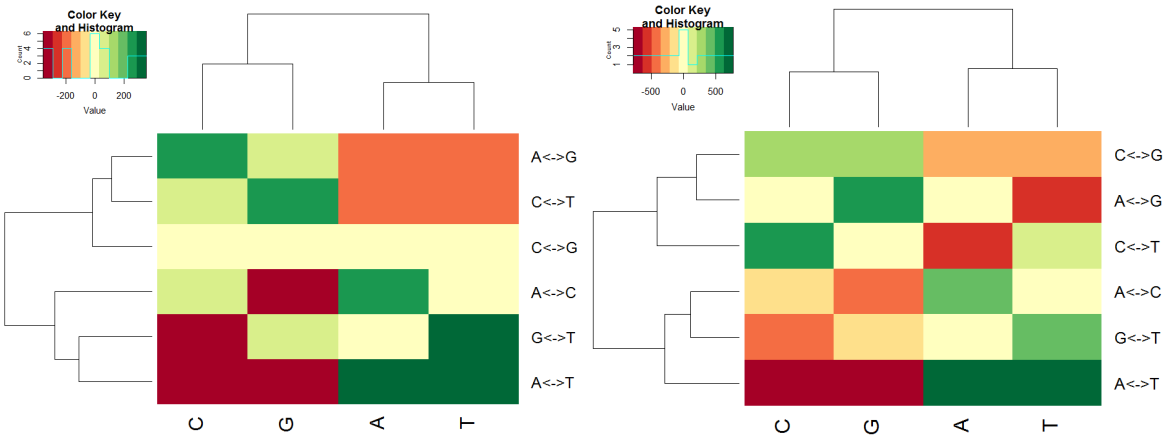


Figura 4.2: *Heatmaps* dos resíduos ajustados do teste de independência entre os nucleótidos ($k = 1$) e as SNVs, com $w = 5$ e $w = 100$.

Os resíduos ajustados resultantes dos teste de independência entre os dinucleótidos ($k = 2$) e as SNVs, para vizinhanças de amplitudes $w = 10$ e $w = 200$, encontram-se em apêndice na tabela A.10). Na figura 4.3 está representado o *heatmap* para quando se considera a vizinhança com amplitude $w = 10$. Observa-se que, por exemplo, a variação $A \leftrightarrow T$ tem uma clara preferência pela ocorrência dos dinucleótidos AA e TT, ao contrário do que acontece com as variações $A \leftrightarrow G$ e $C \leftrightarrow T$. A transversão $A \leftrightarrow T$ parece evitar a ocorrência dos dinucleótidos GC, GG e CC. Relativamente às transições, tem-se que os dinucleótidos CC, TC e CT são os preferidos da variação $C \leftrightarrow T$ sendo AA o preterido, e que a variação $A \leftrightarrow G$ prefere os dinucleótidos GA, AG, GC e GG evitando o TT. Para a transversão, $C \leftrightarrow G$ são os dinucleótidos CG, AG, GG, CC e CT os que apresentam maior frequência. Note-se que para esta SNV é possível formar estruturas CpG. Para as transversões $A \leftrightarrow C$ e $G \leftrightarrow T$ os dinucleótidos preferidos são AA e TT, respetivamente. Observe-se ainda que a variação $A \leftrightarrow T$ é a que mais evita a ocorrência da estrutura CpG. Nos dendrogramas, as SNVs

ficaram agrupadas de igual forma ao que aconteceu para os nucleótidos. De acordo com o corte apresentado na figura, os dinucleótidos formaram 4 grupos, em que 3 deles são, $\{AA\}$, $\{TA, AT, TT\}$, $\{CC, TC, CT\}$. Quando a amplitude da vizinhança aumenta, $w = 200$, perde-se alguma da informação relativa às preferências/preterências das variações, nos locais mais próximos onde ocorreu a SNV. O *heatmap* correspondente está na figura A.4.

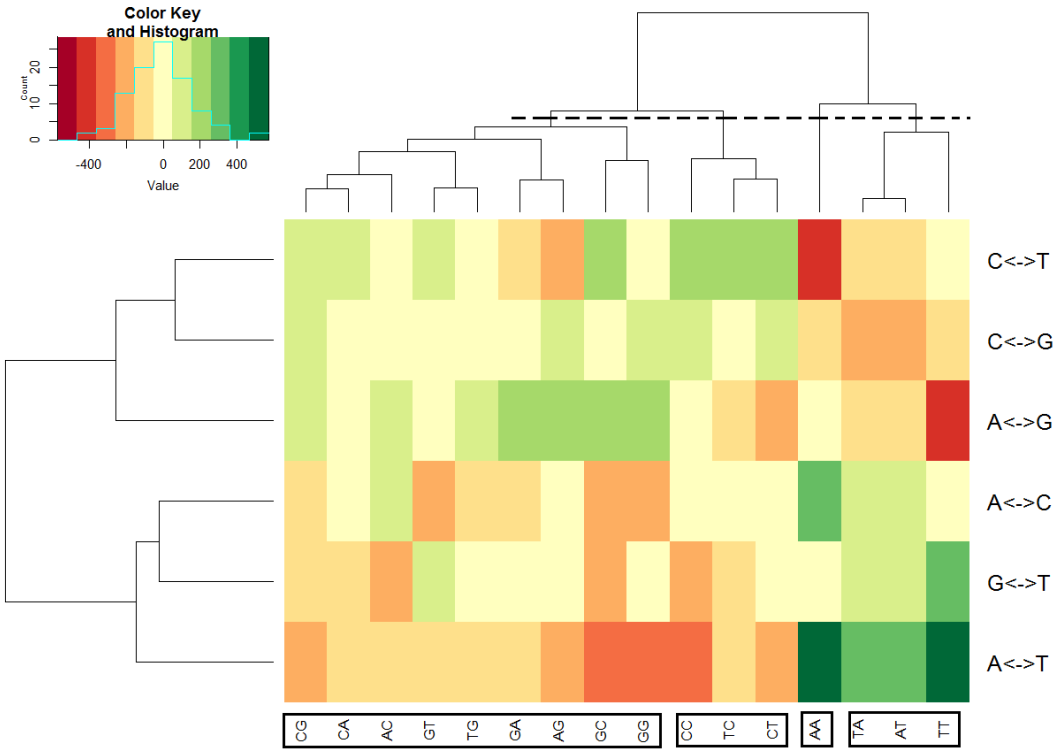


Figura 4.3: *Heatmap* dos resíduos ajustados do teste de independência entre os dinucleótidos ($k = 2$) e as SNVs, com $w = 10$.

Para os trinucleótidos ($k = 3$), também foram calculados os respetivos resíduos ajustados, considerando uma vizinhança de amplitude pequena, por exemplo, $w = 10$. Os valores dos resíduos estão em apêndice nas tabelas A.11 e A.12. No *heatmap* correspondente, representado na figura 4.4, constata-se que são os trinucleótidos TAT, TAA, AAT, ATA, TTA e ATT que mais ocorrem na variação $A \leftrightarrow T$, sendo mais evidente a preferência desta SNV pelos trinucleótidos AAA e TTT. Para além disso, esta transversão parece evitar a ocorrência de GGG, GGC, GCC e CCC. Também as transversões $A \leftrightarrow C$ e $G \leftrightarrow T$ têm mais preferência por AAA e TTT, respetivamente, do que por qualquer outro trinucleótido. Relativamente às transições $C \leftrightarrow T$ e $A \leftrightarrow G$, conclui-se que evitam os trinucleótidos AAA e TTT, respetivamente. Verifica-se também que a variação $C \leftrightarrow T$ prefere a ocorrência dos trinucleótidos GCC, CTC e e CCT enquanto que a outra transição, $A \leftrightarrow G$, prefere GGC, GAG e AGG. Considerando o corte apresentado no dendrograma dos trinucleótidos, resulta que estes ficam divididos em 6 grupos, os quais estão assinalados na figura 4.4.

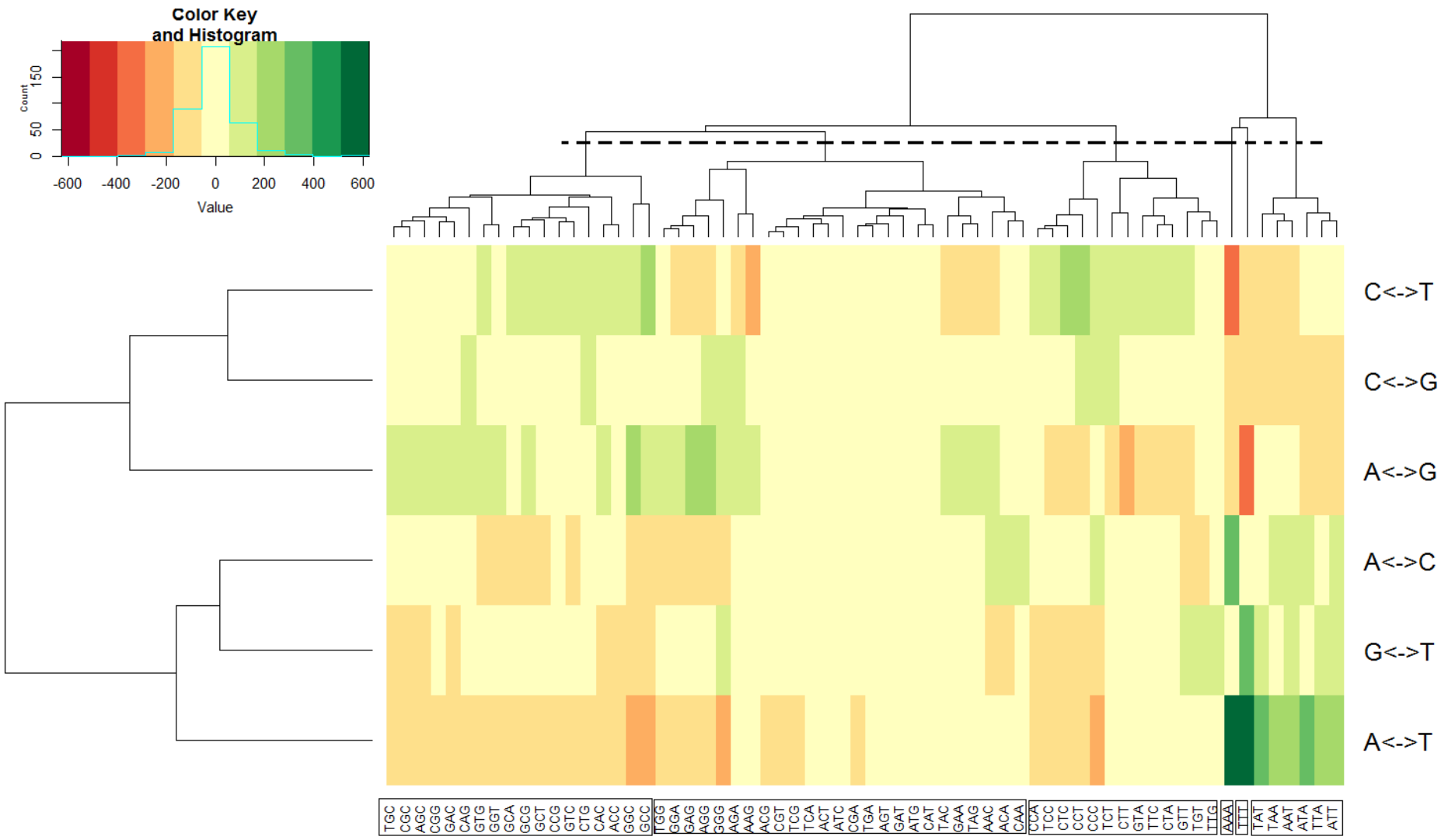


Figura 4.4: *Heatmap* dos resíduos ajustados do teste de independência entre os trinucleótidos ($k = 3$) e as SNVs com $w = 10$.

4.2 Análise dos padrões de frequência em torno de cada variação

Em seguida, para cada uma das SNVs, procedeu-se à contagem de nucleótidos, dinucleótidos e trinucleótidos, localizados d nucleótidos à direita e à esquerda de cada local de variação, de acordo com a amplitude da vizinhança w escolhida. Os resultados obtidos foram organizados em tabelas de contingência.

Considerando, por exemplo, uma vizinhança de amplitude $w = 20$, determinaram-se em cada uma das posições $d = \pm 1, \pm 2, \ldots, \pm 20$, as frequências relativas das contagens dos nucleótidos ($k = 1$), para as transições e para as transversões. Esses resultados encontram-se em apêndice, na tabela A.13. A tabela 4.5 mostra as frequências relativas nas posições mais próximas do local de variação ($d = \pm 1, \pm 2, \pm 3, \pm 4$), para cada uma das SNVs.

Verifica-se que à medida que a posição d se afasta do local de variação, as frequências relativas de cada um dos nucleótidos tendem a aproximar-se das frequências relativas no genoma, ou seja, A (0.2953), C (0.2045), G (0.2046) e T (0.2957). É nas posições muito próximas do local da SNV que a distribuição das palavras de tamanho $k = 1$ é mais díspar da obtida no genoma, ou por serem superiores ao esperado (viés grande e positivo) ou por serem inferiores (viés grande e negativo). Na tabela 4.5 encontram-se destacadas as frequências relativas responsáveis pelas maiores discrepâncias.

SNV	Posição	Nucleótidos				Posição	Nucleótidos			
	d	A	C	G	T	d	A	C	G	T
$A \leftrightarrow G$	-1	0.240	**0.349**	0.164	0.247	1	**0.209**	0.196	0.230	**0.364**
	-2	0.298	0.253	0.225	**0.224**	2	0.316	0.190	0.239	0.255
	-3	0.281	0.214	0.224	0.282	3	0.300	0.185	0.223	0.292
	-4	0.292	0.217	0.203	0.288	4	0.286	0.203	0.218	0.293
$C \leftrightarrow T$	-1	**0.363**	0.231	0.195	**0.210**	1	0.246	0.164	**0.349**	**0.240**
	-2	0.254	0.239	0.190	0.317	2	**0.223**	0.225	0.253	0.298
	-3	0.291	0.223	0.186	0.301	3	0.281	0.223	0.214	0.281
	-4	0.296	0.208	0.214	0.282	4	0.281	0.214	0.209	0.297
$A \leftrightarrow C$	-1	0.312	**0.247**	0.207	**0.234**	1	0.330	**0.248**	**0.153**	0.269
	-2	0.332	0.189	0.196	0.282	2	0.298	0.198	0.184	0.319
	-3	0.331	0.207	0.186	0.277	3	0.275	0.216	0.190	0.319
	-4	0.318	0.225	0.193	0.263	4	0.292	0.207	0.192	0.309
$A \leftrightarrow T$	-1	0.325	0.221	0.203	**0.252**	1	**0.252**	0.202	0.221	0.325
	-2	0.340	0.166	0.180	0.314	2	0.313	0.180	0.167	0.340
	-3	0.330	0.171	0.193	0.306	3	0.305	0.192	0.172	0.331
	-4	0.328	0.178	0.177	0.316	4	0.315	0.177	0.179	0.329
$C \leftrightarrow G$	-1	0.330	**0.156**	0.207	0.308	1	0.305	0.206	**0.158**	0.331
	-2	0.283	**0.258**	0.191	0.269	2	0.268	0.191	**0.257**	0.284
	-3	0.300	0.228	0.196	0.277	3	0.276	0.195	0.229	0.301
	-4	0.283	0.232	0.218	0.266	4	0.265	0.218	0.232	0.284
$G \leftrightarrow T$	-1	0.270	**0.150**	**0.247**	0.333	1	**0.234**	0.207	**0.247**	0.312
	-2	0.318	0.184	0.198	0.300	2	0.282	0.195	0.189	0.333
	-3	0.319	0.190	0.216	0.276	3	0.276	0.186	0.207	0.331
	-4	0.309	0.192	0.207	0.293	4	0.262	0.193	0.226	0.319

Tabela 4.5: Frequências relativas de nucleótidos, nas posições $d = \pm 1, \pm 2, \pm 3, \pm 4$, na vizinhança de cada uma das SNVs.

As figuras 4.5 e 4.6 mostram os padrões de frequência para cada SNV, agrupadas por transições e transversões, respetivamente. Os gráficos representam o viés entre as frequências relativas dos nucleótidos ($k = 1$) e as respetivas frequências relativas no genoma, em função da posição d na vizinhança, com $d = \pm 1, \pm 2, \ldots, \pm 20$. Em apêndice, na figura A.5 podem observar-se os padrões de frequência globais das transições e transversões.

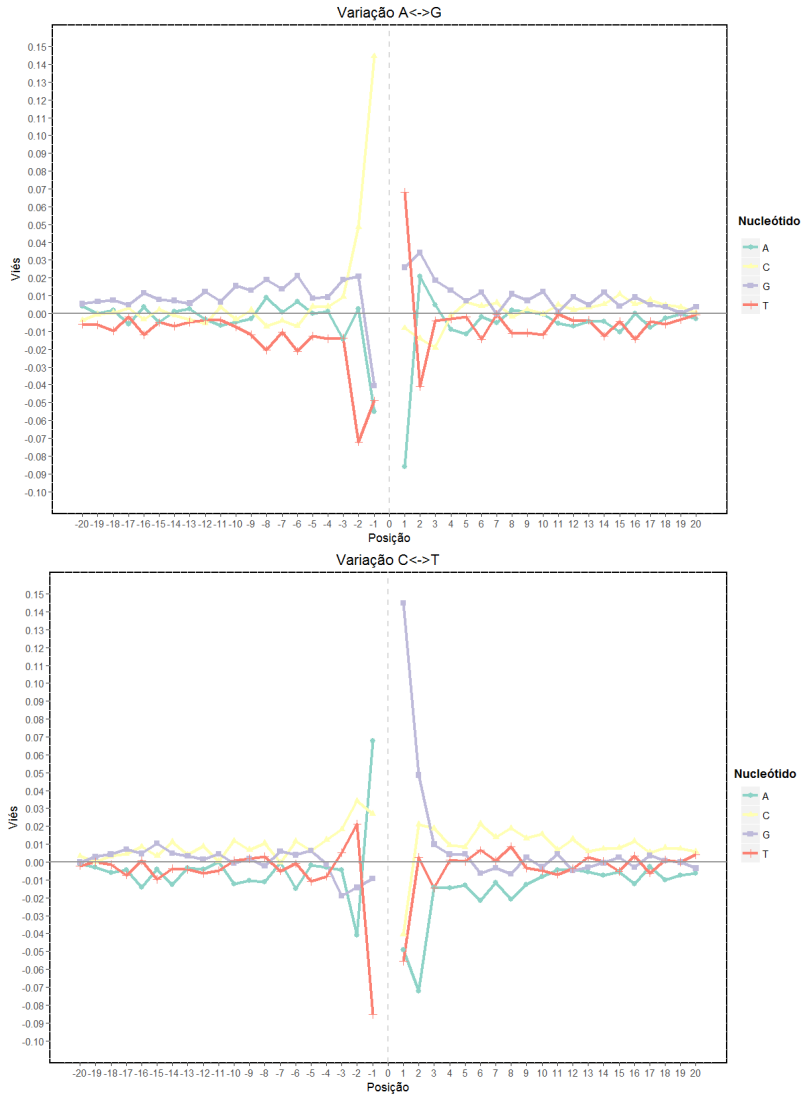


Figura 4.5: Padrões de frequência dos nucleótidos ($k = 1$) para as transições $A \leftrightarrow G$ e $C \leftrightarrow T$, numa vizinhança de amplitude $w = 20$.

Observa-se que na transição $A \leftrightarrow G$, é o nucleótido C, na posição $d = -1$, que tem maior viés relativamente à observada no genoma. Para $C \leftrightarrow T$, é o seu nucleótido complementar G, que na posição simétrica $d = +1$, regista uma frequência relativa superior ao esperado. Por sua vez, na posição $d = +1$, a SNV $A \leftrightarrow G$, tem preferência pela ocorrência do nucleótido T e evita o nucleótido A. A SNV $C \leftrightarrow T$, tem um comportamento simétrico, pois na posição $d = -1$, prefere o nucleótido A e pretere o T. Destaca-se ainda, um viés significativo do nucleótido T, na posição $d = -2$ em $A \leftrightarrow G$ e do nucleótido A, na posição $d = +2$ em $C \leftrightarrow T$.

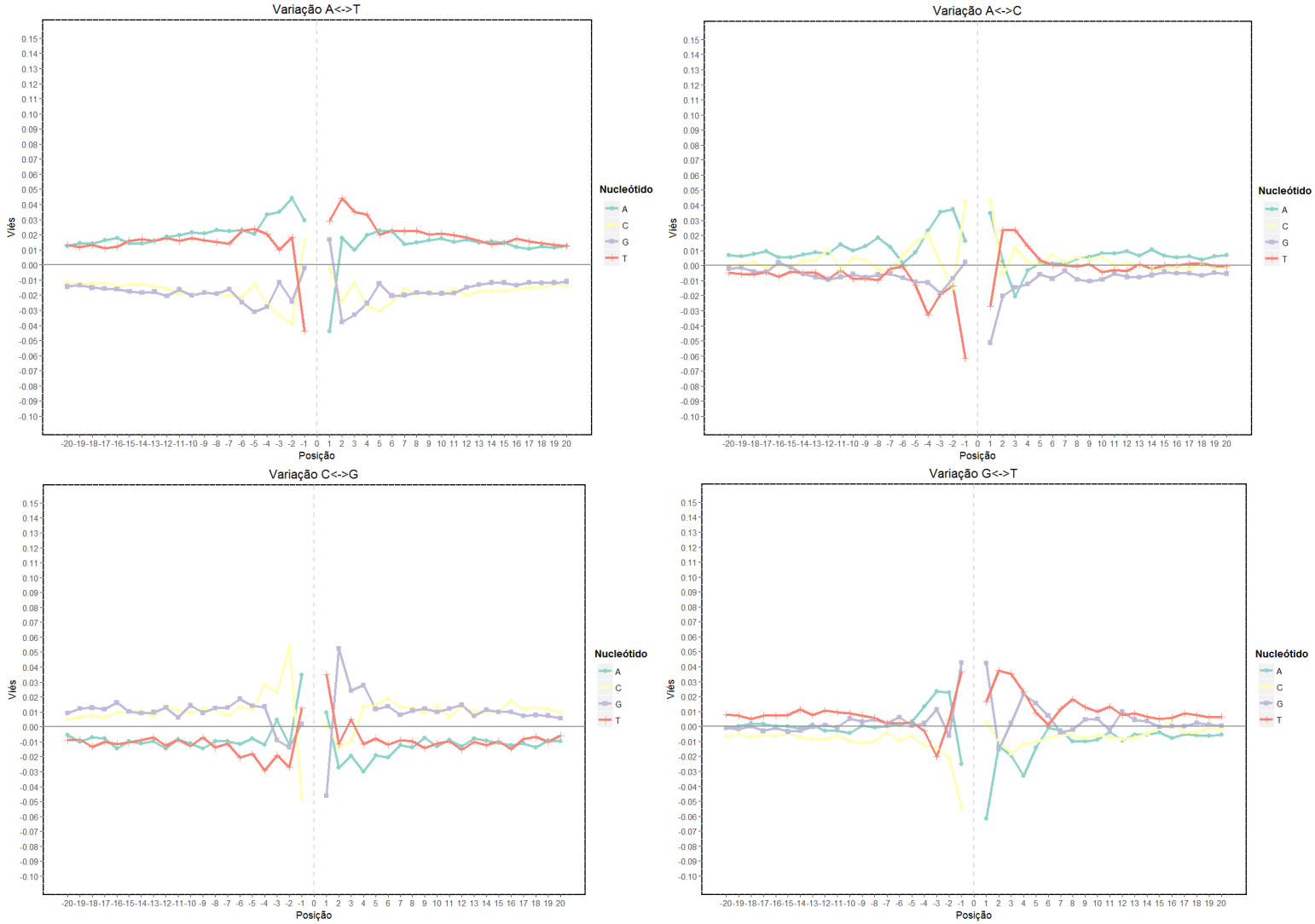


Figura 4.6: Padrões de frequência dos nucleótidos ($k = 1$) para as transversões $A \leftrightarrow C$, $A \leftrightarrow T$, $C \leftrightarrow G$ e $G \leftrightarrow T$, numa vizinhança de amplitude $w = 20$.

Relativamente às transversões, observa-se que é a variação $A \leftrightarrow T$, que para todas as 40 posições da vizinhança $w = 20$, apresenta maiores discrepâncias em relação à distribuição no genoma, uma vez que os seus perfis de frequência são os mais distantes de zero. Observa-se ainda que esta transversão, evita os nucleótidos T e A, nas posições $d = -1$ e $d = +1$, respetivamente, mas prefere esses mesmos nucleótidos, nas posições $d = +2$ e $d = -2$. Na transversão $C \leftrightarrow G$, os nucleótidos C e G apresentam um viés negativo nas posições $d = -1$ e $d = +1$, respetivamente, donde se poderá suspeitar que esta SNV evita os nucleótidos C e G nas posições imediatamente adjacentes ao local de variação, por causa se poderem formar estruturas CpG. Por outro lado, esta SNV, tem preferência pela ocorrência dos nucleótidos C e G, nas posições $d = -2$ e $d = +2$. A variação $G \leftrightarrow T$ tem uma maior preferência pelo nucleótido G nas posições imediatamente adjacentes ao local de variação, evitando também os nucleótidos C e A nas posições $d = -1$ e $d = +1$, respetivamente.

Através dos gráficos apresentados nas figuras 4.5 e 4.6, confirma-se que para uma vizinhança de amplitude $w = 20$ da SNV, é nas posições adjacentes ao local de variação $(\pm 1, \ldots, \pm 4)$ que ocorrem as maiores discrepâncias (viés) para as frequências nucleotídicas, relativamente às frequências de cada um dos nucleótidos A, C, G e T no genoma. Conclui-se que o viés em torno de transições é muito maior do que em torno das transversões.

Assim, para cada SNV, procedeu-se ao teste de ajustamento do χ^2 para avaliar se as frequências nucleotídicas se distribuem uniformemente ao longo da vizinhança. Ou seja, testar a hipótese nula

$$H_0 : p_d^x = p_{0d}^x \;\; com \;\; d = -20, \ldots, -1, +1, \ldots, +20\,,$$

com $x \in \{A, C, G, T\}$ e frequência esperada igual a

$$p_{0d}^x = 1/40\,.$$

Na tabela 4.6 encontram-se, para cada SNV, os resultados obtidos dos testes de ajustamento do χ^2 aplicados às frequências de cada nucleótido. Os valores apresentados correspondem às estatísticas χ^2, à medida da força do efeito ϕ e aos valores-p, onde $g.l. = 39$. Pelos valores-p apresentados, verifica-se que para todas as SNVs se rejeita a hipótese de que as frequências dos nucleótidos se distribuirem uniformememte ao longo de cada uma das posições da vizinhança da variação. Observa-se que os maiores valores para a medida da força do efeito, $\phi \approx 188.23$ e $\phi \approx 188.18$, surgem nas transições, os quais correspondem aos nucleótidos C e G, nas SNVs $A \leftrightarrow G$ e $C \leftrightarrow T$, respetivamente. Confirma-se portanto o que foi observado na figura 4.5 relativamente ao grande viés dos nucleótidos C ou G nas posições imediatamente adjacentes ao local de variação. Relativamente às transversões, os valores de ϕ são menores o que quer dizer que existe uma menor discrepância das contagens nucleotídicas ao longo de cada uma das posições da vizinhança. Contudo, os nucleótidos C e G são os que apresentam os maiores valores de ϕ dentro das transições, neste caso concreto, na variação $C \leftrightarrow G$.

Conclui-se então que para todas as variações há discrepâncias na forma como se distribuem as frequências nucleotídicas ao longo das posições, principalmente no que diz respeito às posições imediatamente adjacentes ao local de variação refletindo que cada tipo SNV tem um contexto de ocorrência específico e que são as transições que apresentam maior viés das frequências de cada nucleótido em relação às frequências no genoma.

SNV	Nucleótidos	χ^2	ϕ	valores-p
$A \leftrightarrow G$	A	476350	**109.13**	≈ 0.000
	C	1417200	**188.23**	
	G	265380	81.45	
	T	598970	**122.37**	
$C \leftrightarrow T$	A	596570	**122.12**	≈ 0.000
	C	267590	81.79	
	G	1416500	**188.18**	
	T	473680	**108.82**	
$A \leftrightarrow C$	A	39405	31.39	≈ 0.000
	C	65881	40.58	
	G	41073	32.04	
	T	74788	43.24	
$A \leftrightarrow T$	A	45792	33.83	≈ 0.000
	C	38234	30.92	
	G	38764	31.13	
	T	45419	33.70	
$C \leftrightarrow G$	A	41448	32.19	≈ 0.000
	C	105640	**51.39**	
	G	99263	**49.82**	
	T	42735	32.69	
$G \leftrightarrow T$	A	74277	43.09	≈ 0.000
	C	45406	33.69	
	G	64694	40.22	
	T	38577	31.06	

Tabela 4.6: Resultados dos testes de ajustamento do χ^2 às contagens dos nucleótidos, para cada tipo de SNV, considerando uma vizinhança de amplitude $w = 20$.

Note-se que tanto nas transições como nas transversões, em torno do local da variação, os nucleótidos complementares (A/T e C/G) parecem exibir padrões de frequência simétricas. Assim, foi testada, para as transições e transversões, a homogeneidade das contagens de cada par de nucleótidos complementares ao longo das posições simétricas. As tabelas de contingência correspondentes às contagens dos nucleótidos complementares, fixando as posições $d = \pm 1, \pm 2, \ldots, \pm 20$, encontram-se em apêndice nas tabelas A.14 e A.15. Os resultados dos testes de homogeneidade aplicados a cada uma dessas tabelas de contingência estão resumidos na tabela 4.7. Tem-se que tanto nas transições como nas transversões, as frequências dos nucleótidos complementares A/T distribuem-se de forma homogénea ao longo das posições simétricas da vizinhança, tal como acontece para os nucleótidos C/G nas transversões. No entanto, nas transições, para C/G rejeitou-se a hipótese de homogeneidade, contudo o valor de V de Cramér ≈ 0.0005 é muito pequeno. Ao proceder à análise de resíduos obtiveram-se, por exemplo, os *heatmaps* da figura 4.7, correspondentes aos pares de nucleótidos complementares A/T e C/G nas transições, de acordo com as posições simétricas. Para o caso em que foi rejeitada a hipótese nula, constata-se que na maioria das posições há uma clara preferência ou pelo nucleótido C ou pelo nucleótido G, à exceção das posições 10 e 16. No caso do par A/T, só nas posições 2, 3 e 10 é que existe uma maior preferência por algum destes nucleótidos.

	Nucleótidos Complementares	χ^2	V	valor-p	$g.l.$
Transições	A/T	21.565	≈ 0.0003	0.3064	19
	C/G	49.923	≈ 0.0005	0.0001	
Transversões	A/T	17.324	≈ 0.0004	0.5679	
	C/G	22.948	≈ 0.0005	0.2396	

Tabela 4.7: Resultados dos testes de homogeneidade entre as contagens de nucleótidos complementares em posições simétricas ao longo da vizinhança, para as transições e transversões.

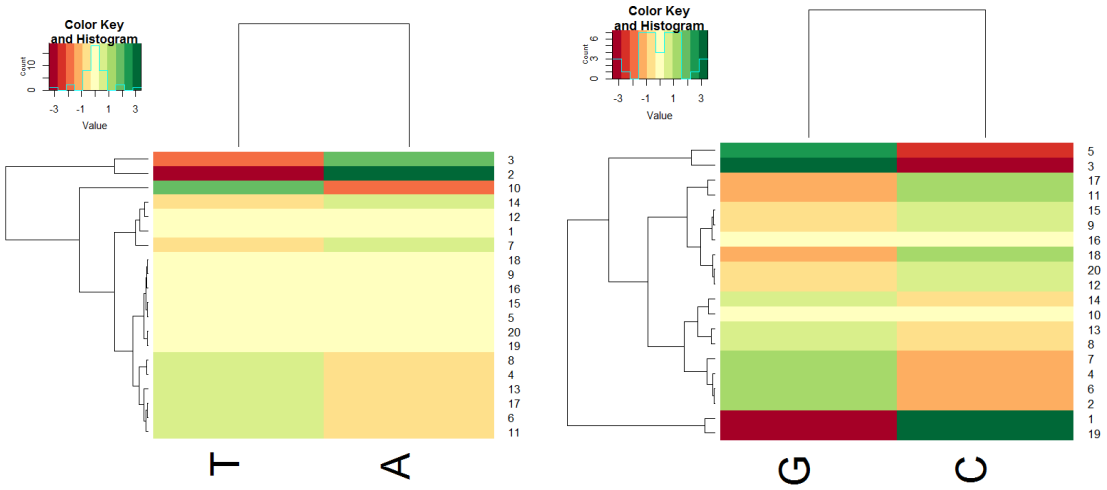


Figura 4.7: *Heatmaps* dos resíduos ajustados do teste de homogeneidade para as contagens dos pares de nucleótidos complementares (A/T e C/G), nas transições.

Em seguida, de forma análoga ao que foi feito para os nucleótidos, na análise dos dinucleótidos ($k = 2$), considerando uma vizinhança de amplitude $w = 20$, começou-se por determinar as frequências relativas das contagens em cada uma das posições $d = \pm 1, \pm 2, \ldots, \pm 20$, para cada uma das SNVs. As frequências relativas dos 16 dinucleótidos, nas posições mais próximas ($d = \pm 1, \pm 2, \pm 3, \pm 4$) do local de variação, encontram-se em apêndice na tabela A.16.

Também para os dinucleótidos, verifica-se que é nas posições mais próximas do local da SNV que as contagens das palavras de tamanho $k = 2$ são mais díspares das frequências relativas no genoma. Constata-se mais uma vez, que à medida que a posição d se afasta do local de variação, as frequências dos dinucleótidos tendem a aproximar-se das do genoma. Note-se que neste caso, o maior valor observado para o viés (≈ 0.0464) é inferior ao maior viés (≈ 0.145) calculado no caso dos nucleótidos.

Nas figuras 4.8 e 4.9 podem observar-se os padrões de frequência dos dinucleótidos, para cada SNV. Os gráficos correspondentes aos padrões de frequência globais para as transições e transversões, encontram-se na figura A.6.

Tem-se que, a transição $A \leftrightarrow G$, na posição $d = -1$, tem uma clara preferência pelo dinucleótido CC, pois este é o que apresenta maior viés, para além disso, também há uma preferência pelos dinucleótidos AC, GC e TC. Ainda nessa mesma posição, observa-se que o dinucleótido preterido é o TT, seguindo-se o TA, TG e AA. Por sua vez, na posição $d = +1$, destacam-se os dinucleótidos TA e TG como preferidos e o AT, AA e AC como preteridos.

Para a transição, $C \leftrightarrow T$, em $d = +1$ é o dinucleótido GG que tem uma frequência mais aumentada, seguindo-se os dinucleótidos GT, GC, GA, no entanto, nessa posição, a variação evita a ocorrência do dinucleótido AA. Para a posição $d = -1$, tem-se TA e CA como os dinucleótidos com o maior viés positivo e AT com o menor. No caso das transversões, de uma forma geral os valores do viés face ao genoma são menores do que nas transições. Observa-se que para $A \leftrightarrow T$, em praticamente todas as posições da vizinhança os perfis de frequência dos dinucleótidos AA, TT, TA e AT são superiores às do genoma. Esta variação evita a ocorrência dos dinucleótidos de complemento-invertido CT e AG, nas posições imediatamente adjacentes à variação. Para a transição $C \leftrightarrow G$, constata-se que o viés dos dinucleótidos de complemento-invertido AG/CT e GA/TC em posições simétricas ($d = -1$ e $d = +1$) é muito similar. Observa-se também que normalmente, os perfis de frequência de AA, TT, AT, TA são inferiores ao que era esperado.

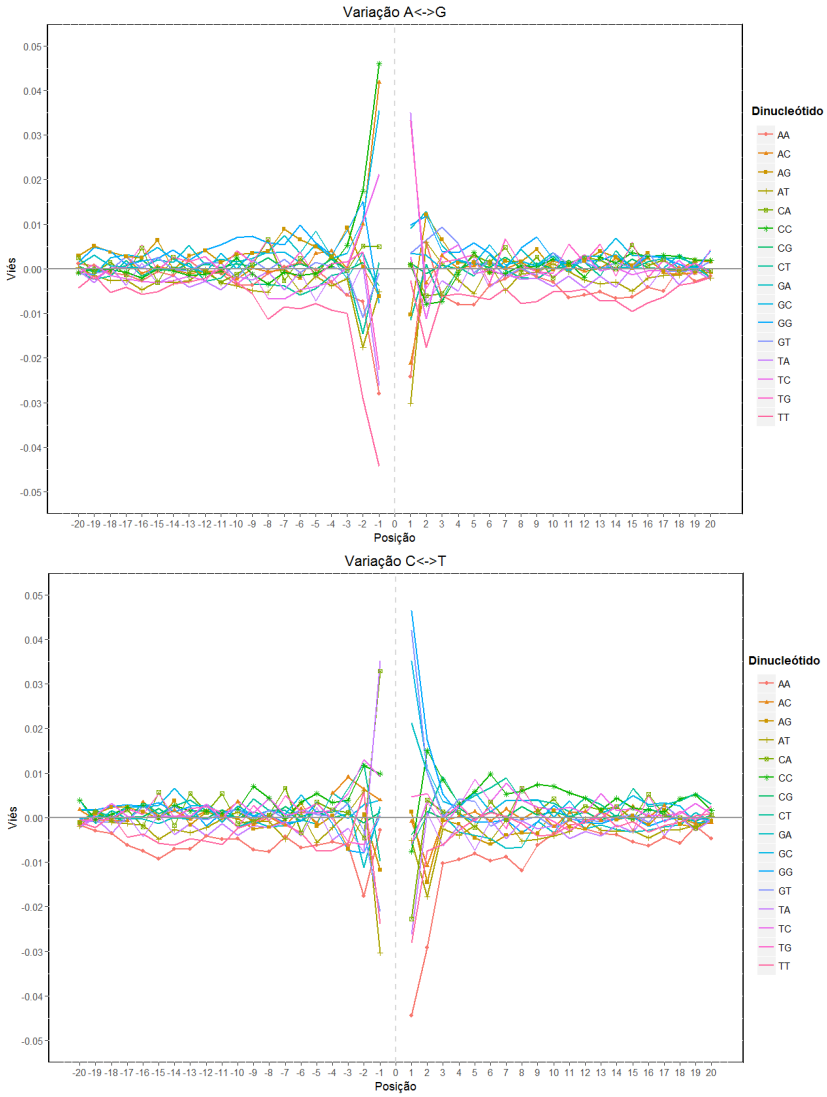


Figura 4.8: Padrões de frequência dos dinucleótidos ($k = 2$) para as transições $A \leftrightarrow G$ e $C \leftrightarrow T$, numa vizinhança de amplitude $w = 20$.

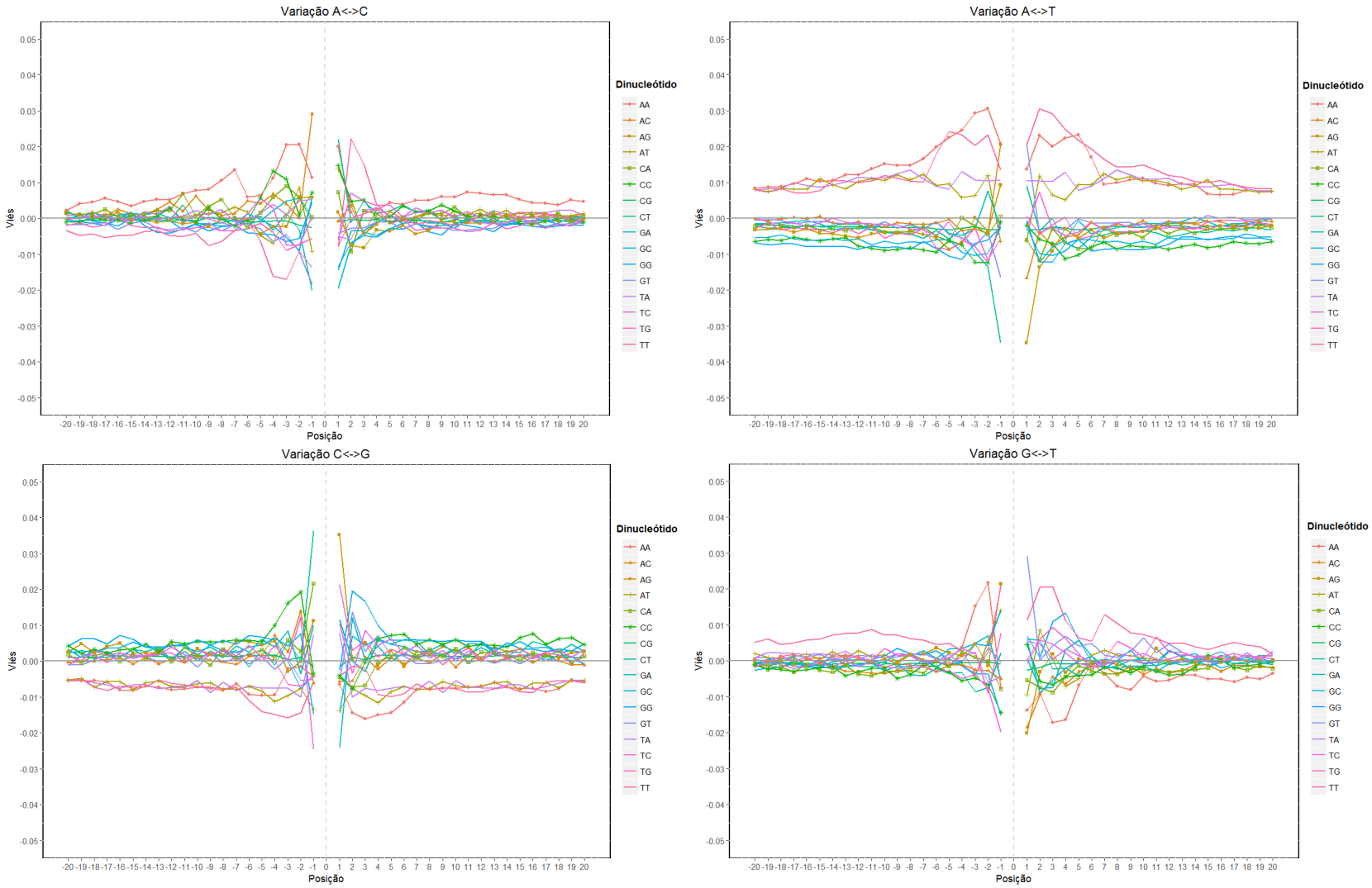


Figura 4.9: Padrões de frequência dos dinucleótidos ($k = 2$) para as transversões $A \leftrightarrow C$, $A \leftrightarrow T$, $C \leftrightarrow G$ e $G \leftrightarrow T$, numa vizinhança de amplitude $w = 20$.

Através dos gráficos dos dinucleótidos, apresentados nas figuras 4.8 e 4.9, pode-se aferir, mais uma vez, que é nas posições adjacentes ao local de variação ($d = \pm 1, \ldots, \pm 4$) da vizinhança que ocorrem as maiores disparidades entre as frequências dinucleotídicas e as do genoma, sendo mais visível nas transições. Confirma-se que cada tipo de SNV tem um contexto específico de ocorrência de dinucleótidos e que existem padrões distintos nas sequências à esquerda e à direita do local de cada SNV.

À semelhança do que foi feito para os nucleótidos, também foi apurado se as frequências dos dinucleótidos se distribuem uniformemente ao longo da vizinhança cada SNV, através da aplicação do teste de ajustamento do χ^2. O teste foi aplicado aos dinucleótidos de cada SNV, considerando as 40 posições da vizinhança $d = \pm 1 \ldots \pm 20$. A tabela 4.8 mostra apenas a medida do efeito ϕ, correspondente a cada dinucleótido. Os respetivos valores da estatística χ^2 encontram-se em apêndice na tabela A.18. Foi rejeitada a hipótese nula para todos os dinucleótidos, pois obteve-se sempre o valor-$p \approx 0.000$.

	SNV					
Dinucleótidos	$A \leftrightarrow C$	$A \leftrightarrow G$	$A \leftrightarrow T$	$C \leftrightarrow G$	$C \leftrightarrow T$	$G \leftrightarrow T$
AA	25.93	68.84	30.36	17.57	89.25	37.82
AC	35.73	**120.20**	32.46	24.43	46.20	23.77
AG	17.17	49.00	35.69	36.59	48.39	32.09
AT	21.85	71.33	20.94	13.77	71.51	22.11
CA	18.82	36.72	13.54	26.61	88.94	13.92
CC	28.69	**122.76**	16.01	32.32	56.58	21.72
CG	8.33	39.49	8.16	11.61	39.30	8.77
CT	32.67	48.89	35.57	37.71	48.63	17.61
GA	26.74	46.09	18.18	35.70	71.50	20.19
GC	24.93	93.68	14.33	35.34	93.26	25.48
GG	20.87	56.53	16.45	32.03	**123.18**	28.18
GT	23.92	46.38	32.27	23.64	**120.29**	35.41
TA	13.30	**103.79**	9.35	24.69	**103.66**	13.23
TC	20.66	71.93	18.09	36.53	46.51	27.32
TG	14.18	89.45	13.77	25.91	36.91	18.85
TT	37.26	89.39	30.00	17.62	67.77	25.57

Tabela 4.8: Valores do coeficiente ϕ do teste de ajustamento do χ^2, aplicado aos dinucleótidos de cada SNV, considerando as posições da vizinhança $d = \pm 1 \ldots \pm 20$.

Na tabela 4.8 observa-se que são as transições que apresentam os valores mais elevados de ϕ, correspondentes aos dinucleótidos AC, CC e TA na variação $A \leftrightarrow G$ e aos dinucleótidos GG, GT e TA na variação $C \leftrightarrow T$, estando de acordo com o que foi dito anteriormente. Também se verifica que o dinucleótido CpG tem preferência pelas variações $A \leftrightarrow G$ e $C \leftrightarrow T$, tendo em conta o valor de ϕ que é maior nas transições.

Também nesta análise quando $k = 2$, tanto nas transições como nas transversões, em torno do local da variação, os dinucleótidos de complemento-invertido (AA/TT, AC/GT, AG/CT, AT/AT, CA/TG, CC/GG, CG/CG, GA/TC, GC/GC, TA/TA) parecem exibir padrões de frequência simétricas. Assim sendo, testou-se a homogeneidade das contagens de cada par de dinucleótidos de complemento-invertido ao longo das posições simétricas, quer para as transições quer para as transversões. Foram organizadas tabelas de contingência

correspondentes às contagens dos dinucleótidos de complemento-invertido, fixando as posições $d = \pm 1, \pm 2, \ldots, \pm 20$. Os resultados dos testes de homogeneidade aplicados a cada uma dessas tabelas de contingência estão resumidos na tabela 4.9, na qual se apresentam o valores da estatística χ^2, de V de Cramér e valores-p, com $g.l. = (20-1)(2-1) = 19$, em todo os pares.

	Dinucleótidos Complemento-invertido	χ^2	V	valores-p
Transições	AA/TT	32.841	0.001	≈ 0.025
	AC/GT	12.008	0.000	≈ 0.885
	AG/CT	37.537	0.001	≈ 0.007
	AT/AT	17.064	0.000	≈ 0.586
	CA/TG	33.025	0.001	≈ 0.024
	CC/GG	71.912	0.001	≈ 0.000
	CG/CG	17.437	0.001	≈ 0.560
	GA/TC	12.269	0.000	≈ 0.874
	GC/GC	97.921	0.001	≈ 0.000
	TA/TA	20.274	0.001	≈ 0.378
Transversões	AA/TT	26.482	0.001	≈ 0.117
	AC/GT	35.181	0.001	≈ 0.013
	AG/CT	29.591	0.001	≈ 0.057
	AT/AT	11.181	0.001	≈ 0.918
	CA/TG	20.709	0.001	≈ 0.353
	CC/GG	24.438	0.001	≈ 0.180
	CG/CG	21.502	0.002	≈ 0.310
	GA/TC	8.982	0.001	≈ 0.974
	GC/GC	25.529	0.001	≈ 0.143
	TA/TA	22.168	0.001	≈ 0.276

Tabela 4.9: Resultados dos testes de homogeneidade entre as contagens de dinucleótidos de complemento-invertido em posições simétricas ao longo da vizinhança, para as transições e transversões.

A partir da tabela 4.9 pode-se aferir que nas transições, para metade dos dinucleótidos de complementos-invertidos, as frequências distribuem-se de forma homogénea ao longo das posições simétricas da vizinhança. Rejeitou-se a hipótese de homogeneidade no caso dos pares AA/TT, AG/CT, CA/TG, CC/GG e GC/GC. Em apêndice, na figura A.7, exemplificam-se os *heatmaps* para os pares AA/TT e CA/TG. No caso das transversões, a hipótese de homogeneidade só foi rejeitada no par AC/GT. Também se constata que em todos os casos, a força da associação é muito fraca. Conclui-se então, que para as transições e transversões, na maioria dos casos, existe homogeneidade entre as frequências dos pares de dinucleótidos de complemento-invertido, ao longo das posições simétricas.

Em seguida, nas figuras 4.10 e 4.11 podem observar-se os padrões de frequência dos trinucleótidos ($k = 3$), para as transições e transversões, considerando uma vizinhança de amplitude $w = 10$. Em cada SNV, calculou-se o viés entre as frequências relativas dos trinucleótidos e as respetivas frequências relativas no genoma, para cada uma das posições $d = \pm 1, \pm 2, \ldots, \pm 10$. Os padrões de frequência globais para as transições e transversões, encontram-se em apêndice, na figura A.8. Conclui-se também para o caso dos trinucleótidos,

que o viés das palavras de tamanho $k = 3$ relativamente às frequências relativas no genoma, é maior nas posições imediatamente adjacentes e que esse viés vai-se aproximando de zero à medida que a posição d se afasta do local de variação. Salienta-se o facto de que o maior viés observado para o caso dos trinucleótidos (≈ 0.025) é inferior aos valores máximos observados nos casos $k = 1$ e $k = 2$. Para além disso, ao contrário do que aconteceu nas situações anteriores, o viés máximo quando $k = 3$ ocorre mais precisamente nas posições $d = -2$ e $d = +2$ da transição $A \leftrightarrow T$. Na posição $d = -1$ a transição $A \leftrightarrow G$ manifesta uma preferência pelos trinucleótidos CAC, CCC, GCC, GGC e evita a ocorrência de AAA, ATT, TTT. Por outro lado, na posição $d = 1$, são os trinucleótidos TAA, TAT, TGA que apresentam uma frequência aumentada. Para a outra transição, $C \leftrightarrow T$, destaca-se o trinucleótido AAA por ser o que tem o maior viés, na posição $d = +1$, e porque na maioria das posições à direita da variação, a sua frequência foi inferior ao esperado.

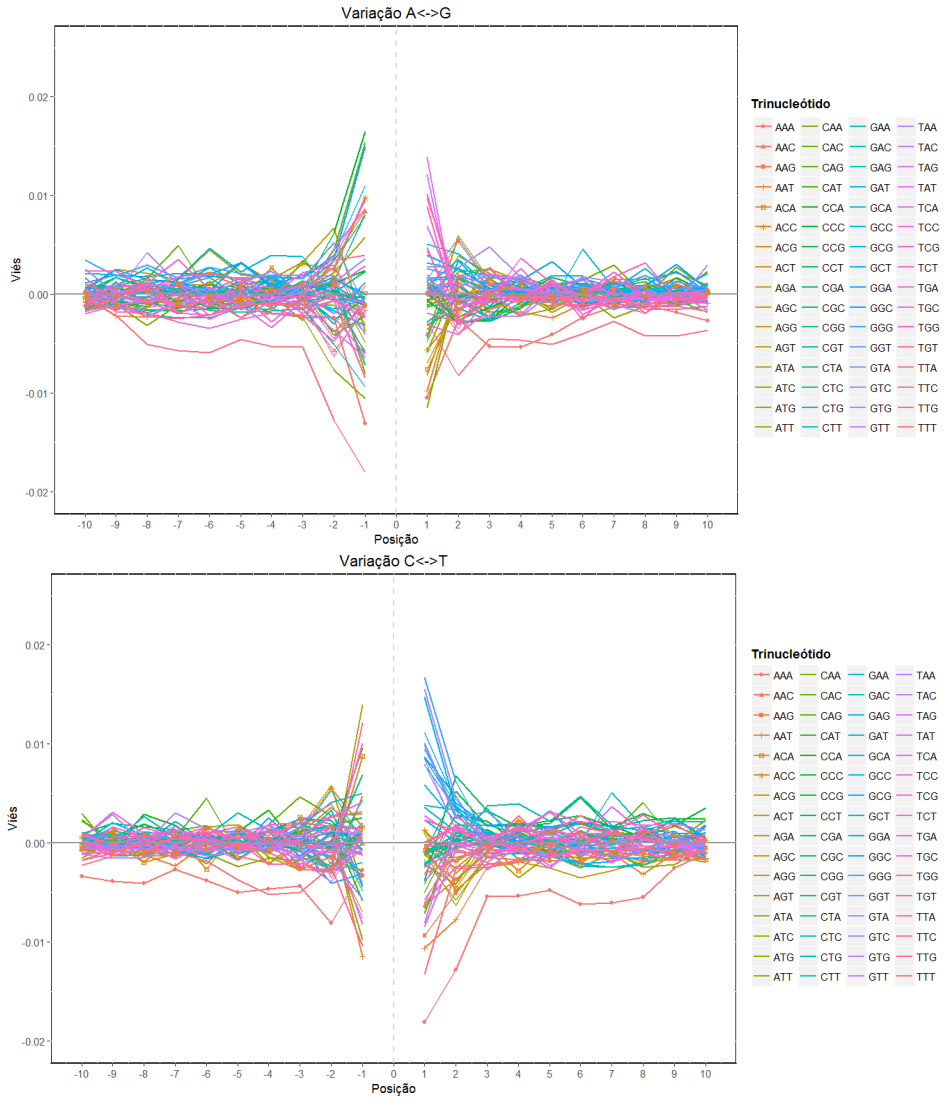


Figura 4.10: Padrões de frequência dos trinucleótidos ($k = 3$) para as transições $A \leftrightarrow G$ e $C \leftrightarrow T$, numa vizinhança com amplitude $w = 10$.

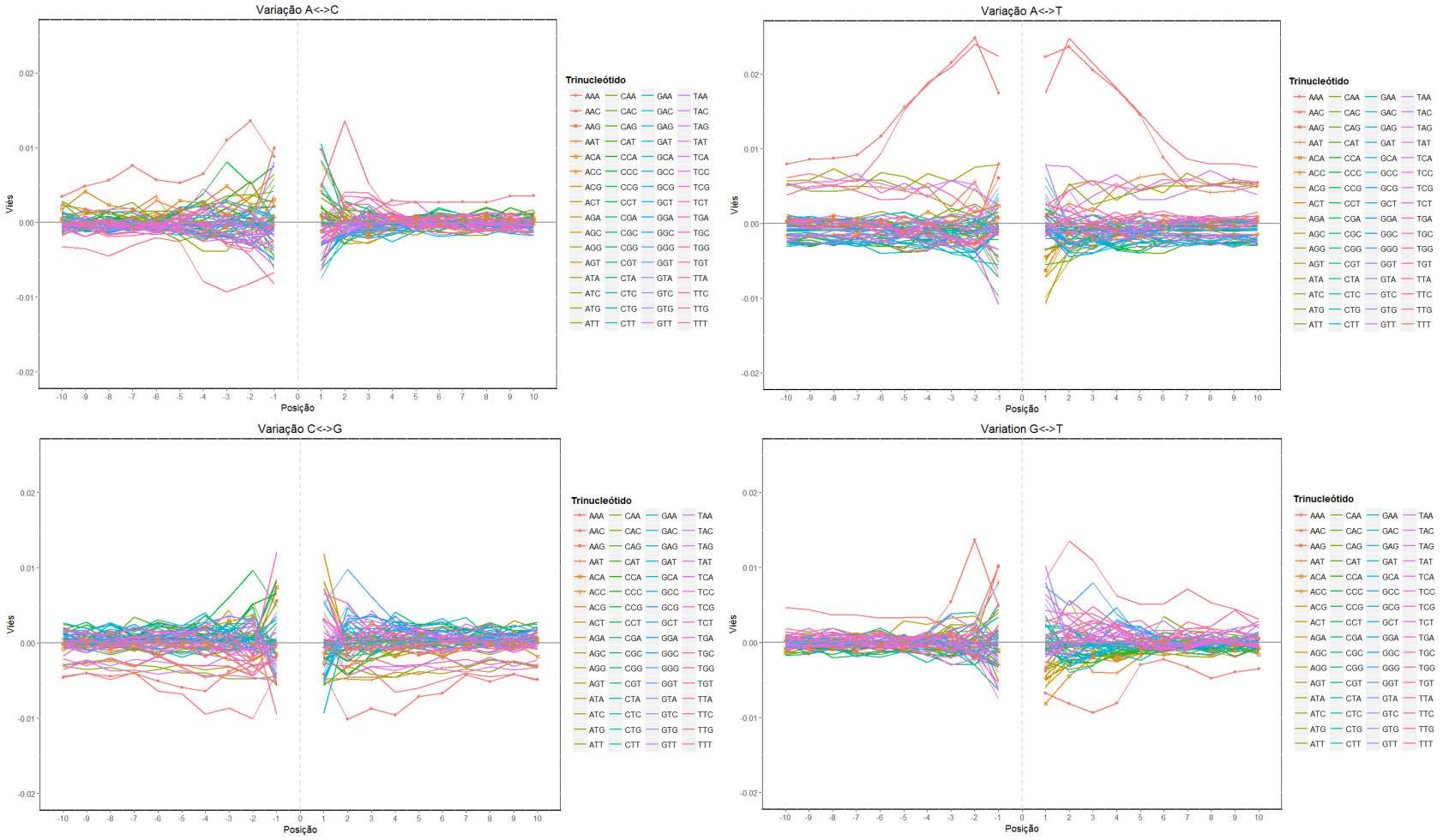


Figura 4.11: Padrões de frequência dos trinucleótidos ($k = 3$) para as transversões $A \leftrightarrow C$, $A \leftrightarrow T$, $C \leftrightarrow G$ e $G \leftrightarrow T$, numa vizinhança com amplitude $w = 10$.

No que diz respeito aos perfis de frequência das transversões, é a SNV $A \leftrightarrow T$ que apresenta um comportamento mais díspar das restantes. Nesta variação os trinucleótidos AAA e TTT têm um viés muito alto em praticamente todas as posições da vizinhança de amplitude $w = 10$, sendo mais notória a discrepância nas posições $d = \pm 2$ e $d = \pm 3$. Na variação $A \leftrightarrow C$, constata-se que nas posições à esquerda do local de variação as frequências do trinucleótido AAA são acima das do genoma, ao contrário do que acontece com o trinucleótido TTT. Na posição $d = +2$, existe uma preferência desta SNV pela ocorrência do trinucleótido TTT. Observa-se que os perfis de frequência da variação $G \leftrightarrow T$ parecem ser simétricos aos da SNV $A \leftrightarrow C$. Verifica-se ainda que para a transversão $C \leftrightarrow G$, os trinucleótidos preferidos são AGA e TCT, nas posições $d = +1$ e $d = -1$, respetivamente. Os trinucleótidos AAA e TTT, correspondem aos trinucleótidos preteridos nas posições $d = +2$ e $d = -2$, respetivamente.

Foi também realizado o teste de ajustamento do χ^2 às frequências dos trinucleótidos para averiguar se estas se distribuem uniformemente ao longo da vizinhança de cada SNV, com amplitude $w = 10$. Constatou-se que em todos os casos foi rejeitada a hipótese nula pois obteve-se valor-$p \approx 0.000$. Os valores para a medida do efeito ϕ por trinucleótido e em cada SNV, encontram-se em anexo na tabela A.19. Verifica-se que é nas transições que se observa a maior força do efeito ϕ. Destacam-se os trinucleótidos CCC ($\phi \approx 120.52$) e GGC ($\phi \approx 94.54$) na variação $A \leftrightarrow G$ e GGG ($\phi \approx 121.63$) e TTA ($\phi \approx 81.82$) na variação $C \leftrightarrow T$.

Também, no caso em que $k = 3$, os trinucleótidos que são complementos invertidos entre si parecem exibir padrões de frequência simétricas ao longo das posições $d = \pm 1, \pm 2, \ldots, \pm 10$, tal como se pode confirmar pela figura A.8.

4.3 Análise do contexto em torno de cada variação e por grupo de prevalência

Surgiu ainda a necessidade de avaliar se para cada tipo de SNV, o contexto da vizinhança é homogéneo em cada um dos grupos de prevalência.

Para cada SNV, as contagens de nucleótidos, dinucleótidos e trinucleótidos ($k = 1, 2, 3$) por grupo de prevalência foram organizadas em tabelas de contingência. Para além disso consideraram-se vizinhanças de diferentes amplitudes ($w = 5, 10, 20, 50, 100, 200$).

A figura 4.12 apresenta os gráficos de barras das frequências relativas da contagem dos nucleótidos por grupo de prevalência e para cada uma das SNV, para uma amplitude $w = 10$. Foram também feitos gráficos análogos para as contagens dos dinucleótidos ($k = 2$) os quais se encontram em apêndice, na figura A.9.

No que diz respeito ao contexto da vizinhança de cada SNV, os padrões apresentados nos grupos de prevalência, são semelhantes aos que foram apresentados anteriormente na seção 4.1. Independentemente de as SNVs serem muito ou pouco prevalentes, as preferências relativamente à contagem de nucleótidos, dinucleótidos e trinucleótidos mantêm-se. Por exemplo, os nucleótidos A e T, continuam a ser os mais frequentes para todas as SNVs, quer sejam muito raras ou comuns. Pela observação dos gráficos, contata-se que fixando a variação, o contexto na vizinhança de cada variação é semelhante entre os grupos de prevalência. Note-se que, em semelhança ao que já foi apresentado, nesta análise do contexto por grupo de prevalência, também se observa que à medida que aumenta a amplitude da vizinhança (de $w = 5$ para $w = 200$), as frequências relativas das palavras de tamanho $k = 1, 2, 3$, tendem a aproximar-se das frequências relativas do genoma.

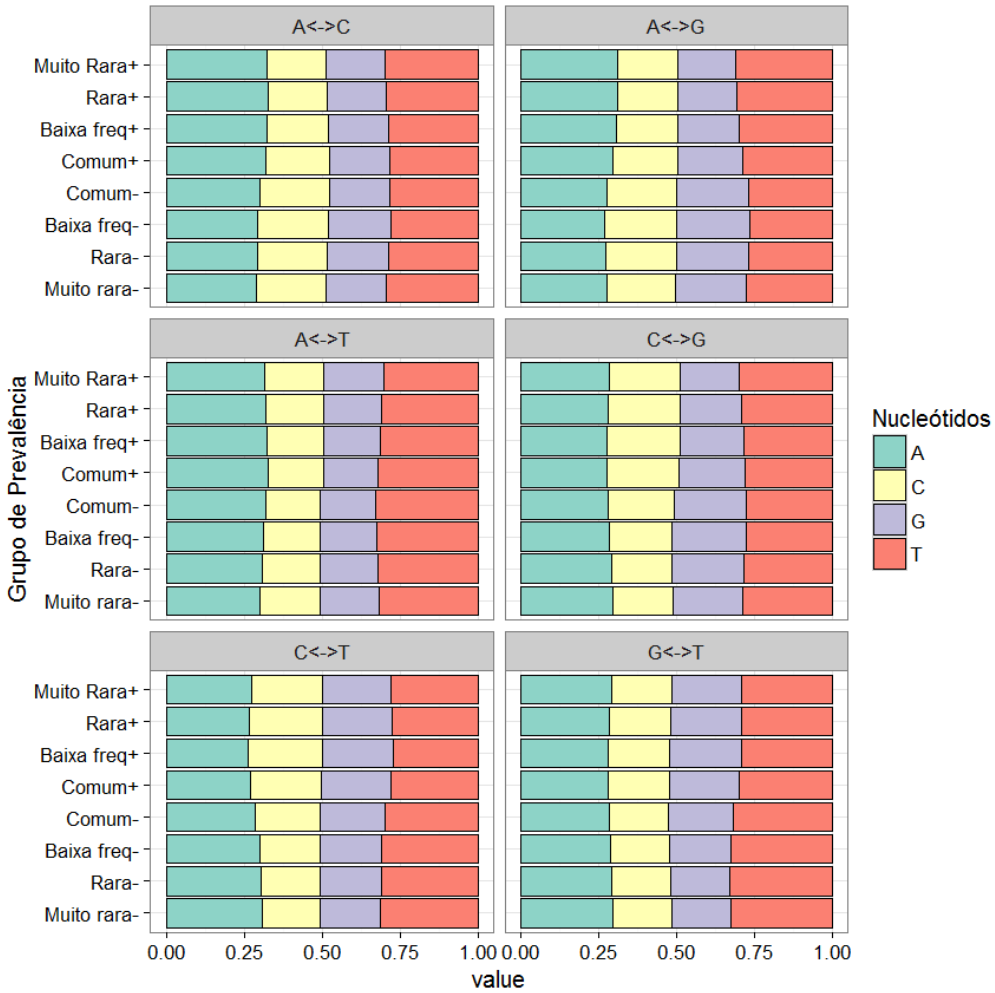


Figura 4.12: Gráficos de barras das frequências relativas da contagem de nucleótidos ($k = 1$) na vizinhança de cada SNV, por grupo de prevalência, considerando uma amplitude $w = 10$.

Para averiguar se existe associação entre as contagens das palavras de tamanho $k = 1, 2, 3$ e os grupos de prevalência, fixando-se a variação, procedeu-se ao teste de independência, considerando as diversas amplitudes das vizinhanças $w = 5, 10, 20, 50, 100, 200$. Na tabela 4.10 apresentam-se os resultados dos testes de independência para cada um dos casos. Esses resultados correspondem à estatística de teste do χ^2, ao valor V de Cramér, ao valor-p e aos graus de liberdade ($g.l.$). Note-se que como são 8 grupos de prevalência, tem-se por exemplo, para o caso dos nucleótidos, $g.l. = (8 - 1)(4 - 1) = 21$. Verifica-se que a hipótese de independência é rejeitada em todos os casos pois obteve-se sempre o valor-$p \approx 0.000$, no entanto, os valores da força de associação são muito pequenos. Constata-se também, nesta analise por grupo de prevalência, que quanto maior for a amplitude da vizinhança w menor é a força da associação entre as contagens e o grupos de prevalência. Só se verificam diferenças significativas nas vizinhanças com amplitude mais pequena ($w = 5$ e $w = 10$), pois, tal como já foi referido, é nas posições muito próximas do local do SNV que a distribuição das palavras é mais díspar da obtida para o genoma. Por outro lado também se observa que, fixando a

amplitude da vizinhança w, a força da associação é maior no caso dos trinucleótidos ($k = 3$), pressupondo que é este tipo de contexto que poderá ter uma maior influência na ocorrência de SNVs. Para além disso, em todas as amplitudes da vizinhança ($w = 5, 10, 20, 50, 100, 200$) e para todos os contextos ($k = 1, 2, 3$), aferiu-se que a força da associação entre o grupo de prevalência e o contexto, é maior para as transições do que nas transversões.

		$k = 1$		$k = 2$		$k = 3$		
w	SNV	χ^2	V	χ^2	V	χ^2	V	valores-p
5	A↔C	73355	0.029	177171	0.032	240471	0.044	≈ 0.000
	A↔G	776862	0.045	1182272	0.041	1384938	0.051	
	A↔T	27072	0.019	106742	0.028	174725	0.041	
	C↔G	129085	0.037	130283	0.027	151629	0.034	
	C↔T	754847	0.045	1156059	0.041	1356688	0.051	
	G↔T	73208	0.029	176573	0.032	237834	0.044	
10	A↔C	137227	0.028	288758	0.028	410239	0.035	≈ 0.000
	A↔G	1211061	0.040	2151822	0.037	2859158	0.045	
	A↔T	40045	0.016	127248	0.020	235882	0.029	
	C↔G	193908	0.032	268109	0.026	339800	0.031	
	C↔T	1184657	0.040	2112190	0.037	2809203	0.045	
	G↔T	135806	0.028	285971	0.028	404353	0.035	
20	A↔C	200871	0.024	426167	0.023	634083	0.029	≈ 0.000
	A↔G	1717942	0.034	3352717	0.032	4773640	0.039	
	A↔T	43960	0.012	130678	0.014	266321	0.020	
	C↔G	208900	0.024	327340	0.020	453423	0.024	
	C↔T	1691061	0.033	3308832	0.031	4713019	0.039	
	G↔T	197145	0.023	418913	0.023	622067	0.029	
50	A↔C	349980	0.020	751390	0.019	1141300	0.024	≈ 0.000
	A↔G	3089100	0.029	6420200	0.027	9491200	0.033	
	A↔T	50156	0.008	142010	0.009	314430	0.014	
	C↔G	252060	0.016	457560	0.015	684780	0.018	
	C↔T	3049800	0.028	6345500	0.027	9385500	0.033	
	G↔T	346720	0.020	744670	0.019	1130300	0.024	
100	A↔C	504740	0.017	1097100	0.016	1684800	0.020	≈ 0.000
	A↔G	4469300	0.024	9556900	0.023	14394000	0.029	
	A↔T	66619	0.007	183970	0.007	401920	0.011	
	C↔G	293580	0.012	586820	0.012	920520	0.015	
	C↔T	4410600	0.024	9436300	0.023	14217000	0.029	
	G↔T	504900	0.017	1098100	0.016	1686000	0.020	
200	A↔C	685970	0.014	1516600	0.013	2345100	0.017	≈ 0.000
	A↔G	6361000	0.021	13842000	0.020	20997000	0.025	
	A↔T	101490	0.006	272980	0.006	558050	0.009	
	C↔G	335560	0.009	724960	0.009	1175500	0.012	
	C↔T	6276700	0.020	13659000	0.020	20720000	0.024	
	G↔T	692230	0.014	1529700	0.014	2360800	0.017	
g.l.		21		105		441		

Tabela 4.10: Resultados dos testes de independência entre a ocorrência de oligonucleótidos e o grupo de prevalência numa vizinhaça de cada SNV, considerando os comprimentos de oligonucleótidos $k = 1, 2, 3$, os vários tipos de SNV e diversas amplitudes para a vizinhança ($w = 5, 10, 20, 50, 100, 200$).

Com a rejeição de H_0 pressupõe-se a realização de uma análise aos resíduos. Em seguida, apenas para a amplitude $w = 10$, em cada um dos casos $k = 1, 2, 3$, apresentam-se os *heatmaps* dos resíduos, para a transição A↔G e a transversão, A↔T, por exemplo. Os *heatmaps* para os nucleótidos, dinucleótidos e trinucleótidos estão representados nas figuras 4.13, respetivamente. Os *heatmaps* apresentados, refletem o efeito da força da associação, maior para A↔G, pois há claramente uma preferência dos grupos de prevalência, por certos nucleótidos, dinucleótidos e trinucleótidos, conforme o caso, ao contrário do que acontece com A↔T. Note-se que por exemplo, no *heatmap* quando $k = 3$, para a variação A↔G, é possível observar-se uma certa simetria pois os grupos de prevalência ficaram agrupados em dois grandes grupos ("+" e "−"), tal como já era esperado. No entanto, para a variação A↔T a maioria das classes dos grupos de prevalência foram agrupadas duas a duas havendo uma distinção evidente entre as preferências das muito raras+ e comuns+, por exemplo.

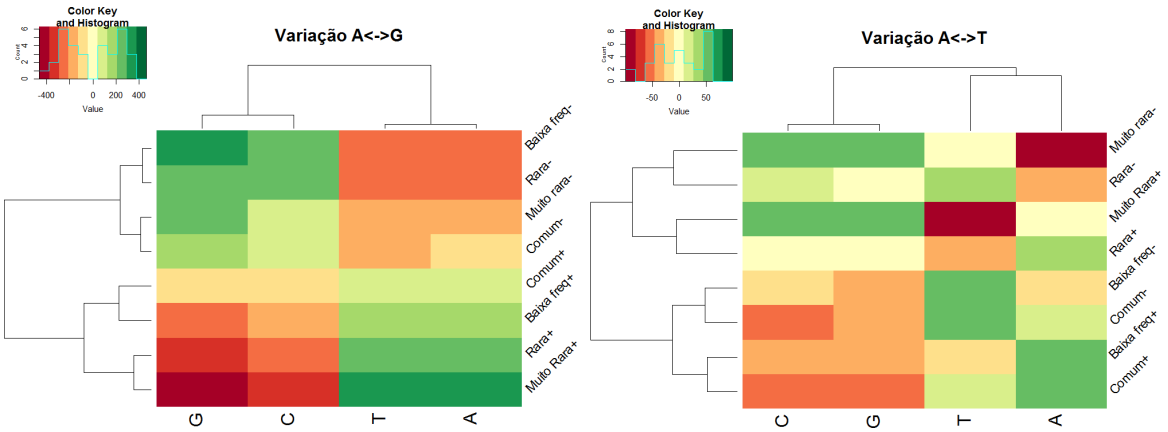


Figura 4.13: *Heatmaps* dos resíduos ajustados do teste de independência entre os nucleótidos ($k = 1$) e os grupos de prevalência, para a transição A↔G e a transversão A↔T, considerando uma vizinhança de amplitude $w = 10$.

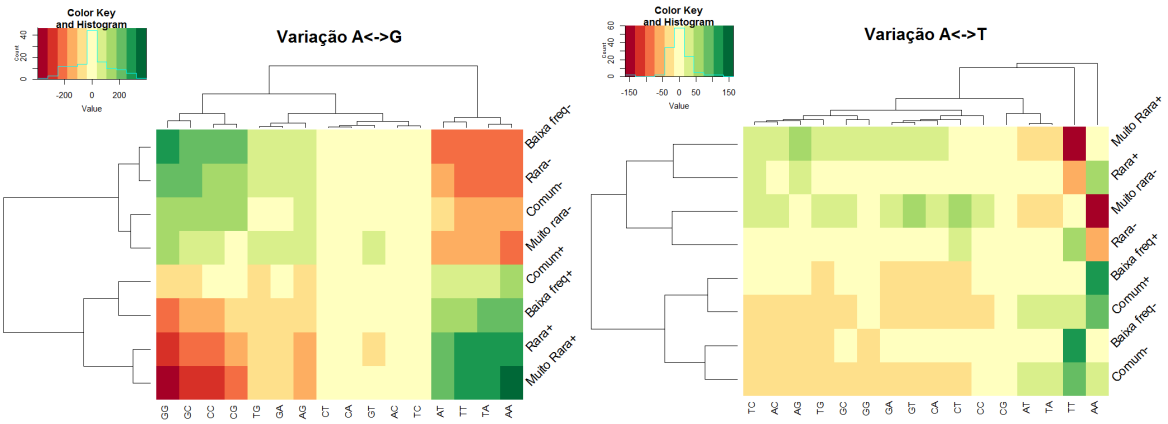


Figura 4.14: *Heatmaps* dos resíduos ajustados do teste de independência entre os dinucleótidos ($k = 2$) e os grupos de prevalência, para a transição A↔G e a transversão A↔T, considerando uma vizinhança de amplitude $w = 10$.

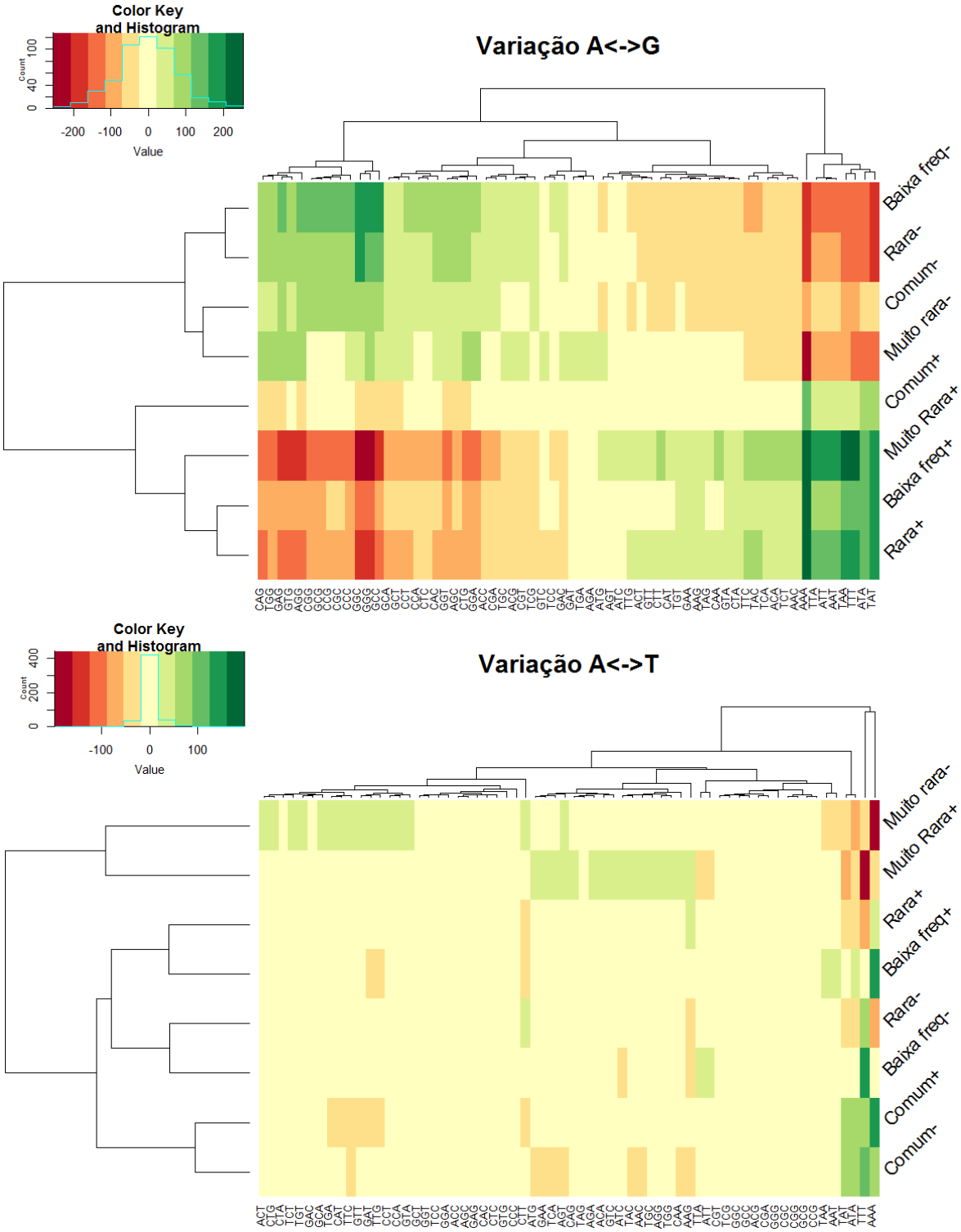


Figura 4.15: *Heatmaps* dos resíduos ajustados do teste de independência entre os trinucleótidos ($k = 3$) e os grupos de prevalência, para a transição A↔G e a transversão A↔T, considerando uma amplitude da vizinhança $w = 10$.

Capítulo 5

Conclusões e trabalho futuro

Ao longo desta dissertação foram apresentados alguns resultados do estudo de aproximadamente 36.8 milhões de SNVs, provenientes do genoma de 1092 indivíduos considerados na fase 1 do projeto 1000 Genomas. O objectivo do trabalho consistiu em caracterizar essas SNVs, relativamente ao contexto na vizinhança da posição onde ocorre cada SNV através de perfis de frequência, de modo a encontrar um padrão que seja indicador do fenómeno. Estes perfis de frequência consistem na contagem de nucleótidos, dinucleótidos e trinucleótidos, isto é, palavras de tamanhos $k = 1, 2, 3$ considerando diversas amplitudes para a vizinhança da variação, como por exemplo $w = 5, 10, 20, 50, 100, 200$. No entanto, iniciou-se essa caracterização com um estudo mais global para averiguar a forma como as SNVs se distribuem ao longo do genoma e de acordo com a sua prevalência, assim como, avaliar se os genótipos associados a cada SNV satisfazem o HWE.

Para aplicar aos dados as metodologias estatísticas apresentadas, foi necessário utilizar e desenvolver várias ferramentas de análise estatística baseadas no software R.

Verificou-se que, ao longo do genoma, as transições ocorrem mais frequentemente do que as transversões. Em seguida, ao organizar as ocorrências de cada tipo de SNV por cromossoma e por grupo de prevalência e aplicar os respetivos testes de homogeneidade, foi rejeitada a hipótese de homogeneidade mas com valores da força da associação muito baixos, levando a concluir que as SNVs se distribuem de forma homogénea ao longo do genoma. Por outro lado, o teste de Kolmogorov-Smirnov para duas amostras independentes permitiu concluir que as prevalências calculadas para cada tipo de SNV provêm de populações com a mesma distribuição. No estudo efetuado à prevalência das variações, constatou-se que a maioria das SNVs são raras ou de baixa prevalência.

Relativamente ao HWE, averiguou-se que para a maioria das SNVs, as frequências dos respetivos genótipos estão de acordo com a lei de Hardy-Weinberg, chegando-se à mesma conclusão quando a análise é feita por cromossoma. Devido ao número elevado de SNVs, na avaliação do HWE, foi necessário optar por metodologias estatísticas mais simples, nomeadamente o teste de ajustamento do χ^2, uma vez que os testes exatos são computacionalmente exaustivos.

Ao proceder à caracterização do contexto da vizinhança da SNV de uma forma global, observou-se que as frequências relativas eram tanto mais próximas das correspondentes no genoma, quanto maior fosse a amplitude da vizinhança. Para todas as vizinhanças w, foi rejeitada a hipótese de independência entre as contagens das palavras de comprimento k e as SNVs. No entanto, apesar da força da associação ser muito pequena para todas as amplitudes

consideradas, observou-se que essa força da associação era maior no caso dos trinucleótidos, $k = 3$, fazendo supor que é este tipo de contexto que poderá ter uma maior influência na ocorrência de SNVs. Verificou-se também através da construção dos dendrogramas que as SNVs correspondentes a transições e transversões, nem sempre ficavam agrupadas, revelando que cada SNV tem uma preferência específica do contexto. Passando para uma análise mais pormenorizada, para cada uma das SNVs, procedeu-se à contagem de nucleótidos, dinucleótidos e trinucleótidos, localizados d nucleótidos à direita e à esquerda de cada local de variação. Foi verificado em todos os casos que, à medida que a posição d se afasta do local de variação, as frequências relativas das palavras tendem a aproximar-se das frequências relativas no genoma e que é nas posições muito próximas do local da SNV que se apresentam as maiores discrepâncias. Os padrões de frequência determinados pelo viés das contagens de cada palavra de tamanho k em cada uma das posições, caracterizam o contexto da SNV e são específicos de cada variação, principalmente no que diz respeito às posições imediatamente adjacentes ao local de SNV. Constata-se que são as transições que apresentam os maiores vieses das frequências de cada nucleótido em relação às frequências no genoma. Para além disso, os pares de nucleótidos, dinucleótidos e trinucleótidos que são complementos invertidos entre si exibem padrões de frequência simétricos.

Realizando o mesmo tipo de análise do contexto da vizinhança de acordo com os grupos de prevalência das SNVs, chegam-se a conclusões semelhantes.

Como trabalho futuro, poder-se-á estender este tipo de análise às SNVs disponibilizadas na fase 3 do P1000G. Outra ideia será considerar todas as posições imediatamente adjacentes a um local de SNV, com nucleótidos C e G, e averiguar se há um aumento significativo da taxa de transições, devido à influência de estruturas CpG. Para além disso, também será interessante explorar e aplicar modelos que estimam as probabilidades de substituição e confirmar por outra via que o contexto em que se considera os trinucleótidos é o que mais influencia a ocorrência das substituições de nucleótido único (SNVs). Também será potencialemente interessante repetir o mesmo tipo de análises, mas diferenciando agora o genoma em zonas codificantes e não codificantes. Finalmente, justifica-se ainda fazer o estudo diferenciado dos padrões de frequência na vizinhança de SNVs consoante verifiquem o HWE, ou não.

Bibliografia

[1000-Genomes-Project, 2016] 1000-Genomes-Project (Consulta fevereiro de 2016). `http://www.1000genomes.org/`.

[1000Genomes, 2010] 1000Genomes, C. P. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073.

[1000Genomes, 2012] 1000Genomes, C. P. (2012). An integrated map of genetic variation from 1092 human genomes. *Nature*, 491:56–65.

[1000Genomes, 2015a] 1000Genomes, C. P. (2015a). A global reference for human genetic variation. *Nature*, 526:68–74.

[1000Genomes, 2015b] 1000Genomes, C. P. (2015b). An integrated map of structural variation in 2504 human genomes. *Nature*, 526:75–81.

[Afreixo et al., 2006] Afreixo, V., Pinheiro, M., Moura, G., and et al. (2006). Statistical, computational and visualization methodologies to unveil gene primary structure features. *Methods Inf Med.*, 45(2):163–168.

[Agresti, 2007] Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. Wiley Series in Probability and Statistics. JohnWiley & Sons, Inc., second edition.

[Agresti, 2010] Agresti, A. (2010). *Analysis of Ordinal Categorical Data*. Wiley Series in Probability and Statistics. JohnWiley & Sons, Inc., second edition.

[Bioconductor, 2016] Bioconductor (Consulta junho de 2016). `https://www.bioconductor.org/`.

[Bogacka, 2016] Bogacka, B. (Consulta junho de 2016). *Kolmogorov-Smirnov Tests*. `http://www.maths.qmul.ac.uk/~bb/CTS_Chapter3_Students.pdf`.

[Carlson et al., 2015] Carlson, M., Obenchain, V., and et al. (Consulta fevereiro de 2015). *Intermediate R/Bioconductor for Sequence Analysis*. `http://bioconductor.org/help/course-materials/2013/SeattleFeb2013/IntermediateSequenceAnalysis2013.pdf`.

[Chargaff, 1950] Chargaff, E. (1950). Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia*, 6(6):201–209.

[Cramér, 1946] Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.

[dbSNP, 2016] dbSNP (Consulta abril de 2016). http://www.ncbi.nlm.nih.gov/projects/SNP/.

[Draghici, 2011] Draghici, S. (2011). *Statistics and Data Analysis for Microarrays Using R and Bioconductor*. Chapman Hall/CRC Mathematical and Computational Biology. Chapman and Hall/CRC, 2nd edition.

[Everitt, 1977] Everitt, B. (1977). *The Analysis of Contingency Tables*. Springer-Science+Business Media, B.V.

[Everitt and et al., 2011] Everitt, B. and et al. (2011). *Cluster analysis*. Wiley, 5th edition.

[Everitt and Hothorn, 2011] Everitt, B. and Hothorn, T. (2011). *An Introduction to Applied Multivariate Analysis with R*. Springer-Verlag New York, 1st edition.

[Feero et al., 2010] Feero, W. G., Guttmacher, A. E., and Collins, F. S. (2010). Genomic Medicine - An Updated Primer. *N Engl J Med*, 362:2001–2011.

[Foulkes, 2009] Foulkes, A. S. (2009). *Applied Statistical Genetics with R: For Population-based Association Studies*. Use R. Springer-Verlag New York.

[Gentleman, 2009] Gentleman, R. (2009). *R Programing for Bioinformatics*. Chapman & Hall/CRC.

[Gibbons and Chakraborti, 2010] Gibbons, J. D. and Chakraborti, S. (2010). *Nonparametric Statistical Inference*. CRC Press, 5th edition.

[Gibbons and Pratt, 1981] Gibbons, J. D. and Pratt, J. W. (1981). *Concepts of Nonparametic Theory*. Springer-Verlag New York Inc.

[Goemana and Solari, 2012] Goemana, J. J. and Solari, A. (2012). Tutorial in biostatistics: multiple hypothesis testing in genomics. *Statistics in Medicine*, 0:1–27. http://www.few.vu.nl/~mavdwiel/HDDA/tutorial_multtest.pdf.

[Graffelman, 2015] Graffelman, J. (2015). Exploring diallelic genetic markers: The Hardy-Weinberg package. *Journal of Statistical Software*, 64(3):1–23.

[Graffelman, 2016] Graffelman, J. (Consulta outubro de 2016). *Package "Hardy-Weinberg"*. https://cran.r-project.org/web/packages/HardyWeinberg/HardyWeinberg.pdf.

[Graffelman and Camarena, 2016] Graffelman, J. and Camarena, J. (Consulta outubro de 2016). *Hardy-Weinberg Equilibrium and the Ternary Plot*. http://dugi-doc.udg.edu/bitstream/handle/10256/737/GRAFFELMAN_JAN_extended_NEW_16_maig_08.pdf?sequence=4.

[Hitchcock, 2016] Hitchcock, D. B. (Consulta outubro de 2016). *Yates and Contingency Tables: 75 Years Later*. http://people.stat.sc.edu/Hitchcock/yates75tech.pdf.

[Hsu, 1996] Hsu, J. C. (1996). *Multiple Comparisons: Theory and methods*. Springer-Science+Business Media, B.V.

[Jiang et al., 2008] Jiang, Z., Xiao-Lin, W., Zhang, M., and et al. (2008). The complementary neighborhood patterns and methylation-to-mutation likelihood structures of 15.110 Single-Nucleotide Polymorphisms in the bovine genome. *Genetics*, 180(1):639–647.

[Kateri, 2014] Kateri, M. (2014). *Contingency Table Analysis: Methods and Implementation Using R*. Birkhäuser Basel. Statistics for Industry and Technology.

[Kim and Misra, 2007] Kim, S. and Misra, A. (2007). SNP genotyping: technologies and biomedical applications. *Annu. Rev. Biomed. Eng.*, 9:289–320.

[Murteira, 1990] Murteira, B. J. (1990). *Probabilidades e Estatística. Volume II.* McGraw-Hill, 2nd edition.

[NCBI, 2016] NCBI (Consulta fevereiro de 2016). *Human-Sapiens-Genome*. `ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/`.

[NEI, 2016] NEI (Consulta maio de 2016). `https://neuroendoimmune.files.wordpress.com/2014/03/snp.png`.

[NHGRI et al., 2011] NHGRI, Green, E. D., and Guyer, M. S. (2011). Charting a course for genomic medicine from base pairs to bedside. *Nature*, 470:204–213.

[Plyler et al., 2015] Plyler, Z. E., Hill, A. E., McAtee, C. W., and et al. (2015). SNP formation bias in the murine genome provides evidence for parallel evolution. *Genome Biology and Evolution*, 7(9):2506–2519.

[Pollard et al., 2016] Pollard, K. S., Gilbert, H. N., and et al. (Consulta outubro de 2016). *Package "multtest"*. `http://www.bioconductor.org/packages/release/bioc/manuals/multtest/man/multtest.pdf`.

[R-Project, 2016] R-Project (Consulta junho de 2016). `http://cran.r-project.org/`.

[Reece, 2004] Reece, R. J. (2004). *Analysis of genes and genomes*. John Wiley and Sons Inc.

[Regateiro, 2007] Regateiro, F. (2007). *Manual de Genética Médica*. Imprensa da Universidade de Coimbra.

[Reis, 2001] Reis, E. (2001). *Estatística Multivariada Aplicada*. Edições Sílabo, 2nd edition.

[Santner and Duffy, 1989] Santner, T. J. and Duffy, D. E. (1989). *The Statistical Analysis of Discrete Data*. Springer Texts in Statistics. Springer-Verlag New York, 1st edition.

[Sheskin, 2003] Sheskin, D. J. (2003). *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall, third edition.

[Squassina et al., 2010] Squassina, A., Manchia, M., Manolopoulos, V., and et al. (2010). Realities and expectations of pharmacogenomics and personalized medicine: impact of translating genetic knowledge into clinical practice. *Pharmacogenomics*, 11(8):1149–1167.

[Stram, 2014] Stram, D. O. (2014). *Design, Analysis, and Interpretation of Genome-Wide Association Scans*. Statistics for Biology and Health. Springer-Verlag New York.

[UCSC, 2016] UCSC (Consulta fevereiro de 2016). *Genome Browser*. `https://genome.ucsc.edu/`.

[VCF.file, 2016] VCF.file (Consulta fevereiro de 2016). `http://samtools.github.io/hts-specs/VCFv4.1.pdf`.

[Voight and Aggarwala, 2016] Voight, B. F. and Aggarwala, V. (2016). An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nature Genetics*, 48:1–7.

[Watson and Crick, 1953] Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

[Weir, 1996] Weir, B. S. (1996). *Genetic Data Analysis II. Methods for Discrete Population Genetic Data*. Sinauer Associates.

[Winslow, 2016] Winslow, T. (Consulta junho de 2016). *DNA structure*. `http://www.cancer.gov/images/cdr/Live/CDR761781.jpg`.

[Zeggini and Morris, 2015] Zeggini, E. and Morris, A. (2015). *Assessing Rare Variation in Complex Traits. Design and Analysis of Genetic Studies*. Springer-Verlag New York.

[Zhang and Zhao, 2004] Zhang, F. and Zhao, Z. (2004). The influence of neighboring-nucleotide composition on single nucleotide polymorphisms (SNPs) in the mouse genome and its comparison with human SNPs. *Genomics.*, 84(5):785–795.

Apêndice A

Gráficos e tabelas adicionais

Neste apêndice apresentam-se algumas tabelas e gráficos adicionais que servem de complemento aos resultados apresentados nos capítulos 3 e 4.

Cromossoma	$A \leftrightarrow C$	$A \leftrightarrow G$	$A \leftrightarrow T$	$C \leftrightarrow G$	$C \leftrightarrow T$	$G \leftrightarrow T$	Total
1	232542	1004783	191545	234222	1002597	231271	2896960
2	261384	1087808	223066	273536	1083925	260170	3189889
3	221613	907312	190695	223983	899661	220737	2664001
4	224880	886297	201061	206983	889768	225011	2634000
5	204180	826799	177196	200437	827344	202877	2438833
6	192051	797911	165431	185290	799095	190953	2330731
7	173649	729079	146332	185279	725279	172663	2132281
8	177951	703282	150284	200507	703056	176288	2111368
9	130986	534285	108184	154605	537266	129155	1594481
10	146646	625498	120226	150252	626189	145336	1814147
11	151245	622963	124114	153706	625112	150144	1827284
12	143036	608802	119353	142415	604132	141954	1759692
13	110495	453579	97124	101624	447579	109137	1319538
14	97584	419342	82083	101810	412637	97556	1211012
15	86489	372127	69446	101800	372878	86219	1088959
16	93545	393378	70750	135477	386872	92584	1172606
17	74800	358968	56268	86360	356595	74128	1007119
18	85947	361499	73535	82312	358992	85686	1047971
19	57390	278052	41072	71275	276013	58170	781972
20	63392	293298	49529	68263	286525	63946	824953
21	40450	170993	34148	39644	172783	39806	497824
22	34240	170421	23708	43538	169398	34066	475371
Total	3004495	12606476	2515150	3143318	12563696	2987857	**36820992**

Tabela A.1: Número de ocorrências de cada tipo de SNV por cromossoma.

	SNV					
Cromossoma	$A \leftrightarrow C$	$A \leftrightarrow G$	$A \leftrightarrow T$	$C \leftrightarrow G$	$C \leftrightarrow T$	$G \leftrightarrow T$
1	-8.591	16.699	-15.381	-28.662	18.236	-8.528
2	2.349	-5.335	12.012	2.564	-5.557	2.843
3	9.847	-6.391	21.997	-7.823	-12.509	10.636
4	23.248	-20.903	53.583	-40.906	-12.111	26.401
5	12.534	-11.436	27.858	-18.403	-6.724	12.078
6	4.622	-0.094	16.699	-33.131	5.461	4.523
7	-0.875	-1.419	1.905	8.210	-3.388	-0.935
8	14.679	-29.266	17.032	51.406	-25.960	12.876
9	2.604	-19.829	-2.346	53.570	-11.591	-0.682
10	-3.848	7.036	-11.149	-12.582	11.537	-5.225
11	5.942	-4.235	-2.115	-6.205	2.600	5.192
12	-1.553	10.310	-2.594	-21.580	6.041	-2.368
13	9.147	3.375	24.564	-34.970	-4.975	6.696
14	-4.156	9.202	-2.338	-5.196	-1.116	-2.410
15	-8.412	-1.439	-19.041	30.769	2.696	-7.642
16	-7.325	-16.000	-34.777	118.815	-26.196	-8.825
17	-27.232	30.148	-50.167	1.391	27.610	-28.104
18	1.576	5.646	7.663	-25.362	2.955	2.351
19	-26.794	24.876	-55.924	18.490	22.171	-22.117
20	-15.953	25.479	-30.110	-8.613	11.844	-12.215
21	-0.892	1.660	0.808	-14.575	8.790	-3.084
22	-24.259	23.589	-50.712	15.447	22.159	-24.102

Tabela A.2: Resíduos ajustados do teste de homogeneidade entre tipo de SNV e cromossoma.

	SNV					
Grupo	$A \leftrightarrow C$	$A \leftrightarrow G$	$A \leftrightarrow T$	$C \leftrightarrow G$	$C \leftrightarrow T$	$G \leftrightarrow T$
Muito rara−	136.3298	379.5641	18.3264	9.2590	-407.1581	-115.7085
Rara−	102.6930	407.8863	-23.1794	-20.6276	-372.1460	-123.0648
Baixa freq−	80.2465	348.6480	-7.4253	-6.3987	-334.0852	-92.8223
Comum−	44.4747	130.0192	10.7646	6.6403	-148.9136	-28.7056
Comum+	-30.0532	-153.4284	12.5484	8.5722	132.6237	46.1053
Baixa freq+	-94.5529	-333.5434	-9.6425	-0.3528	349.3856	77.0284
Rara+	-121.1664	-370.3505	-23.5113	-14.5666	403.3248	101.3732
Muito Rara+	-118.6989	-409.2089	20.4556	16.5210	379.3062	135.6960

Tabela A.3: Resíduos ajustados do teste de independência entre grupos de prevalência e SNVs.

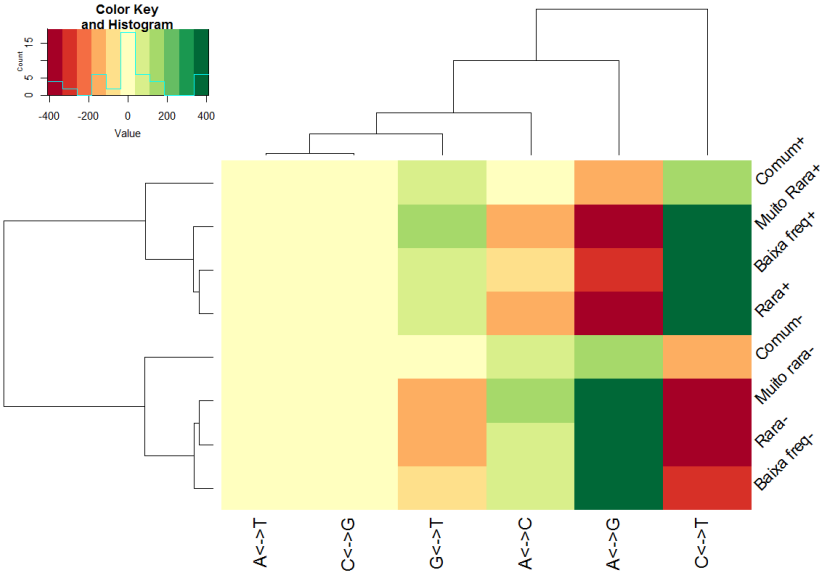


Figura A.1: *Heatmap* dos resíduos ajustados do teste de independência entre grupos de prevalência e cada tipo de SNV.

		SNV					
w	$k = 1$	$A \leftrightarrow C$	$A \leftrightarrow G$	$A \leftrightarrow T$	$C \leftrightarrow G$	$C \leftrightarrow T$	$G \leftrightarrow T$
5	A	9280819	35372389	7900236	9032657	35268661	8513791
	C	6496778	27971742	4663713	6659875	27133993	5644473
	G	5692483	27247155	4672052	6672206	27893773	6460519
	T	8574866	35473468	7915494	9068442	35340532	9259786
100	A	180239045	738869870	153979210	180541857	725817733	174956607
	C	123529003	521168087	97349832	133531224	529311111	120091117
	G	120892679	531615538	97458204	133665361	519613929	122877822
	T	176237180	729638941	154241620	180924341	737993051	179644965

Tabela A.4: Contagens de nucleótidos ($k = 1$) na vizinhança de cada SNV, considerando $w = 5$ e $w = 100$.

		SNV					
w	$k = 2$	$A \leftrightarrow C$	$A \leftrightarrow G$	$A \leftrightarrow T$	$C \leftrightarrow G$	$C \leftrightarrow T$	$G \leftrightarrow T$
10	AA	5746309	20979294	5291280	5031292	19567736	5101868
	AC	2887509	11906895	2174806	2879791	11565981	2493221
	AG	3689599	16526338	2891562	4215773	15083787	3783472
	AT	4209549	16484857	3857655	3873820	16426607	4188078
	CA	4020234	16691346	3078050	4204654	16785538	3738955
	CC	3041114	12440381	2004547	3244348	13069244	2582730
	CG	491625	2372260	321338	617073	2359094	487989
	CT	3809839	15157028	2888983	4228374	16487660	3673648
	GA	3095749	14233615	2550565	3468218	13409277	3229687
	GC	2186231	10794894	1610290	2552381	10756521	2169884
	GG	2604168	13131844	2004927	3251537	12413515	3019156
	GT	2519479	11647167	2184241	2892187	11932211	2885048
	TA	3609566	14438958	3468466	3369957	14397012	3591640
	TC	3248416	13483452	2550164	3474282	14202393	3078676
	TG	3770995	16900803	3088507	4217474	16689832	4012838
	TT	5150514	19727413	5307296	5058563	21000113	5744535

Tabela A.5: Contagens dos dinucleótidos ($k = 2$) na vizinhança de cada SNV, considerando $w = 10$.

		SNV					
w	$k = 2$	$A \leftrightarrow C$	$A \leftrightarrow G$	$A \leftrightarrow T$	$C \leftrightarrow G$	$C \leftrightarrow T$	$G \leftrightarrow T$
200	AA	119688938	482635230	103239252	116623175	470991262	114329365
	AC	61348089	253286107	50184658	63171773	253294993	58673955
	AG	83743891	358778120	68233469	89666869	348570047	83098088
	AT	92977831	377679691	82912050	91015722	376583633	92487185
	CA	87884279	366177986	71511927	91815044	366642400	85090989
	CC	62362068	265816577	47558265	69353461	272043528	60573672
	CG	11297279	51460670	8136034	13687567	51183878	11222069
	CT	83645781	350157739	68276586	89743728	357866926	83355258
	GA	71133594	304031887	58601023	75411852	294211969	70679432
	GC	50338488	220904674	39162473	56466380	219993466	50045841
	GG	61016466	273244003	47602268	69451453	264889178	61998820
	GT	59169408	254882066	50344897	63374217	253221091	61180026
	TA	79001135	319929024	71394725	76708374	319084016	78593615
	TC	71134273	295537085	58620244	75454910	303242787	70781963
	TG	85757365	368751117	71696751	92044177	365725395	87591770
	TT	115284916	474091502	103550825	117046341	482790499	119461184

Tabela A.6: Contagens dos dinucleótidos ($k = 2$) na vizinhança de cada SNV, considerando $w = 200$.

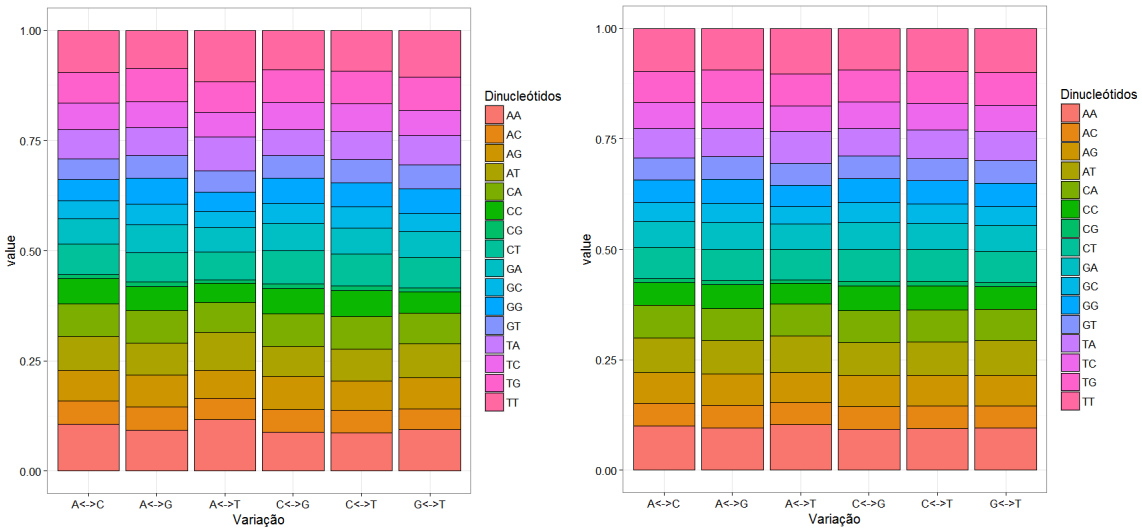


Figura A.2: Gráficos de barras das frequências relativas da contagem de dinucleótidos ($k = 2$) na vizinhança de cada SNV, considerando $w = 10$ e $w = 200$.

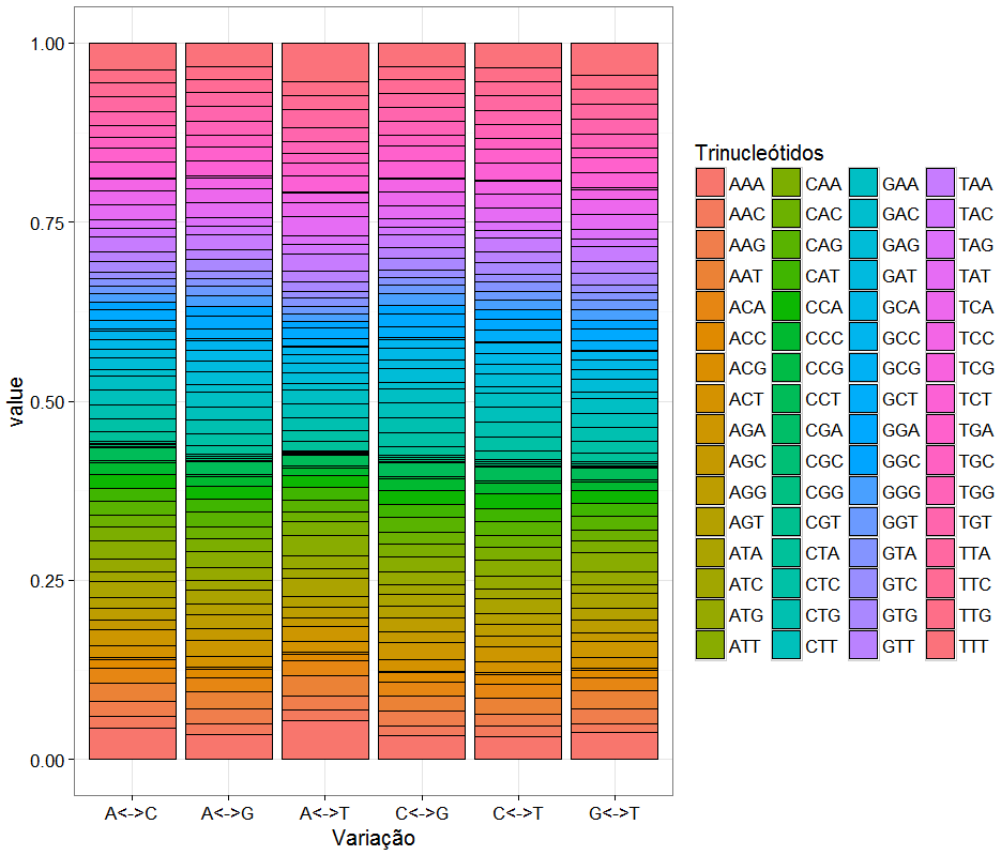


Figura A.3: Gráfico de barras das frequências relativas das contagens de trinucleótidos ($k = 3$) na vizinhança de cada SNV, considerando $w = 10$.

	SNV					
$k = 3$	$A \leftrightarrow C$	$A \leftrightarrow G$	$A \leftrightarrow T$	$C \leftrightarrow G$	$C \leftrightarrow T$	$G \leftrightarrow T$
AAA	0.0441	0.0350	0.0538	0.0329	0.0321	0.0368
AAC	0.0163	0.0149	0.0147	0.0143	0.0138	0.0134
AAG	0.0201	0.0207	0.0201	0.0201	0.0177	0.0203
AAT	0.0257	0.0230	0.0288	0.0213	0.0220	0.0250
ACA	0.0215	0.0199	0.0198	0.0198	0.0199	0.0186
ACC	0.0122	0.0123	0.0100	0.0124	0.0126	0.0103
ACG	0.0023	0.0026	0.0020	0.0026	0.0025	0.0023
ACT	0.0159	0.0153	0.0147	0.0164	0.0159	0.0151
AGA	0.0222	0.0230	0.0213	0.0236	0.0207	0.0225
AGC	0.0137	0.0150	0.0117	0.0149	0.0143	0.0126
AGG	0.0166	0.0194	0.0149	0.0192	0.0168	0.0182
AGT	0.0152	0.0159	0.0148	0.0164	0.0154	0.0159
ATA	0.0218	0.0189	0.0261	0.0171	0.0199	0.0206
ATC	0.0136	0.0135	0.0128	0.0129	0.0137	0.0125
ATG	0.0181	0.0179	0.0181	0.0173	0.0176	0.0181
ATT	0.0250	0.0221	0.0288	0.0213	0.0231	0.0258
CAA	0.0204	0.0180	0.0189	0.0182	0.0180	0.0178
CAC	0.0162	0.0167	0.0134	0.0160	0.0166	0.0137
CAG	0.0198	0.0209	0.0173	0.0216	0.0198	0.0202
CAT	0.0180	0.0176	0.0181	0.0173	0.0179	0.0182
CCA	0.0192	0.0185	0.0160	0.0198	0.0201	0.0174
CCC	0.0160	0.0140	0.0111	0.0158	0.0148	0.0121
CCG	0.0027	0.0029	0.0018	0.0031	0.0031	0.0025
CCT	0.0182	0.0169	0.0150	0.0193	0.0194	0.0166
CGA	0.0021	0.0023	0.0017	0.0024	0.0022	0.0020
CGC	0.0023	0.0027	0.0015	0.0028	0.0027	0.0020
CGG	0.0025	0.0031	0.0018	0.0031	0.0029	0.0027
CGT	0.0023	0.0026	0.0020	0.0027	0.0026	0.0023
CTA	0.0134	0.0120	0.0133	0.0125	0.0131	0.0125
CTC	0.0179	0.0169	0.0148	0.0185	0.0188	0.0156
CTG	0.0202	0.0198	0.0173	0.0217	0.0209	0.0198
CTT	0.0204	0.0178	0.0201	0.0202	0.0208	0.0202
GAA	0.0196	0.0205	0.0198	0.0194	0.0190	0.0199
GAC	0.0093	0.0104	0.0087	0.0101	0.0098	0.0091
GAG	0.0156	0.0188	0.0148	0.0185	0.0169	0.0178
GAT	0.0126	0.0137	0.0128	0.0129	0.0135	0.0136
GCA	0.0137	0.0154	0.0128	0.0153	0.0157	0.0142
GCC	0.0115	0.0135	0.0092	0.0130	0.0142	0.0108
GCG	0.0020	0.0027	0.0015	0.0028	0.0027	0.0023
GCT	0.0127	0.0143	0.0117	0.0150	0.0151	0.0138
GGA	0.0144	0.0169	0.0137	0.0168	0.0154	0.0160
GGC	0.0109	0.0143	0.0092	0.0130	0.0135	0.0115

Tabela A.7: Frequências relativas das contagens dos trinucleótidos ($k = 3$) na vizinhança de cada SNV, considerando $w = 10$.

	SNV					
$k = 3$	$A \leftrightarrow C$	$A \leftrightarrow G$	$A \leftrightarrow T$	$C \leftrightarrow G$	$C \leftrightarrow T$	$G \leftrightarrow T$
GGG	0.0121	0.0148	0.0110	0.0158	0.0140	0.0158
GGT	0.0103	0.0127	0.0100	0.0124	0.0124	0.0122
GTA	0.0111	0.0112	0.0120	0.0109	0.0121	0.0118
GTC	0.0091	0.0099	0.0087	0.0102	0.0104	0.0094
GTG	0.0138	0.0167	0.0135	0.0160	0.0168	0.0163
GTT	0.0135	0.0139	0.0148	0.0144	0.0150	0.0164
TAA	0.0215	0.0201	0.0250	0.0183	0.0192	0.0209
TAC	0.0118	0.0121	0.0120	0.0109	0.0112	0.0110
TAG	0.0125	0.0131	0.0134	0.0125	0.0120	0.0134
TAT	0.0206	0.0199	0.0261	0.0172	0.0190	0.0218
TCA	0.0195	0.0198	0.0194	0.0201	0.0202	0.0195
TCC	0.0160	0.0154	0.0137	0.0168	0.0169	0.0144
TCG	0.0020	0.0022	0.0017	0.0025	0.0023	0.0021
TCT	0.0225	0.0208	0.0213	0.0237	0.0231	0.0223
TGA	0.0195	0.0203	0.0194	0.0201	0.0198	0.0195
TGC	0.0142	0.0157	0.0128	0.0153	0.0154	0.0137
TGG	0.0175	0.0202	0.0161	0.0198	0.0186	0.0193
TGT	0.0187	0.0200	0.0200	0.0199	0.0201	0.0216
TTA	0.0209	0.0193	0.0251	0.0183	0.0201	0.0215
TTC	0.0200	0.0190	0.0199	0.0194	0.0205	0.0197
TTG	0.0179	0.0181	0.0190	0.0183	0.0182	0.0205
TTT	0.0370	0.0323	0.0540	0.0331	0.0352	0.0444

Tabela A.8: Frequências relativas das contagens dos trinucleótidos ($k = 3$) na vizinhança de cada SNV, considerando $w = 10$ (Continuação).

		SNV					
w	$k = 1$	$A \leftrightarrow C$	$A \leftrightarrow G$	$A \leftrightarrow T$	$C \leftrightarrow G$	$C \leftrightarrow T$	$G \leftrightarrow T$
5	A	287.69	-170.79	321.21	15.52	-166.39	-15.37
	C	39.79	287.21	-354.59	-21.62	87.17	-340.61
	G	-336.39	86.73	-352.60	-18.63	284.73	36.97
	T	-18.72	-168.07	319.70	20.97	-170.69	290.38
100	A	301.60	-36.77	413.05	-70.60	-308.86	-11.54
	C	28.05	104.48	-457.54	75.79	279.51	-349.85
	G	-346.62	280.27	-456.76	76.95	103.27	24.38
	T	-14.94	-309.15	409.14	-66.76	-35.43	304.04

Tabela A.9: Resíduos ajustados do teste de independência entre os nucleótidos ($k = 1$) e as SNVs com $w = 5$ e $w = 100$.

		SNV					
w	$k = 2$	$A \leftrightarrow C$	$A \leftrightarrow G$	$A \leftrightarrow T$	$C \leftrightarrow G$	$C \leftrightarrow T$	$G \leftrightarrow T$
10	AA	346.82	-42.58	569.85	-113.56	-420.35	45.89
	AC	77.73	110.60	-98.80	-9.38	-1.43	-166.76
	AG	-44.26	228.90	-159.40	148.82	-217.79	19.73
	AT	112.74	-95.46	298.71	-166.04	-95.94	113.39
	CA	33.36	25.10	-139.59	33.47	72.48	-108.20
	CC	45.11	-5.72	-324.91	84.51	235.67	-230.79
	CG	-72.54	78.61	-205.28	68.95	74.20	-73.62
	CT	20.23	-217.26	-163.14	153.05	227.75	-44.09
	GA	-99.58	186.74	-116.95	31.89	-80.77	-9.01
	GC	-182.33	196.55	-328.29	-9.78	195.36	-184.67
	GG	-229.14	237.39	-326.45	86.62	-5.43	39.59
	GT	-166.93	-5.24	-99.26	-9.73	115.21	78.09
	TA	64.05	-80.02	337.86	-164.02	-77.46	65.05
	TC	-11.03	-77.05	-119.36	32.90	186.12	-101.61
	TG	-109.44	73.42	-139.72	32.90	24.99	34.16
	TT	45.04	-420.50	567.42	-111.35	-42.48	349.05
200	AA	495.69	43.95	788.58	-354.38	-547.82	11.96
	AC	133.60	-4.41	-53.85	0.45	64.71	-189.34
	AG	-55.65	368.71	-289.99	180.23	-244.44	-77.02
	AT	239.89	-236.85	845.36	-452.30	-224.03	241.92
	CA	75.05	9.38	-191.02	62.15	123.73	-193.51
	CC	-147.63	-34.86	-812.87	390.69	519.33	-341.94
	CG	-210.92	198.51	-625.77	338.09	180.14	-214.13
	CT	-76.38	-239.97	-292.70	179.61	366.27	-55.75
	GA	-23.97	350.99	-152.71	99.21	-296.45	-31.60
	GC	-241.49	242.83	-697.52	303.74	228.40	-242.83
	GG	-334.10	528.64	-811.14	397.92	-49.83	-154.98
	GT	-191.26	62.18	-52.41	2.28	-1.23	131.92
	TA	235.35	-249.08	915.22	-474.20	-231.36	238.32
	TC	-31.16	-291.01	-156.61	97.07	350.20	-25.77
	TG	-194.68	125.49	-188.57	64.42	6.42	74.25
	TT	5.24	-555.44	786.22	-351.81	54.61	498.79

Tabela A.10: Resíduos ajustados do teste de independência entre os dinucleótidos ($k = 2$) e as SNVs com $w = 10$ e $w = 200$.

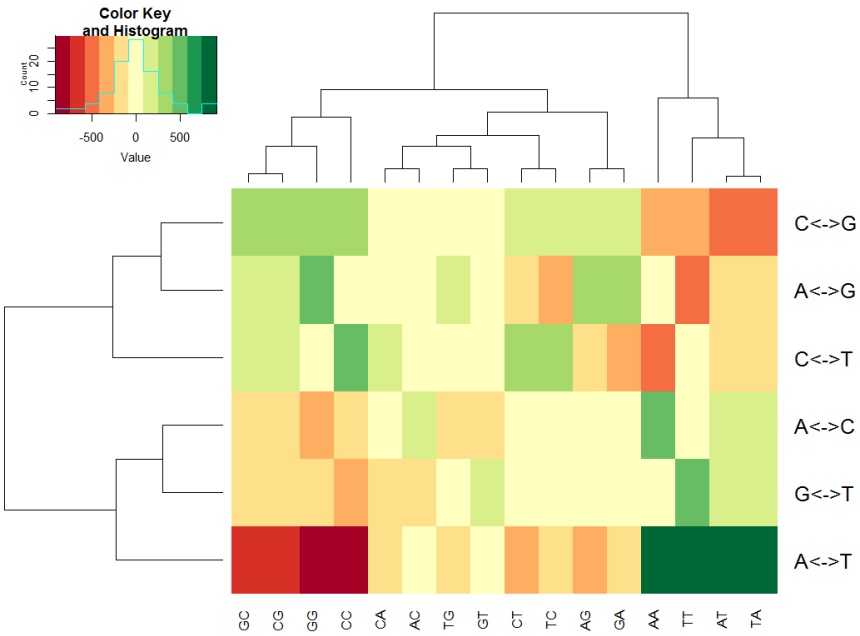


Figura A.4: *Heatmap* dos resíduos ajustados do teste de independência entre os dinucleótidos ($k = 2$) e as SNVs com $w = 200$.

	SNV					
$k = 3$	$A \leftrightarrow C$	$A \leftrightarrow G$	$A \leftrightarrow T$	$C \leftrightarrow G$	$C \leftrightarrow T$	$G \leftrightarrow T$
AAA	312.29	-94.82	625.42	-124.94	-363.00	31.99
AAC	111.44	64.72	13.11	-10.50	-93.33	-63.51
AAG	31.88	152.45	26.26	33.18	-227.46	39.86
AAT	116.93	-31.10	238.63	-100.27	-149.58	78.68
ACA	84.73	-0.56	-2.51	-5.21	-4.67	-68.21
ACC	4.37	38.34	-125.68	18.63	84.59	-120.84
ACG	-24.02	34.02	-61.70	20.99	16.50	-28.16
ACT	16.87	-36.12	-46.24	49.53	37.47	-27.18
AGA	6.63	113.55	-31.95	76.84	-157.81	20.94
AGC	-30.39	114.96	-139.33	42.82	9.17	-100.32
AGG	-68.21	199.45	-146.17	75.12	-136.30	16.63
AGT	-28.53	37.78	-44.36	46.50	-35.25	17.56
ATA	93.17	-127.35	289.75	-151.24	-9.43	31.35
ATC	12.74	16.81	-32.06	-35.39	45.85	-55.76
ATG	18.80	16.86	13.96	-30.46	-25.31	14.08
ATT	78.93	-149.45	238.34	-98.64	-33.31	119.16
CAA	114.95	-33.93	31.61	-2.37	-33.16	-25.49
CAC	8.87	84.10	-140.73	-6.47	69.02	-138.25
CAG	-19.83	89.92	-137.71	74.88	-48.64	-1.33
CAT	12.83	-25.49	14.90	-27.99	15.23	19.88
CCA	13.97	-54.12	-143.10	42.56	143.66	-80.75

Tabela A.11: Resíduos ajustados do teste de independência entre os trinucleótidos ($k = 3$) e as SNVs com $w = 10$.

	SNV					
$k = 3$	$A \leftrightarrow C$	$A \leftrightarrow G$	$A \leftrightarrow T$	$C \leftrightarrow G$	$C \leftrightarrow T$	$G \leftrightarrow T$
CCC	105.20	-37.61	-176.34	95.77	89.96	-131.40
CCG	-24.09	10.30	-128.93	41.01	75.07	-46.94
CCT	17.04	-137.45	-144.52	76.56	199.97	-70.26
CGA	-14.81	41.43	-62.82	38.17	-3.62	-31.88
CGC	-37.35	59.74	-134.26	35.64	56.37	-76.68
CGG	-45.83	78.57	-129.54	41.37	6.94	-25.29
CGT	-27.68	16.96	-62.40	21.55	33.45	-24.20
CTA	48.01	-108.00	38.15	-7.97	69.92	-8.95
CTC	18.34	-84.52	-136.71	53.92	176.54	-106.93
CTG	-1.55	-48.92	-137.42	79.72	87.43	-19.87
CTT	41.22	-226.71	24.91	33.51	151.91	31.57
GAA	-3.59	96.90	4.64	-19.21	-92.33	10.90
GAC	-39.41	93.36	-74.85	18.68	-7.40	-59.85
GAG	-105.86	178.16	-133.87	53.85	-87.80	17.55
GAT	-53.52	46.05	-33.03	-35.40	14.92	14.48
GCA	-83.86	44.29	-120.97	11.65	92.27	-53.26
GCC	-96.07	74.18	-220.89	-2.09	179.32	-137.73
GCG	-76.41	58.10	-132.75	35.22	57.35	-37.34
GCT	-98.51	8.07	-141.85	45.05	115.78	-31.34
GGA	-84.39	141.40	-113.75	56.40	-69.79	7.43
GGC	-138.40	183.07	-222.75	-2.19	73.13	-98.32
GGG	-130.55	91.25	-175.82	98.77	-36.71	97.41
GGT	-122.83	83.98	-126.19	17.20	40.29	6.23
GTA	-33.61	-67.13	29.52	-43.69	85.95	18.54
GTC	-59.89	-6.89	-77.70	19.04	94.45	-39.68
GTG	-137.64	69.79	-138.65	-8.61	84.34	7.17
GTT	-64.18	-96.83	11.89	-10.95	70.64	110.18
TAA	68.73	-10.78	228.76	-99.62	-113.29	37.15
TAC	16.83	83.96	29.54	-43.86	-64.02	-34.02
TAG	-11.07	69.05	40.54	-9.36	-106.20	47.63
TAT	32.27	-11.12	287.34	-149.19	-125.56	92.19
TCA	-21.13	-14.95	-24.07	10.57	45.79	-20.92
TCC	8.62	-69.87	-113.78	55.27	143.52	-87.87
TCG	-31.74	-0.99	-65.18	40.02	39.11	-15.10
TCT	17.87	-154.68	-35.47	78.88	112.29	7.95
TGA	-20.42	47.07	-23.49	8.94	-17.37	-18.61
TGC	-52.51	92.88	-122.28	11.55	43.93	-83.90
TGG	-83.37	146.90	-144.56	42.51	-54.96	13.78
TGT	-69.11	-7.92	-2.05	-5.74	4.68	82.69
TTA	37.34	-112.11	228.29	-99.85	-11.53	68.69
TTC	11.38	-91.18	4.58	-20.38	96.43	-3.77
TTG	-26.35	-36.16	30.03	-2.75	-29.65	115.81
TTT	31.19	-362.78	624.26	-121.98	-96.42	314.64

Tabela A.12: Resíduos ajustados do teste de independência entre os trinucleótidos ($k = 3$) e as SNVs com $w = 10$ (continuação).

Posição	Nucleótidos			
d	A	C	G	T
-20	0.297	0.204	0.207	0.292
-19	0.294	0.204	0.209	0.293
-18	0.293	0.206	0.210	0.290
-17	0.290	0.208	0.211	0.291
-16	0.290	0.207	0.213	0.290
-15	0.291	0.207	0.214	0.288
-14	0.290	0.210	0.211	0.290
-13	0.295	0.205	0.209	0.291
-12	0.292	0.206	0.212	0.291
-11	0.292	0.207	0.210	0.291
-10	0.287	0.209	0.212	0.292
-9	0.288	0.209	0.212	0.291
-8	0.294	0.206	0.213	0.287
-7	0.295	0.202	0.214	0.288
-6	0.291	0.207	0.217	0.285
-5	0.294	0.210	0.212	0.284
-4	0.294	0.213	0.209	0.285
-3	0.286	0.218	0.205	0.291
-2	0.276	0.246	0.208	0.270
-1	0.301	0.290	0.180	0.229
0				
1	0.228	0.180	0.290	0.302
2	0.270	0.208	0.246	0.277
3	0.291	0.204	0.219	0.286
4	0.284	0.209	0.213	0.295
5	0.283	0.212	0.210	0.295
6	0.284	0.217	0.207	0.292
7	0.287	0.214	0.203	0.296
8	0.286	0.213	0.207	0.294
9	0.290	0.212	0.210	0.289
10	0.291	0.212	0.209	0.287
11	0.290	0.210	0.207	0.292
12	0.290	0.212	0.207	0.292
13	0.290	0.209	0.206	0.295
14	0.289	0.211	0.210	0.290
15	0.287	0.214	0.208	0.291
16	0.289	0.213	0.208	0.290
17	0.290	0.211	0.209	0.290
18	0.289	0.211	0.207	0.293
19	0.291	0.210	0.205	0.294
20	0.291	0.207	0.205	0.297

Posição	Nucleótidos			
d	A	C	G	T
-20	0.298	0.202	0.203	0.297
-19	0.297	0.202	0.204	0.296
-18	0.299	0.203	0.204	0.295
-17	0.300	0.202	0.203	0.296
-16	0.297	0.202	0.206	0.295
-15	0.297	0.203	0.202	0.297
-14	0.297	0.204	0.201	0.299
-13	0.299	0.202	0.201	0.298
-12	0.297	0.205	0.202	0.297
-11	0.300	0.201	0.201	0.298
-10	0.298	0.202	0.204	0.296
-9	0.299	0.202	0.202	0.297
-8	0.302	0.200	0.203	0.294
-7	0.301	0.200	0.204	0.296
-6	0.298	0.203	0.204	0.295
-5	0.301	0.207	0.199	0.293
-4	0.309	0.208	0.200	0.283
-3	0.319	0.200	0.198	0.283
-2	0.317	0.201	0.192	0.290
-1	0.309	0.192	0.216	0.283
0				
1	0.282	0.216	0.193	0.309
2	0.289	0.192	0.201	0.318
3	0.282	0.198	0.201	0.320
4	0.282	0.200	0.209	0.309
5	0.292	0.199	0.208	0.301
6	0.295	0.204	0.203	0.298
7	0.295	0.204	0.200	0.301
8	0.293	0.204	0.201	0.302
9	0.296	0.202	0.202	0.300
10	0.295	0.204	0.202	0.299
11	0.297	0.201	0.202	0.300
12	0.295	0.202	0.206	0.297
13	0.297	0.202	0.203	0.299
14	0.297	0.202	0.204	0.297
15	0.296	0.203	0.204	0.298
16	0.294	0.206	0.203	0.297
17	0.295	0.203	0.203	0.299
18	0.294	0.204	0.203	0.299
19	0.295	0.204	0.203	0.297
20	0.296	0.203	0.202	0.298

Tabela A.13: Frequências relativas de nucleótidos nas posições $d = \pm 1, \pm 2, \ldots, \pm 20$, para as transições (à esquerda) e transversões (à direita), em que o local de variação corresponde à posição 0.

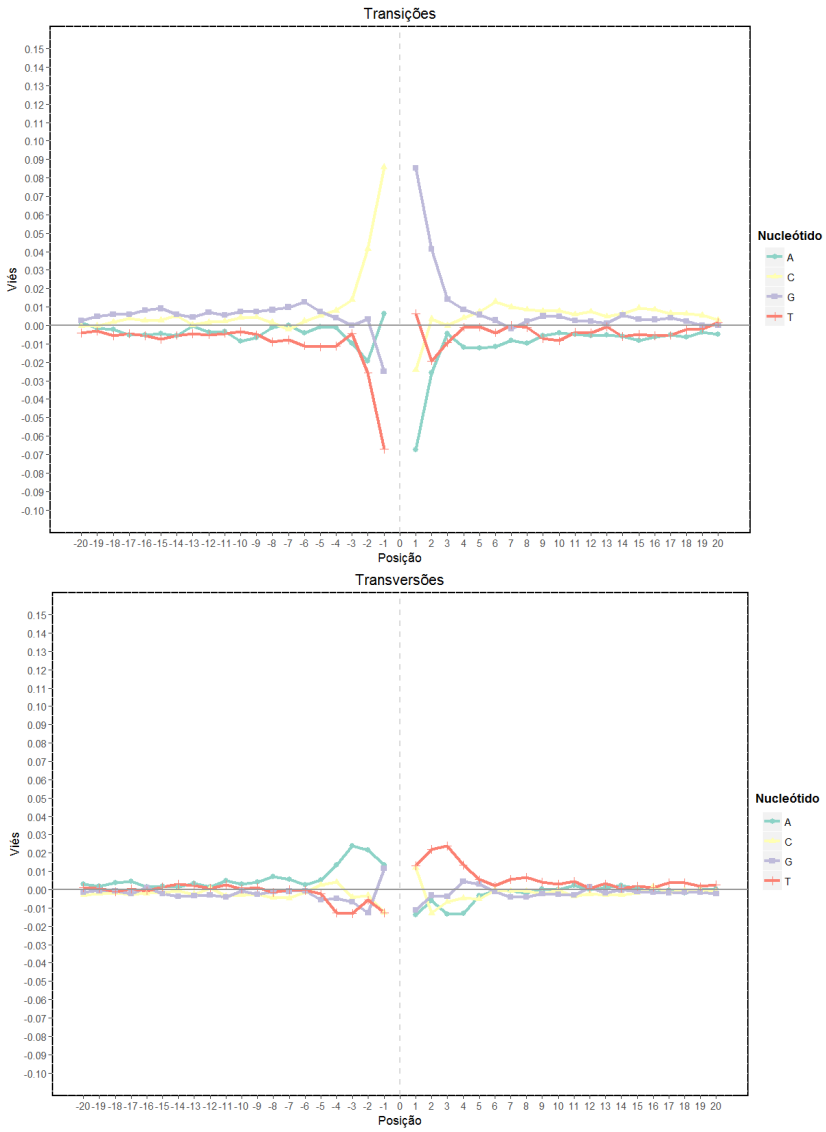


Figura A.5: Padrões de frequência globais dos nucleótidos ($k = 1$) para as transições e transversões, numa vizinhança com $w = 20$.

Posição d	Nucleótidos complementares A	T
±1	5735054	5755825
±2	6787022	6799729
±3	7315770	7334708
±4	7136163	7164768
±5	7121645	7148337
±6	7138545	7166181
±7	7221819	7246245
±8	7192558	7221575
±9	7289981	7317344
±10	7326963	7361546
±11	7306170	7334647
±12	7290804	7316590
±13	7304046	7332793
±14	7281218	7306521
±15	7230404	7257415
±16	7276712	7304167
±17	7299964	7328179
±18	7272774	7300116
±19	7335899	7363862
±20	7314901	7342700

Posição d	Nucleótidos complementares C	G
±1	4535376	4521833
±2	5232930	5232798
±3	5142073	5148708
±4	5249244	5249337
±5	5333027	5336333
±6	5468225	5467932
±7	5396478	5397326
±8	5362118	5360518
±9	5340462	5335353
±10	5340830	5337238
±11	5295267	5287421
±12	5332474	5325652
±13	5265895	5264791
±14	5305313	5302667
±15	5385011	5379727
±16	5360437	5355511
±17	5307409	5298529
±18	5305018	5298077
±19	5284125	5270642
±20	5221618	5215013

Tabela A.14: Tabela de contingência das contagens dos nucleótidos complementares A/T e C/G nas posições $d = \pm1, \pm2, \ldots, \pm20$, para as transições.

Posição d	Nucleótidos complementares A	T
±1	3280142	3298677
±2	3369241	3378607
±3	3283930	3292340
±4	3286459	3298022
±5	3400572	3418026
±6	3431907	3442361
±7	3433012	3445254
±8	3417393	3428467
±9	3446127	3460359
±10	3439474	3450250
±11	3464146	3477336
±12	3442428	3456117
±13	3455142	3470910
±14	3465774	3478825
±15	3450598	3464158
±16	3424167	3438583
±17	3435015	3449834
±18	3420251	3433723
±19	3439420	3453051
±20	3446247	3459827

Posição d	Nucleótidos complementares C	G
±1	2520289	2518439
±2	2233432	2235025
±3	2303235	2304980
±4	2328826	2328624
±5	2323054	2318197
±6	2378896	2373724
±7	2372963	2373220
±8	2373169	2369653
±9	2354355	2352421
±10	2377854	2375759
±11	2339986	2336377
±12	2354189	2347760
±13	2349603	2345461
±14	2348269	2341367
±15	2361248	2358666
±16	2400527	2396543
±17	2364507	2359601
±18	2377259	2373743
±19	2379918	2377697
±20	2370129	2367815

Tabela A.15: Tabela de contingência das contagens dos nucleótidos complementares A/T e C/G nas posições $d = \pm1, \pm2, \ldots, \pm20$, para as transversões.

SNV	Posição	Dinucleótidos															
	d	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
$A \leftrightarrow G$	-1	0.070	0.092	0.064	0.072	0.078	0.098	0.006	0.071	0.054	0.078	0.044	0.049	0.039	0.081	0.050	0.054
	-2	0.090	0.060	0.070	0.060	0.078	0.069	0.011	0.055	0.063	0.054	0.067	0.040	0.067	0.070	0.077	0.069
	-3	0.092	0.050	0.079	0.075	0.072	0.057	0.010	0.069	0.057	0.046	0.061	0.050	0.059	0.060	0.074	0.088
	-4	0.095	0.054	0.073	0.073	0.075	0.053	0.012	0.068	0.062	0.045	0.055	0.051	0.064	0.056	0.074	0.089
$C \leftrightarrow T$	-1	0.070	0.092	0.064	0.072	0.078	0.098	0.006	0.071	0.054	0.078	0.044	0.049	0.039	0.081	0.050	0.054
	-2	0.090	0.060	0.070	0.060	0.078	0.069	0.011	0.055	0.063	0.054	0.067	0.040	0.067	0.070	0.077	0.069
	-3	0.092	0.050	0.079	0.075	0.072	0.057	0.010	0.069	0.057	0.046	0.061	0.050	0.059	0.060	0.074	0.088
	-4	0.095	0.054	0.073	0.073	0.075	0.053	0.012	0.068	0.062	0.045	0.055	0.051	0.064	0.056	0.074	0.089
$A \leftrightarrow C$	-1	0.109	0.079	0.076	0.068	0.073	0.059	0.007	0.050	0.059	0.049	0.057	0.032	0.071	0.060	0.067	0.084
	-2	0.118	0.050	0.076	0.086	0.079	0.053	0.008	0.067	0.065	0.034	0.047	0.041	0.071	0.052	0.065	0.089
	-3	0.118	0.054	0.068	0.078	0.082	0.063	0.009	0.071	0.064	0.037	0.046	0.046	0.067	0.052	0.064	0.081
	-4	0.109	0.057	0.068	0.070	0.078	0.065	0.009	0.067	0.062	0.041	0.048	0.044	0.070	0.062	0.069	0.082
$A \leftrightarrow T$	-1	0.119	0.071	0.079	0.071	0.073	0.051	0.007	0.035	0.057	0.040	0.050	0.034	0.076	0.059	0.067	0.112
	-2	0.128	0.046	0.066	0.089	0.068	0.040	0.007	0.057	0.067	0.033	0.046	0.047	0.076	0.047	0.061	0.121
	-3	0.127	0.047	0.070	0.084	0.070	0.040	0.008	0.061	0.056	0.032	0.045	0.043	0.076	0.051	0.070	0.118
	-4	0.122	0.047	0.063	0.083	0.073	0.045	0.007	0.067	0.055	0.033	0.041	0.045	0.079	0.053	0.066	0.121
$C \leftrightarrow G$	-1	0.093	0.044	0.081	0.064	0.094	0.049	0.008	0.106	0.069	0.028	0.048	0.045	0.073	0.035	0.069	0.092
	-2	0.096	0.064	0.070	0.070	0.075	0.071	0.011	0.070	0.056	0.050	0.044	0.045	0.056	0.072	0.066	0.084
	-3	0.096	0.053	0.067	0.068	0.079	0.068	0.010	0.075	0.068	0.047	0.052	0.051	0.058	0.060	0.066	0.082
	-4	0.088	0.058	0.075	0.066	0.073	0.062	0.012	0.069	0.063	0.049	0.058	0.048	0.058	0.064	0.073	0.083
$G \leftrightarrow T$	-1	0.090	0.045	0.091	0.091	0.065	0.038	0.009	0.072	0.054	0.028	0.066	0.050	0.061	0.040	0.080	0.119
	-2	0.120	0.048	0.069	0.083	0.072	0.045	0.010	0.063	0.066	0.039	0.056	0.054	0.060	0.052	0.064	0.101
	-3	0.113	0.048	0.075	0.073	0.074	0.047	0.009	0.061	0.064	0.039	0.057	0.047	0.068	0.056	0.075	0.095
	-4	0.102	0.049	0.071	0.076	0.075	0.047	0.010	0.067	0.063	0.041	0.052	0.049	0.069	0.054	0.074	0.101

Tabela A.16: Frequências relativas dos dinucleótidos, nas posições $d = \pm 1, \pm 2, \ldots, \pm 4$ numa vizinhança em torno do local de variação, para cada uma das SNVs.

SNV	Posição	Dinucleótidos															
	d	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
$A \leftrightarrow G$	1	0.074	0.029	0.060	0.047	0.073	0.053	0.011	0.058	**0.068**	0.046	0.062	0.054	0.101	0.062	0.106	0.095
	2	0.104	0.047	0.082	0.083	0.067	0.044	0.009	0.071	0.072	0.046	0.064	0.057	0.058	0.048	0.068	0.080
	3	0.092	0.053	0.077	0.078	0.067	0.045	0.011	0.063	0.065	0.043	0.056	0.060	0.063	0.063	0.074	0.092
	4	0.090	0.050	0.071	0.075	0.072	0.051	0.010	0.070	0.061	0.045	0.056	0.056	0.061	0.065	0.075	0.092
$C \leftrightarrow T$	1	0.074	0.029	0.060	0.047	0.073	0.053	0.011	0.058	0.068	0.046	0.062	0.054	0.101	0.062	0.106	0.095
	2	0.104	0.047	0.082	0.083	0.067	0.044	0.009	0.071	0.072	0.046	0.064	0.057	0.058	0.048	0.068	0.080
	3	0.092	0.053	0.077	0.078	0.067	0.045	0.011	0.063	0.065	0.043	0.056	0.060	0.063	0.063	0.074	0.092
	4	0.090	0.050	0.071	0.075	0.072	0.051	0.010	0.070	0.061	0.045	0.056	0.056	0.061	0.065	0.075	0.092
$A \leftrightarrow C$	1	0.118	0.050	0.072	0.091	0.080	0.067	0.009	0.092	0.040	0.028	0.039	0.046	0.061	0.053	0.065	0.090
	2	0.099	0.054	0.062	0.082	0.063	0.057	0.010	0.069	0.052	0.039	0.045	0.048	0.060	0.066	0.073	0.120
	3	0.094	0.046	0.062	0.073	0.075	0.057	0.009	0.075	0.056	0.039	0.047	0.048	0.068	0.064	0.074	0.113
	4	0.100	0.049	0.067	0.076	0.074	0.052	0.010	0.071	0.055	0.042	0.047	0.049	0.069	0.063	0.075	0.102
$A \leftrightarrow T$	1	0.111	0.034	0.035	0.071	0.067	0.050	0.007	0.079	0.059	0.040	0.051	0.071	0.076	0.057	0.073	0.119
	2	0.121	0.046	0.056	0.089	0.061	0.046	0.007	0.066	0.047	0.033	0.040	0.046	0.076	0.067	0.068	0.129
	3	0.118	0.043	0.061	0.084	0.070	0.045	0.008	0.070	0.051	0.032	0.040	0.048	0.076	0.057	0.071	0.127
	4	0.120	0.045	0.067	0.082	0.066	0.041	0.007	0.063	0.053	0.034	0.045	0.047	0.079	0.055	0.073	0.123
$C \leftrightarrow G$	1	0.091	0.045	0.105	0.063	0.068	0.048	0.008	0.081	0.035	0.028	0.050	0.045	0.073	0.070	0.094	0.094
	2	0.083	0.045	0.070	0.070	0.065	0.044	0.011	0.070	0.072	0.050	0.072	0.064	0.056	0.056	0.076	0.097
	3	0.082	0.051	0.075	0.068	0.066	0.052	0.010	0.067	0.060	0.047	0.069	0.053	0.058	0.068	0.078	0.096
	4	0.083	0.048	0.069	0.066	0.072	0.058	0.012	0.076	0.064	0.049	0.062	0.058	0.058	0.064	0.074	0.089
$G \leftrightarrow T$	1	0.084	0.032	0.050	0.068	0.067	0.057	0.007	0.076	0.060	0.048	0.059	0.080	0.071	0.059	0.073	0.109
	2	0.089	0.041	0.067	0.086	0.065	0.046	0.008	0.076	0.052	0.034	0.053	0.051	0.070	0.065	0.079	0.119
	3	0.081	0.046	0.072	0.078	0.064	0.045	0.009	0.068	0.052	0.037	0.063	0.055	0.066	0.064	0.082	0.119
	4	0.081	0.044	0.067	0.070	0.068	0.048	0.009	0.068	0.063	0.041	0.066	0.057	0.069	0.062	0.079	0.109

Tabela A.17: Frequências relativas dos dinucleótidos, nas posições $d = \pm 1, \pm 2, \ldots, \pm 4$ numa vizinhança em torno do local de variação, para cada uma das SNVs.

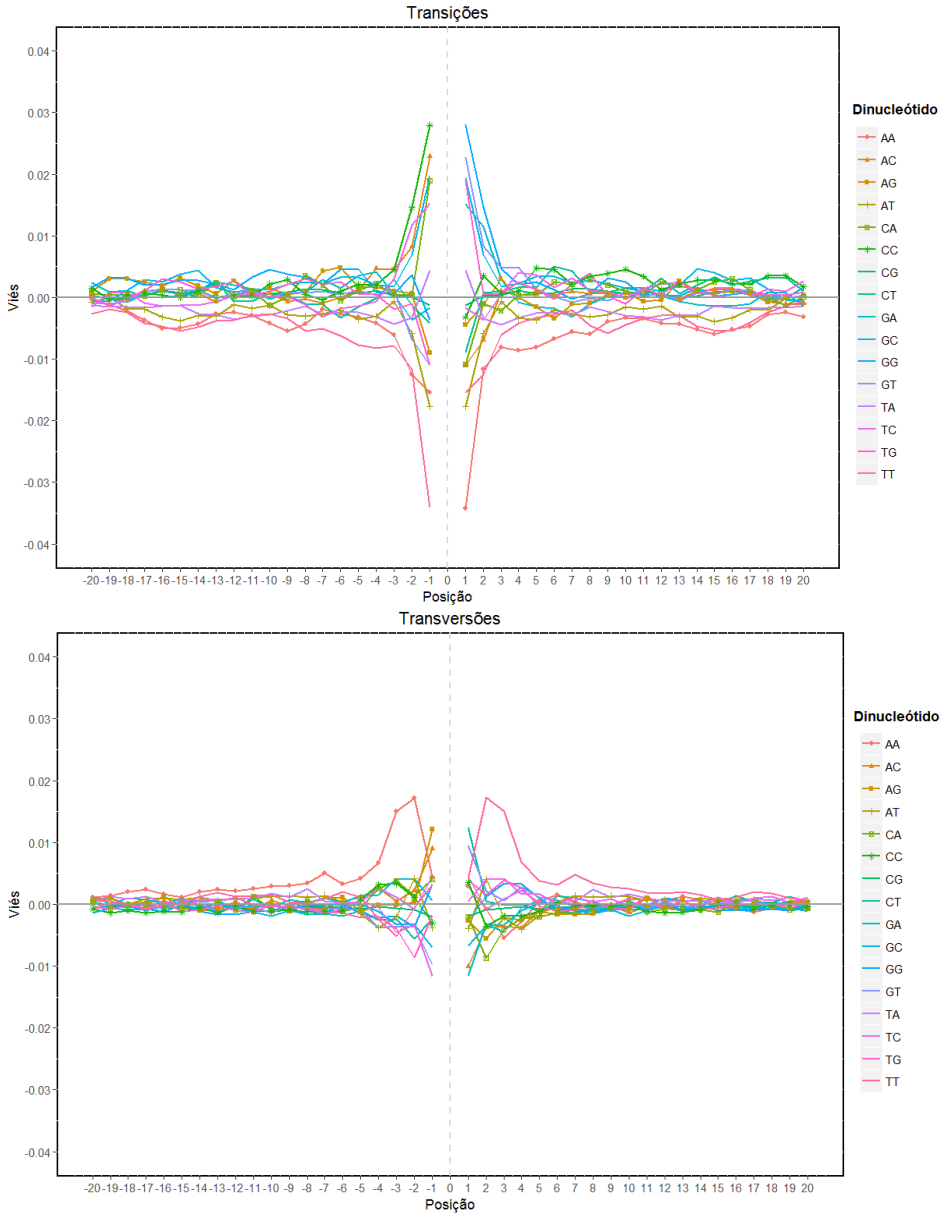


Figura A.6: Padrões de frequência globais dos dinucleótidos ($k = 2$) para as transições e transversões, numa vizinhança com $w = 20$.

	SNV					
Dinucleótidos	$A \leftrightarrow C$	$A \leftrightarrow G$	$A \leftrightarrow T$	$C \leftrightarrow G$	$C \leftrightarrow T$	$G \leftrightarrow T$
AA	26890	189539	36877	12342	318619	57205
AC	51071	577960	42156	23876	85360	22597
AG	11796	96051	50965	53551	93677	41198
AT	19089	203505	17535	7588	204536	19560
CA	14164	53940	7334	28321	316391	7751
CC	32928	602818	10253	41775	128070	18863
CG	2773	62367	2661	5396	61770	3073
CT	42687	95598	50606	56881	94580	12408
GA	28612	84971	13220	50967	204494	16303
GC	24852	351066	8210	49950	347922	25971
GG	17424	127829	10824	41046	606970	31762
GT	22896	86028	41646	22361	578799	50160
TA	7075	430935	3494	24379	429797	7005
TC	17072	206946	13097	53377	86514	29849
TG	8048	320025	7587	26859	54508	14208
TT	55522	319658	36004	12423	183721	26159

Tabela A.18: Valores da estatística do χ^2 resultantes da aplicação do teste de ajustamento do χ^2 a cada dinucleótido em cada SNV, considerando as 40 posições da vizinhança $d = \pm 1 \ldots \pm 20$.

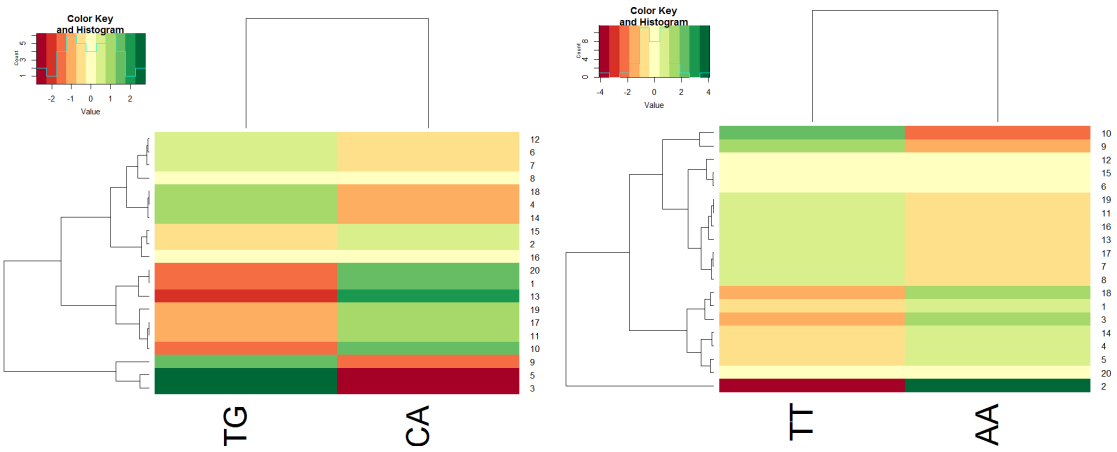


Figura A.7: *Heatmaps* dos resíduos ajustados do teste de homogeneidade para as contagens dos dinucleótidos complementos-invertidos (AA/TT e CA/TG), nas transições.

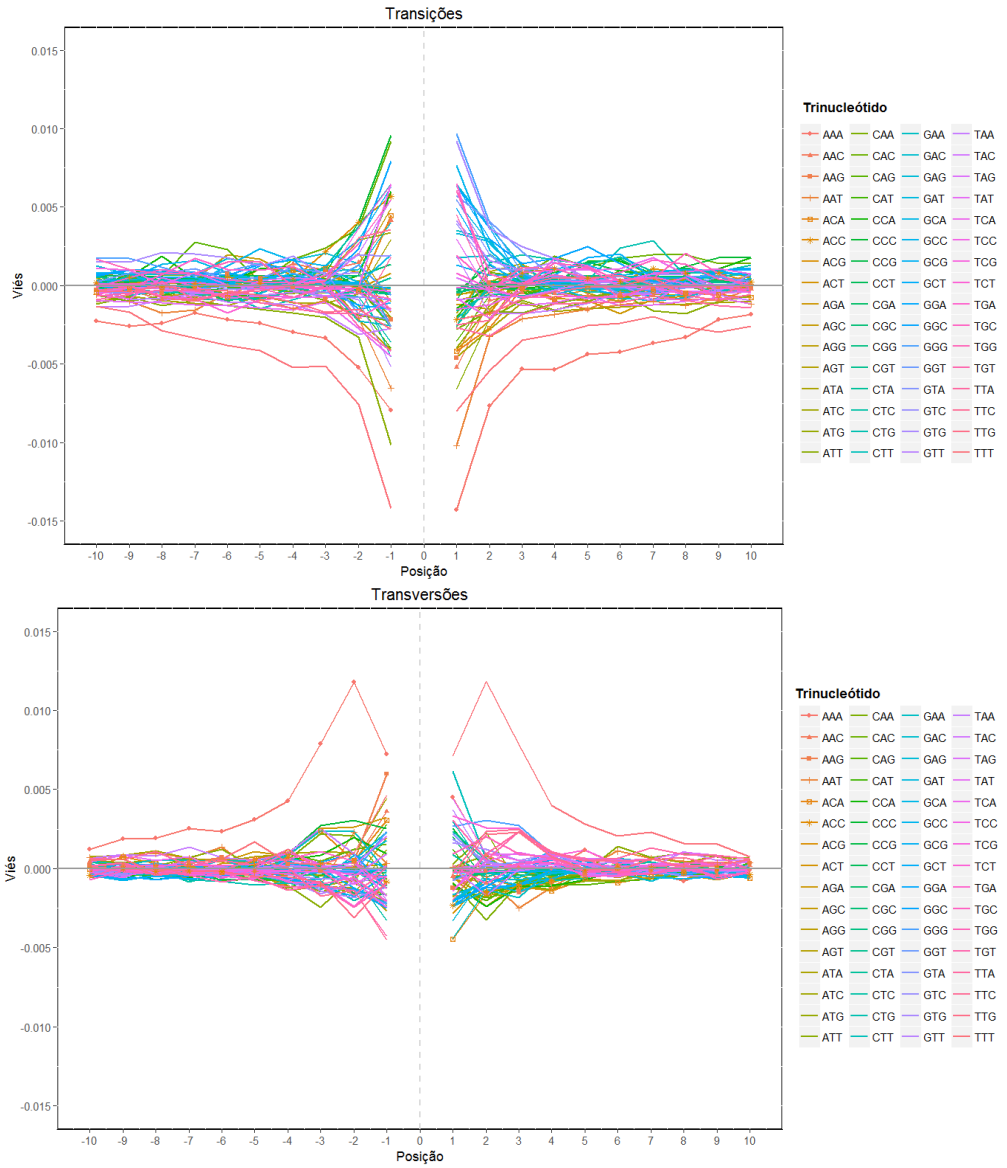


Figura A.8: Padrões de frequência globais dos trinucleótidos ($k = 3$) para as transições e transversões, numa vizinhança com $w = 10$.

	SNV					
Trinucleótidos	$A \leftrightarrow C$	$A \leftrightarrow G$	$A \leftrightarrow T$	$C \leftrightarrow G$	$C \leftrightarrow T$	$G \leftrightarrow T$
AAA	26.76	61.67	46.10	22.46	73.02	46.49
AAC	30.78	64.00	27.54	14.97	53.46	11.72
AAG	11.66	31.75	22.95	19.82	52.77	31.09
AAT	20.16	53.40	17.30	12.10	79.13	25.82
ACA	18.83	49.43	19.95	27.73	66.62	23.80
ACC	13.48	86.89	15.80	20.00	43.23	19.72
ACG	6.42	31.35	7.16	5.35	19.07	9.62
ACT	14.20	53.55	22.19	27.63	54.94	14.71
AGA	19.24	36.14	30.05	36.37	48.21	23.61
AGC	16.97	64.70	20.07	28.42	31.91	22.74
AGG	16.79	51.98	25.77	24.92	32.04	23.27
AGT	14.67	55.53	22.05	26.80	53.72	14.18
ATA	18.04	62.85	15.00	19.90	93.87	14.25
ATC	16.42	62.95	12.26	19.95	34.43	13.10
ATG	9.92	51.37	16.05	8.77	33.49	18.79
ATT	25.45	79.10	17.67	12.23	53.05	20.67
CAA	17.03	39.53	12.93	14.81	42.29	22.02
CAC	20.80	97.61	18.10	23.34	51.38	11.48
CAG	15.73	40.38	13.58	17.10	55.24	17.89
CAT	18.53	33.73	16.03	8.82	51.21	10.19
CCA	19.47	33.98	10.20	27.89	59.37	12.55
CCC	37.17	**120.52**	15.88	33.36	38.29	15.16
CCG	9.22	31.93	5.85	9.25	27.22	6.23
CCT	23.49	33.56	25.63	25.23	51.89	17.30
CGA	5.99	22.41	6.66	8.47	26.62	7.18
CGC	5.01	43.69	3.80	9.29	35.65	7.09
CGG	7.03	27.69	5.67	9.35	32.54	8.66
CGT	9.74	19.09	7.16	5.37	31.58	6.01
CTA	9.19	42.65	9.16	14.25	52.67	11.09
CTC	17.71	69.93	16.88	27.09	44.03	21.78
CTG	18.04	55.34	13.84	17.04	41.09	16.20
CTT	31.97	52.75	23.07	20.25	31.30	11.83
GAA	24.48	34.71	11.91	30.02	42.87	23.01
GAC	24.97	63.56	21.61	26.04	36.24	10.34
GAG	21.24	44.25	17.52	26.39	70.01	17.30
GAT	12.65	34.29	12.30	19.94	62.95	16.50
GCA	18.55	24.14	7.42	23.31	68.72	18.03
GCC	19.28	77.85	14.76	28.34	93.63	15.57
GCG	6.62	36.17	3.63	8.92	43.54	5.65
GCT	22.63	32.04	20.09	29.28	64.58	16.86

Tabela A.19: Valores do coeficiente ϕ do teste de ajustamento do χ^2, aplicado a cada trinucleótido e em cada SNV, considerando as posições da vizinhança $d = \pm 1 \ldots \pm 10$.

	SNV					
Trinucleótidos	$A \leftrightarrow C$	$A \leftrightarrow G$	$A \leftrightarrow T$	$C \leftrightarrow G$	$C \leftrightarrow T$	$G \leftrightarrow T$
GGA	18.19	29.91	10.67	18.88	77.70	19.55
GGC	15.18	**94.54**	14.75	28.45	78.44	19.88
GGG	14.65	38.75	15.69	33.89	**121.63**	36.76
GGT	19.06	44.00	15.49	19.95	87.40	13.29
GTA	9.52	38.59	15.74	15.02	80.24	29.53
GTC	10.47	36.77	21.42	26.64	62.71	25.18
GTG	12.11	51.94	17.90	22.36	97.55	20.03
GTT	12.20	53.60	27.33	14.27	64.03	31.08
TAA	15.20	82.06	11.86	13.30	48.51	21.71
TAC	30.03	79.39	14.81	14.99	38.47	9.74
TAG	11.07	53.05	9.86	15.03	43.01	9.14
TAT	13.91	93.53	14.43	19.84	63.48	17.52
TCA	16.17	37.50	9.95	16.93	64.25	13.30
TCC	19.56	77.75	10.20	19.55	30.42	18.65
TCG	7.23	27.65	6.92	8.66	23.16	6.87
TCT	23.67	48.10	30.04	37.15	35.83	19.49
TGA	13.23	64.39	9.68	16.90	37.44	15.85
TGC	18.33	68.80	7.70	23.57	24.87	19.12
TGG	12.73	60.18	10.48	27.87	34.22	19.57
TGT	24.20	67.09	19.91	27.10	50.12	17.82
TTA	21.79	49.02	11.95	12.69	**81.82**	14.85
TTC	23.13	43.18	11.80	30.73	34.99	24.45
TTG	22.03	41.94	13.13	14.40	39.54	16.71
TTT	45.79	73.14	45.66	22.37	60.96	24.85

Tabela A.20: Valores do coeficiente ϕ do teste de ajustamento do χ^2, aplicado a cada trinucleótido e em cada SNV, considerando as posições da vizinhança $d = \pm 1 \ldots \pm 10$. (Continuação)

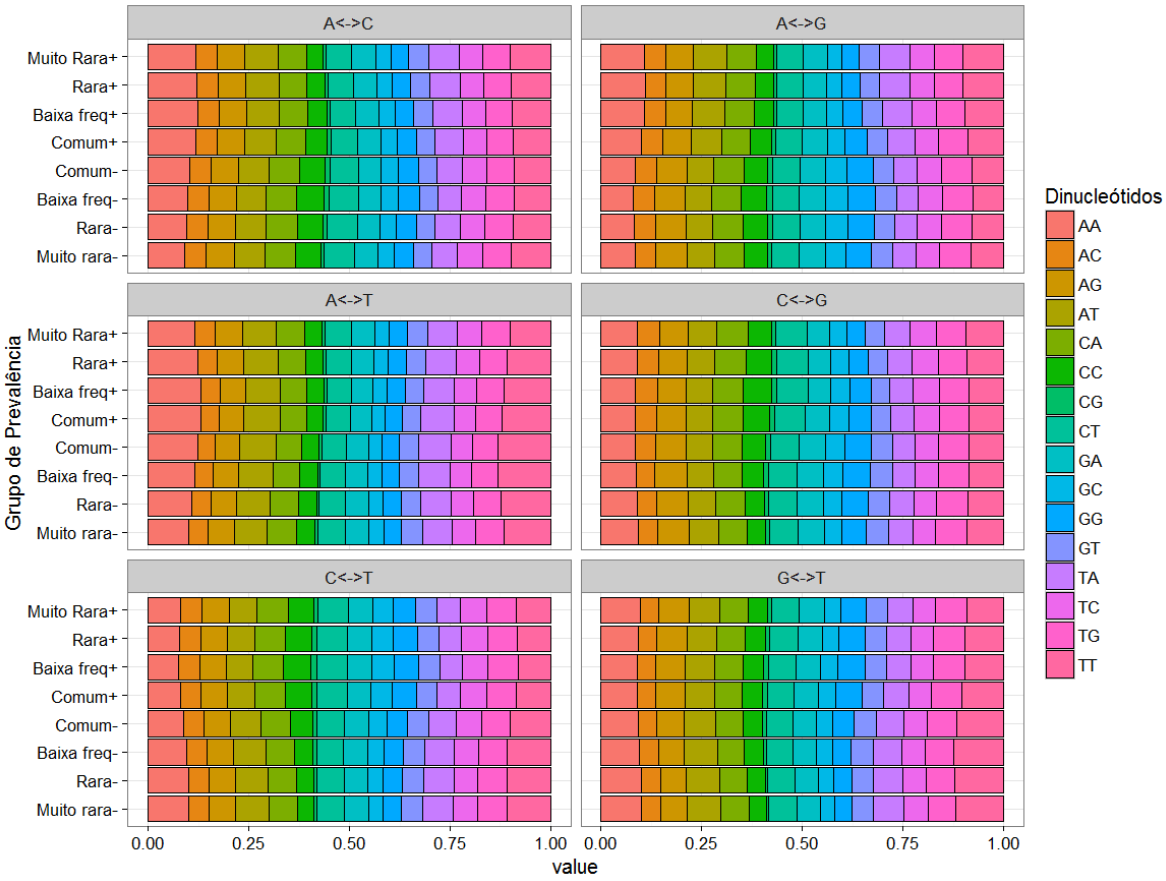


Figura A.9: Gráficos de barras das frequências relativas da contagem de dinucleótidos ($k = 2$) na vizinhança de cada SNV, por grupo de prevalência, considerando uma amplitude $w = 10$.

Apêndice B

Funções desenvolvidas em R

Neste apêndice apresentam-se as funções mais importantes que foram desenvolvidas em R, com uma breve explicação do funcionamento de cada uma, assim como, algum do código que foi efetuado para a análise deste trabalho.

Função B.1. `get.file`: Permite ler os dados resultantes do pré-processamento dos `VCFs` extraídos do projeto 1000G.

```
get.file <- function(path, nrows=-1) {
t <- read.table(path, header=TRUE, comment.char="",
colClasses=c("factor","integer","NULL","factor","factor",rep("integer",10)), nrows=nrows)
t$REF = factor(t$REF, NUCLEOTIDES)
t$ALT = factor(t$ALT, NUCLEOTIDES)
t=cbind(t[c(1:4)],t$"C0.0.",t$"C0.1."+t$"C1.0.",t$"C1.1.",t$"C0."+t$"C1."+rowSums(t[,11:14]))
names(t)[1] <- "CHROM"
names(t)[5:8] <- c("C0", "C1", "C2", "OTHER")
t
}
```

Função B.2. `norm.file`: Normaliza os dados lidos pela função anterior e faz a transformação das colunas.

```
norm.file <- function(tbl) {
u <- (as.character(tbl$REF) > as.character(tbl$ALT))
t <- tbl
t[u,] <- tbl[u, c(1,2, 4,3, 7,6,5, 8)]
names(t) = c("CHROM","POS","x","y","Cxx","Cxy","Cyy","OTHER")
t$var <- interaction(t$x, t$y, sep="<->", drop=TRUE, lex.order=TRUE)
n <- rowSums(t[,c("Cxx","Cxy","Cyy")])
nx <- (2*t$Cxx+t$Cxy)
t$Px <- nx/(2*n)
t$Py <- 1.0-t$Px
t
}
```

Função B.3. `ALL.var.Gr.Px`: Agrupa as variações de acordo com o grupo de prevalência.

```
ALL.var.Gr.Px<- function(w) {
q = w[,c(1,2,9,10)]
q1 = q[which(q$Px <= 0.001), ]
gr1 <- cbind(q1,as.matrix(rep("Muito␣rara-",dim(q1)[1])))
names(gr1)[5]=c("Grupos_Prevalência")
q2 <- q[which(q$Px > 0.001 & q$Px <= 0.005),  ]
gr2 <- cbind(q2,as.matrix(rep("Rara-",dim(q2)[1])))
names(gr2)[5]=c("Grupos_Prevalência")
q3 <- q[which(q$Px > 0.005 & q$Px <= 0.05),  ]
```

```
gr3 <- cbind(q3,as.matrix(rep("Baixa␣freq-",dim(q3)[1])))
names(gr3)[5]=c("Grupos_Prevalência")
q4 <- q[which(q$Px > 0.05 & q$Px <= 0.5),  ]
gr4 <- cbind(q4,as.matrix(rep("Comum-",dim(q4)[1])))
names(gr4)[5]=c("Grupos_Prevalência")
q5 <- q[which(q$Px > 0.5 & q$Px <= 0.95),  ]
gr5 <- cbind(q5,as.matrix(rep("Comum+",dim(q5)[1])))
names(gr5)[5]=c("Grupos_Prevalência")
q6 <- q[which(q$Px > 0.95 & q$Px <= 0.995),  ]
gr6 <- cbind(q6,as.matrix(rep("Baixa␣freq+",dim(q6)[1])))
names(gr6)[5]=c("Grupos_Prevalência")
q7 <- q[which(q$Px > 0.995 & q$Px <= 0.999),  ]
gr7 <- cbind(q7,as.matrix(rep("Rara+",dim(q7)[1])))
names(gr7)[5]=c("Grupos_Prevalência")
q8 = q[which(q$Px > 0.999), ]
gr8 <- cbind(q8,as.matrix(rep("Muito␣Rara+",dim(q8)[1])))
names(gr8)[5]=c("Grupos_Prevalência")
q.ALL.Gr=rbind(gr1,gr2,gr3,gr4,gr5,gr6,gr7,gr8)
q.ALL.Gr
}
```

Função B.4. `snp.neighborhood`: Mostra a sequência de nucleótidos na vinhança do local onde ocorreu a SNV.

```
snp.neighborhood <-
function(snptbl, winbeg=-10, winend=-winbeg, pattern="chr%s.fa") {
s <- rep(DNAStringSet(""), nrow(snptbl))
for (chr in levels(snptbl$CHROM)) {
chrfile <- sprintf(pattern, chr)
rec <- which(snptbl$CHROM==chr)
if (length(rec) > 0) {
seqset <- rep(readDNAStringSet(chrfile, nrec=1), length(rec))
s[rec,] <- subseq(seqset, pmax(1, snptbl[rec,"POS"]+winbeg),
pmin(width(seqset), snptbl[rec,"POS"]+winend))}}
s
}
```

Função B.5. `oligonucleotideFrequency.snp.neighborhood`: Contabiliza o número total de oligonucleótidos ($k = 1, 2, 3$) na vizinhança do local onde ocorreu a SNV.

```
oligonucleotideFrequency.snp.neighborhood <-
function(snptbl, k=1, win=10, pattern="chr%s.fa") {
sb <- snp.neighborhood(snptbl, winbeg=-win, winend=-1, pattern=pattern)
sa <- snp.neighborhood(snptbl, winbeg=+1, winend=+win, pattern=pattern)
freqs <- oligonucleotideFrequency(sb, k) + oligonucleotideFrequency(sa, k)
}
```

Função B.6. `sum.chr`: Calcula a soma das contagens dos oligonucleótidos ($k = 1, 2, 3$) na vizinhança do local onde ocorreu a SNV.

```
sum_chr<-
function(x, k , win) {
var = levels(x$var)
sum = matrix(nrow=length(var), ncol= 4^k)
rownames(sum) = var
for (s in var) {
var.file <- x[which(x$var==s), -c(4,5)]
freqs = oligonucleotideFrequency.snp.neighborhood(var.file, k, win)
d=ncol(freqs)
for (j in 1:d) {
sum[s, j]=as.numeric(sum(freqs[,j]))
colnames(sum)=colnames(freqs)}}
sum
}
```

Função B.7. `neighborhood.pattern`: Determina o padrão de frequências para cada variação contabilizando o número de oligonucleótidos em cada uma das posições da vizinhança do local onde ocorreu a SNV.

```
neighborhood.pattern <-function(tbl, k, win, pattern="chr%s.fa") {
center <- win+k
s <- snp.neighborhood(tbl, winbeg=-(win+k-1), winend=+(win+k-1), pattern=pattern)
m <- matrix(nrow = 2*win+1, ncol=4^k)
rownames(m) = -win:win
levels = cart.pow(NUCLEOTIDES, k)
show(levels)
for (d in 1:win) {
trb <- table(factor(as.character(subseq(s, center-(d+k-1), center-d)), levels=levels))
m[win+1-d, ] <- trb
tra <- table(factor(as.character(subseq(s, center+d, center+(d+k-1))), levels=levels))
m[win+1+d, ] <- tra
stopifnot(names(tra) == names(trb))}
colnames(m) <- names(tra)
m}
```

Função B.8. `sum.chr.grp`: Calcula a soma das contagens dos oligonucleótidos ($k = 1, 2, 3$) na vizinhança do local onde ocorreu a SNV, de acordo com o grupo de prevalência.

```
sum.chr.grp<- function(x, k , win) {
grp = levels(x$Grupos_Prevalência)
sum.grp = matrix(nrow=length(grp), ncol= 4^k)
rownames(sum.grp) = grp
for(s in grp) {
grp.file <- x[which(x$Grupos_Prevalência==s), ]
freqs = oligonucleotideFrequency.snp.neighborhood(grp.file, k, win)
for(j in 1:ncol(freqs)) {
sum.grp[s, j]= colSums(freqs)[j]
colnames(sum.grp)=colnames(freqs)}}
sum.grp}
```

Código em R

Packages, funções auxiliares e leitura dos dados.

```
library(BSgenome)
library("BSgenome.Hsapiens.UCSC.hg19")
library(Biostrings)
library(ggplot2)
library(gplots)
library(xtable)
library(reshape2)
library(vcd)
library(multtest)
library(HardyWeinberg)

Phi=function(z){as.numeric(sqrt(chisq.test(z)$statistic/(sum(z))))}

Resid_ajust=function(z){
Vr=matrix(nrow=nrow(z), ncol=ncol(z))
for (i in 1:nrow(z)) for(j in 1:ncol(z))
Vr[i,j]=((1-(rowSums(z))[i]/sum(z))*(1-(colSums(z))[j]/sum(z)))
Ra=matrix(nrow=nrow(z), ncol=ncol(z))
colnames(Ra)=colnames(z)
rownames(Ra)=rownames(z)
for (i in 1:nrow(z)) for(j in 1:ncol(z))
Ra[i,j]= (chisq.test(z)$residuals)[i,j]/sqrt(Vr[i,j])
Ra}
```

```
HWE.data=function(z){HWdata=as.matrix(z[, 5:7])
colnames(HWdata)=c("xx","xy","yy")
z=HWdata
z}

table.HWE.chr=function(z){table (z <0.05)}

prop.table.HWE.chr=function(z){round(prop.table (z)*100, 2)}

get.var=function(x,v){x[which(x$var==v), ]}

cart.prod <- function(f) {
sort( apply( expand.grid(f), 1, function(x) {paste(x,collapse="")} ) )}

cart.pow <- function(a, exp) {exp <- as.integer(exp)
if (exp == 0) {r <- ""
} else {
r <- cart.prod(a, cart.pow(a, exp-1))}
r}

filenames.chr=c("chr1-snp-stats.txt","chr2-snp-stats.txt","chr3-snp-stats.txt",
"chr4-snp-stats.txt","chr5-snp-stats.txt","chr6-snp-stats.txt","chr7-snp-stats.txt",
"chr8-snp-stats.txt","chr9-snp-stats.txt","chr10-snp-stats.txt","chr11-snp-stats.txt",
"chr12-snp-stats.txt","chr13-snp-stats.txt","chr14-snp-stats.txt","chr15-snp-stats.txt",
"chr16-snp-stats.txt","chr17-snp-stats.txt","chr18-snp-stats.txt","chr19-snp-stats.txt",
"chr20-snp-stats.txt","chr21-snp-stats.txt","chr22-snp-stats.txt")

list.var.chr=list()
for (i in filenames.chr){list.var.chr[[i]]= norm.file(get.file(i))}
list.var.Prev=lapply(list.var.chr, function(q){q=transform(q, Prev = pmin(q$Px, q$Py))})
```

Análise global das variações.

```
ALL.var = rbind(list.var.Prev[[1]], list.var.Prev[[2]],list.var.Prev[[3]],
list.var.Prev[[4]],list.var.Prev[[5]], list.var.Prev[[6]],list.var.Prev[[7]],
list.var.Prev[[8]],list.var.Prev[[9]],list.var.Prev[[10]], list.var.Prev[[11]],
list.var.Prev[[12]], list.var.Prev[[13]], list.var.Prev[[14]], list.var.Prev[[15]],
list.var.Prev[[16]], list.var.Prev[[17]], list.var.Prev[[18]],
list.var.Prev[[19]],list.var.Prev[[20]],list.var.Prev[[21]],list.var.Prev[[22]])

Frq.abs=table(ALL.var$CHROM,ALL.var$var)
fa=addmargins(Frq.abs)
dataframe.f.abs <- melt(Frq.abs)
boxplot1=ggplot(dataframe.f.abs, aes(x=Var2,y=value, fill=Var2))+geom_boxplot()+
theme_bw()+ labs(fill = "Variação")+ylab("Frequência␣Absoluta")
+ xlab("Variação")
print(boxplot1)
total_abs_chrom=as.vector(rowSums(Frq.abs))
Frq.rel= Frq.abs/total_abs_chrom
dataframe.f.rel <- melt(Frq.rel)
barplot1 <- ggplot(ALL.var, aes(CHROM)) + geom_bar(colour="black", aes(fill = var),
position = "fill")+ theme_bw() + labs(fill="Variação")
+ ylab("Frequência␣Relativa") + xlab("Cromossoma")
print(barplot1)
summary(assocstats(Frq.abs))
Resi_aj=chisq.test(Frq.abs)$stdres
dataf.resid.ajust<- melt(Resi_aj)
m=mean(dataf.resid.ajust$value)
sd=sd(dataf.resid.ajust$value)
hist_resid.ajust1= ggplot(dataf.resid.ajust, aes(value)) + scale_x_continuous(limits =
c(-75, 125))+ geom_histogram(aes(y = ..density..),color="black", fill="lightblue",
binwidth=10)+ theme_bw() + xlab("Resíduos␣Ajustados")
+ ylab("Frequência␣Absoluta") + stat_function(fun=dnorm, args=list(mean= m, sd=sd ))
print(hist_resid.ajust1)
y= quantile(dataf.resid.ajust$value, c(0.25, 0.75))
```

```
x= qnorm( c(0.25, 0.75))
slope=diff(y) / diff(x)
int=y[1] - slope * x[1]
qqplot_resid.ajust1= ggplot(dataf.resid.ajust, aes(sample = value, colour = Var2))
+ stat_qq(size=2.5)+ theme_bw()+labs(colour="Variação") + ylab("Resíduos␣ajustados")
+ xlab("Quantis␣teóricos") + geom_abline(intercept=int, slope=slope)
print(qqplot_resid.ajust1)
heatmap.2(Resi_aj,col=brewer.pal(11,"RdYlGn"),trace="none")
```

Estudo da variável prevalência.

```
Prev= c(list.var.Prev[[1]]$Prev,list.var.Prev[[2]]$Prev,list.var.Prev[[3]]$Prev,
list.var.Prev[[4]]$Prev, list.var.Prev[[5]]$Prev,list.var.Prev[[6]]$Prev,
list.var.Prev[[7]]$Prev, list.var.Prev[[8]]$Prev,list.var.Prev[[9]]$Prev,
list.var.Prev[[10]]$Prev, list.var.Prev[[11]]$Prev, list.var.Prev[[12]]$Prev,
list.var.Prev[[13]]$Prev, list.var.Prev[[14]]$Prev, list.var.Prev[[15]]$Prev,
list.var.Prev[[16]]$Prev, list.var.Prev[[17]]$Prev, list.var.Prev[[18]]$Prev,
list.var.Prev[[19]]$Prev, list.var.Prev[[20]]$Prev,list.var.Prev[[21]]$Prev,
list.var.Prev[[22]]$Prev)

table(Prev==0)
hist_Prev=hist(Prev, col="ligthgreen", xlab="Prevalência", ylab="Frequência␣Absoluta",
main="Histograma␣da␣prevalência␣das␣variações")
t1=table(Prev <= 0.02)
s=summary(Prev)
t2=table(Prev <= s[2])
boxplot_var_semilog= ggplot(ALL.var,aes(var,Prev,fill=var))+ geom_boxplot()+
ylab("Prevalência") + xlab("Variação") +scale_y_log10()
print(boxplot_var_semilog)

vector_var=levels(ALL.var$var)
l_prev=list()
for ( i in vector_var) { l_prev[[i]] = ALL.var[which(ALL.var$var == i),"Prev"]}
test_KS_D=matrix(nrow=length(vector_var),ncol=length(vector_var))
test_KS_p=matrix(nrow=length(vector_var),ncol=length(vector_var))
colnames(test_KS_D)=vector_var
rownames(test_KS_D)=vector_var
colnames(test_KS_p)=vector_var
rownames(test_KS_p)=vector_var
for ( i in vector_var){for ( j in vector_var){ks = ks.test(l_prev[[i]],l_prev[[j]])
test_KS_D[i,j] = ks$statistic
test_KS_p[i,j] = ks$p.value}}
test_KS_D
test_KS_p
vector_pvals = as.vector(test_KS_p)
pvals_adjstBONF = p.adjust(vector_pvals, "bonferroni")
pvals_adjst = mt.rawp2adjp(vector_pvals, c("Bonferroni","SidakSS") )
allpvals=as.matrix(pvals_adjst$adjp)
```

Análise por grupo de prevalência.

```
ALL.var.Gr=ALL.var.Gr.Px(ALL.var)
Frq.abs.Grp=table(ALL.var.Gr$Grupos_Prevalência,ALL.var.Gr$var)
total_abs_var=colSums(Frq.abs.Grp)
(total_abs_var/nrow(ALL.var.Gr))*100
Frq.rel.Var= t(t(Frq.abs.Grp)/colSums(Frq.abs.Grp))
dataf.Frq.rel.Var<- melt(Frq.rel.Var)
Frq.rel.Grp= Frq.abs.Grp/rowSums(Frq.abs.Grp)
barplot.grp <- ggplot(ALL.var.Gr, aes(Grupos_Prevalência)) + geom_bar(colour="black",
aes(fill = var), position = "fill")+ coord_flip()+ theme_bw() + labs(fill="Variação")
+ ylab("Frequência␣Relativa") + xlab("Grupo␣de␣Prevalência")
print(barplot.grp)

summary(assocstats(Frq.abs.Grp))
Resi_aj_gr=chisq.test(Frq.abs.Grp)$stdres
dataf.resid.ajustGrp<- melt(Resi_aj_gr)
```

```
m.grp=mean(dataf.resid.ajustGrp$value)
sd.grp=sd(dataf.resid.ajustGrp$value)
hist_resid.ajustgrp= ggplot(dataf.resid.ajustGrp, aes(value)) +
geom_histogram(aes(y = ..density..),color="black", fill="lightblue", binwidth=40)+
theme_bw() + xlab("Resíduos␣Ajustados") + stat_function(fun=dnorm,
args=list(mean= m.grp, sd=sd.grp ))
print(hist_resid.ajustgrp)
heatmap.2(Resi_aj_gr,srtRow=45,  offsetRow=-0.5, offsetCol=0,
 col=brewer.pal(11,"RdYlGn"),trace="none")
```

Equilíbrio de Hardy-Weinberg

```
list.HWE.data.chr=lapply(list.var.chr, HWE.data)
list.HWE.Chisq.stat.chr=lapply(list.HWE.data.chr, function(z){HWChisqStats(z,
x.linked=FALSE, pvalues=F)} )
list.HWE.Chisqp.values.chr=lapply(list.HWE.data.chr, function(z){HWChisqStats(z,
x.linked=FALSE, pvalues=TRUE)} )
list.HWE.Chisq.D.chr=lapply(list.HWE.data.chr,function(z){HWChisqMat
(z,cc=0,verbose=F)$Dvec} )

vector.HWE.Chisq.stat.chr=c(list.HWE.Chisq.stat.chr[[1]],
list.HWE.Chisq.stat.chr[[2]], list.HWE.Chisq.stat.chr[[3]],
list.HWE.Chisq.stat.chr[[4]], list.HWE.Chisq.stat.chr[[5]],
list.HWE.Chisq.stat.chr[[6]], list.HWE.Chisq.stat.chr[[7]],
list.HWE.Chisq.stat.chr[[8]], list.HWE.Chisq.stat.chr[[9]],
list.HWE.Chisq.stat.chr[[10]], list.HWE.Chisq.stat.chr[[11]],
list.HWE.Chisq.stat.chr[[12]], list.HWE.Chisq.stat.chr[[13]],
list.HWE.Chisq.stat.chr[[14]],list.HWE.Chisq.stat.chr[[15]],
list.HWE.Chisq.stat.chr[[16]], list.HWE.Chisq.stat.chr[[17]],
list.HWE.Chisq.stat.chr[[18]], list.HWE.Chisq.stat.chr[[19]],
list.HWE.Chisq.stat.chr[[20]],list.HWE.Chisq.stat.chr[[21]],
list.HWE.Chisq.stat.chr[[22]])

vector.HWE.Chisq.p.values.chr=c(list.HWE.Chisqp.values.chr[[1]],
list.HWE.Chisqp.values.chr[[2]], list.HWE.Chisqp.values.chr[[3]],
list.HWE.Chisqp.values.chr[[4]], list.HWE.Chisqp.values.chr[[5]],
list.HWE.Chisqp.values.chr[[6]], list.HWE.Chisqp.values.chr[[7]],
list.HWE.Chisqp.values.chr[[8]], list.HWE.Chisqp.values.chr[[9]],
list.HWE.Chisqp.values.chr[[10]], list.HWE.Chisqp.values.chr[[11]],
list.HWE.Chisqp.values.chr[[12]],  list.HWE.Chisqp.values.chr[[13]],
list.HWE.Chisqp.values.chr[[14]], list.HWE.Chisqp.values.chr[[15]],
list.HWE.Chisqp.values.chr[[16]],  list.HWE.Chisqp.values.chr[[17]],
list.HWE.Chisqp.values.chr[[18]],  list.HWE.Chisqp.values.chr[[19]],
list.HWE.Chisqp.values.chr[[20]], list.HWE.Chisqp.values.chr[[21]],
list.HWE.Chisqp.values.chr[[22]])

table(list.HWE.Chisqp.values.chr[[1]]<0.05)
which(is.na(list.HWE.Chisqp.values.chr[[1]]))
options(digits=10)
summary(vector.HWE.Chisq.p.values.chr)
summary(vector.HWE.Chisq.stat.chr)
t=table(vector.HWE.Chisq.p.values.chr<0.05)
list.HWE.Chisq.stat.chr.summary=lapply(list.HWE.Chisq.stat.chr,summary,digits=3)
list.HWE.Chisqp.values.chr.summary=lapply(list.HWE.Chisqp.values.chr,summary,digits=3)

HWGenotypePlot(list.HWE.data.chr[[16]],plottype=1, pch=22, xlab="pxx", ylab="pxy",
main="Frequências␣heterozigóticas/homozigóticas")
HWGenotypePlot(list.HWE.data.chr[[16]],plottype=2, pch=22, xlab="pxx", ylab="pyy",
main="Frequências␣homozigóticas/homozigóticas")

list.table.HWE.chr=lapply(list.HWE.Chisqp.values.chr, table.HWE.chr)
list.prop.table.HWE.chr=lapply(list.table.HWE.chr, prop.table.HWE.chr)
HWTernaryPlot(list.HWE.data.chr[[11]], vbounds=T, pch = 19)
HWTernaryPlot(list.HWE.data.chr[[22]], vbounds=T, pch = 19)
```

```
vector_var=ALL.var$var
data.stats.Var=data.frame(vector_var, vector.HWE.Chisq.stat.chr)
data.pvalues.Var=data.frame(vector_var, vector.HWE.Chisq.p.values.chr)

data.stats_AC= data.stats.Var[which(data.stats.Var$vector_var == "A<->C" ),  ]
data.pvalues_AC= data.pvalues.Var[which(data.pvalues.Var$vector_var == "A<->C" ),  ]
t_AC=table(data.pvalues_AC$vector.HWE.Chisq.p.values.chr < 0.05)
t.prop_AC=round(prop.table (t_AC)*100, 2)
options(digits=10)
summary(data.pvalues_AC$vector.HWE.Chisq.p.values.chr)
summary(data.stats_AC$vector.HWE.Chisq.stat.chr)

data.stats_AG= data.stats.Var[which(data.stats.Var$vector_var == "A<->G" ),  ]
data.pvalues_AG= data.pvalues.Var[which(data.pvalues.Var$vector_var == "A<->G" ),  ]
t_AG=table(data.pvalues_AG$vector.HWE.Chisq.p.values.chr < 0.05)
t.prop_AG=round(prop.table (t_AG)*100, 2)
options(digits=10)
summary(data.pvalues_AG$vector.HWE.Chisq.p.values.chr)
summary(data.stats_AG$vector.HWE.Chisq.stat.chr)

data.stats_AT= data.stats.Var[which(data.stats.Var$vector_var == "A<->T" ),  ]
data.pvalues_AT= data.pvalues.Var[which(data.pvalues.Var$vector_var == "A<->T" ),  ]
t_AT=table(data.pvalues_AT$vector.HWE.Chisq.p.values.chr < 0.05)
t.prop_AT=round(prop.table (t_AT)*100, 2)
options(digits=10)
summary(data.pvalues_AT$vector.HWE.Chisq.p.values.chr)
summary(data.stats_AT$vector.HWE.Chisq.stat.chr)

data.stats_CG= data.stats.Var[which(data.stats.Var$vector_var == "C<->G" ),  ]
data.pvalues_CG= data.pvalues.Var[which(data.pvalues.Var$vector_var == "C<->G" ),  ]
t_CG=table(data.pvalues_CG$vector.HWE.Chisq.p.values.chr < 0.05)
t.prop_CG=round(prop.table (t_CG)*100, 2)
options(digits=10)
summary(data.pvalues_CG$vector.HWE.Chisq.p.values.chr)
summary(data.stats_CG$vector.HWE.Chisq.stat.chr)

data.stats_CT= data.stats.Var[which(data.stats.Var$vector_var == "C<->T" ),  ]
data.pvalues_CT= data.pvalues.Var[which(data.pvalues.Var$vector_var == "C<->T" ),  ]
t_CT=table(data.pvalues_CT$vector.HWE.Chisq.p.values.chr < 0.05)
t.prop_CT=round(prop.table (t_CT)*100, 2)
options(digits=10)
summary(data.pvalues_CT$vector.HWE.Chisq.p.values.chr)
summary(data.stats_CT$vector.HWE.Chisq.stat.chr)

data.stats_GT= data.stats.Var[which(data.stats.Var$vector_var == "G<->T" ),  ]
data.pvalues_GT= data.pvalues.Var[which(data.pvalues.Var$vector_var == "G<->T" ),  ]
t_GT=table(data.pvalues_GT$vector.HWE.Chisq.p.values.chr < 0.05)
t.prop_GT=round(prop.table (t_GT)*100, 2)
options(digits=10)
summary(data.pvalues_GT$vector.HWE.Chisq.p.values.chr)
summary(data.stats_GT$vector.HWE.Chisq.stat.chr)

data.pvaluesTrans= data.pvalues.Var[which(data.pvalues.Var$vector_var == "A<->G" |
 data.pvalues.Var$vector_var == "C<->T"),  ]
t_Trans=table(data.pvaluesTrans$vector.HWE.Chisq.p.values.chr < 0.05)
t.prop_Trans=round(prop.table (t_Trans)*100, 2)

data.pvaluesTransV= data.pvalues.Var[which(data.pvalues.Var$vector_var == "A<->T" |
data.pvalues.Var$vector_var == "A<->C" | data.pvalues.Var$vector_var == "C<->G" |
data.pvalues.Var$vector_var == "G<->T"),  ]

data.pvaluesTransV=rbind(data.pvalues_AT,data.pvalues_AC,data.pvalues_CG,data.pvalues_GT)
t_TransV=table(data.pvaluesTransV$vector.HWE.Chisq.p.values.chr < 0.05)
t.prop_TransV=round(prop.table (t_TransV)*100, 8)
```

Análise global do contexto na vizinhança de cada SNV.

```
list.sum1<- lapply(list.var.chr,function(z) {sum.chr(z, 1, 5) } )
ALLSum1_5=Reduce('+', list.sum1)
list.sum2<- lapply(list.var.chr, function(z) {sum.chr(z, 2, 5) } )
ALLSum2_5=Reduce('+', list.sum2)
list.sum3<- lapply(list.var.chr, function(z) {sum.chr(z, 3, 5) } )
ALLSum3_5=Reduce('+', list.sum3)
list.sum1<- lapply(list.var.chr,function(z) {sum.chr(z, 1, 10) } )
ALLSum1_10=Reduce('+', list.sum1)
list.sum2<- lapply(list.var.chr, function(z) {sum.chr(z, 2, 10) } )
ALLSum2_10=Reduce('+', list.sum2)
list.sum3<- lapply(list.var.chr, function(z) {sum.chr(z, 3, 10) } )
ALLSum3_10=Reduce('+', list.sum3)
list.sum1<- lapply(list.var.chr,function(z) {sum.chr(z, 1, 20) } )
ALLSum1_20=Reduce('+', list.sum1)
list.sum2<- lapply(list.var.chr, function(z) {sum.chr(z, 2, 20) } )
ALLSum2_20=Reduce('+', list.sum2)
list.sum3<- lapply(list.var.chr, function(z) {sum.chr(z, 3, 20) } )
ALLSum3_20=Reduce('+', list.sum3)
list.sum1<- lapply(list.var.chr,function(z) {sum.chr(z, 1, 50) } )
ALLSum1_50=Reduce('+', list.sum1)
list.sum2<- lapply(list.var.chr, function(z) {sum.chr(z, 2, 50) } )
ALLSum2_50=Reduce('+', list.sum2)
list.sum3<- lapply(list.var.chr, function(z) {sum.chr(z, 3, 50) } )
ALLSum3_50=Reduce('+', list.sum3)
list.sum1<- lapply(list.var.chr, function(z) {sum.chr(z, 1, 100) })
ALLSum1_100=Reduce('+', list.sum1)
list.sum2<- lapply(list.var.chr, function(z) {sum.chr(z, 2, 100) })
ALLSum2_100=Reduce('+', list.sum2)
list.sum3<- lapply(list.var.chr, function(z) {sum.chr(z, 3, 100) })
ALLSum3_100=Reduce('+', list.sum3)
list.sum1<- lapply(list.var.chr, function(z) {sum.chr(z, 1, 200) })
ALLSum1_200=Reduce('+', list.sum1)
list.sum2<- lapply(list.var.chr, function(z) {sum.chr(z, 2, 200) })
ALLSum2_200=Reduce('+', list.sum2)
list.sum3<- lapply(list.var.chr function(z) {sum.chr(z, 3, 200)})
ALLSum3_200=Reduce('+', list.sum3)

list.ALLSum.5=list(ALLSum1_5,ALLSum2_5,ALLSum3_5)
list.ALLSum.10=list(ALLSum1_10,ALLSum2_10,ALLSum3_10)
list.ALLSum.20=list(ALLSum1_20,ALLSum2_20,ALLSum3_20)
list.ALLSum.50=list(ALLSum1_50,ALLSum2_50,ALLSum3_50)
list.ALLSum.100=list(ALLSum1_100,ALLSum2_100,ALLSum3_100)
list.ALLSum.200=list(ALLSum1_200,ALLSum2_200,ALLSum3_200)
list.ALLSum.5.FrqRel=lapply(list.ALLSum.5, function(w){prop.table(w,1)})
list.ALLSum.10.FrqRel=lapply(list.ALLSum.10, function(w){prop.table(w,1)})
list.ALLSum.20.FrqRel=lapply(list.ALLSum.20, function(w){prop.table(w,1)})
list.ALLSum.50.FrqRel=lapply(list.ALLSum.50, function(w){prop.table(w,1)})
list.ALLSum.100.FrqRel=lapply(list.ALLSum.100, function(w){prop.table(w,1)})
list.ALLSum.200.FrqRel=lapply(list.ALLSum.200, function(w){prop.table(w,1)})
list.fr.ALL5=lapply(list.ALLSum.5.FrqRel, melt)
list.fr.ALL10=lapply(list.ALLSum.10.FrqRel, melt)
list.fr.ALL20=lapply(list.ALLSum.20.FrqRel, melt)
list.fr.ALL50=lapply(list.ALLSum.50.FrqRel, melt)
list.fr.ALL100=lapply(list.ALLSum.100.FrqRel, melt)
list.fr.ALL200=lapply(list.ALLSum.200.FrqRel, melt)

vector.win<-c('5', '10', '20','50','100','200')
var.data.w1=as.vector(rep(vector.win,each=6*4^1))
var.data.w2=as.vector(rep(vector.win,each=6*(4^2)))
var.data.w3=as.vector(rep(vector.win,each=6*(4^3)))

barplot1.ALL5 <- ggplot(list.fr.ALL5[[1]], aes(x=Var1,value,fill=Var2)) +
geom_bar(colour="black",stat = "identity" )+
theme_bw()+ scale_fill_brewer(palette="Set3")+ labs(fill="Nucleótidos")+
```

```
xlab("Variação")
print(barplot1.ALL5)
barplot1.ALL100 <- ggplot(list.fr.ALL100[[1]], aes(x=Var1,value,fill=Var2)) +
geom_bar(colour="black",stat = "identity" )+
theme_bw()+ scale_fill_brewer(palette="Set3")+ labs(fill="Nucleótidos") +
xlab("Variação")
print(barplot1.ALL100)
barplot2.ALL10 <- ggplot(list.fr.ALL10[[2]], aes(x=Var1,value,fill=Var2)) +
geom_bar(colour="black",stat = "identity" )+
theme_bw()+ labs(fill="Dinucleótidos") + xlab("Variação")
print(barplot2.ALL10)
barplot2.ALL200 <- ggplot(list.fr.ALL200[[2]], aes(x=Var1,value,fill=Var2)) +
geom_bar(colour="black",stat = "identity" )+
theme_bw()+  labs(fill="Dinucleótidos") + xlab("Variação")
print(barplot2.ALL200)
barplot3.ALL10 <- ggplot(list.fr.ALL10[[3]], aes(x=Var1,value,fill=Var2)) +
geom_bar(colour="black",stat = "identity" )+
theme_bw()+ labs(fill="Trinucleótidos") + xlab("Variação")
print(barplot3.ALL10)

list.chisqtest.ALL5=lapply(list.ALLSum.5, assocstats)
list.chisqtest.ALL10=lapply(list.ALLSum.10, assocstats)
list.chisqtest.ALL20=lapply(list.ALLSum.20, assocstats)
list.chisqtest.ALL50=lapply(list.ALLSum.50, assocstats)
list.chisqtest.ALL100=lapply(list.ALLSum.100, assocstats)
list.chisqtest.ALL200=lapply(list.ALLSum.200, assocstats)
list.Resi_aj.ALL5=lapply(list.ALLSum.5,Resid_ajust)
list.Resi_aj.ALL10=lapply(list.ALLSum.10,Resid_ajust)
list.Resi_aj.ALL20=lapply(list.ALLSum.20,Resid_ajust)
list.Resi_aj.ALL50=lapply(list.ALLSum.50,Resid_ajust)
list.Resi_aj.ALL100=lapply(list.ALLSum.100,Resid_ajust)
list.Resi_aj.ALL200=lapply(list.ALLSum.200,Resid_ajust)

L.data.Resi_aj.ALL5=lapply(list.Resi_aj.ALL5, melt)
L.data.Resi_aj.ALL10=lapply(list.Resi_aj.ALL10, melt)
L.data.Resi_aj.ALL20=lapply(list.Resi_aj.ALL20, melt)
L.data.Resi_aj.ALL50=lapply(list.Resi_aj.ALL50, melt)
L.data.Resi_aj.ALL100=lapply(list.Resi_aj.ALL100, melt)
L.data.Resi_aj.ALL200=lapply(list.Resi_aj.ALL200, melt)
heatmap.2(list.Resi_aj.ALL5[[1]],col=brewer.pal(11,"RdYlGn"),trace="none")
heatmap.2(list.Resi_aj.ALL100[[1]],col=brewer.pal(11,"RdYlGn"),trace="none")
heatmap.2(list.Resi_aj.ALL10[[2]],col=brewer.pal(11,"RdYlGn"),trace="none")
heatmap.2(list.Resi_aj.ALL200[[2]],col=brewer.pal(11,"RdYlGn"),trace="none")
heatmap.2(list.Resi_aj.ALL10[[3]],col=brewer.pal(11,"RdYlGn"),trace="none")
```

Frequências globais dos nuclótidos no genoma.

```
Tbl_frq_esp_Nucl = alphabetFrequency(Hsapiens$chr1)+alphabetFrequency(Hsapiens$chr2)+
alphabetFrequency(Hsapiens$chr3)+alphabetFrequency(Hsapiens$chr4)+
alphabetFrequency(Hsapiens$chr5)+alphabetFrequency(Hsapiens$chr6)+
alphabetFrequency(Hsapiens$chr7)+alphabetFrequency(Hsapiens$chr8)+
alphabetFrequency(Hsapiens$chr9)+alphabetFrequency(Hsapiens$chr10)+
alphabetFrequency(Hsapiens$chr11)+alphabetFrequency(Hsapiens$chr12)+
alphabetFrequency(Hsapiens$chr13)+alphabetFrequency(Hsapiens$chr14)+
alphabetFrequency(Hsapiens$chr15)+alphabetFrequency(Hsapiens$chr16)+
alphabetFrequency(Hsapiens$chr17)+alphabetFrequency(Hsapiens$chr18)+
alphabetFrequency(Hsapiens$chr19)+alphabetFrequency(Hsapiens$chr20)+
alphabetFrequency(Hsapiens$chr21)+alphabetFrequency(Hsapiens$chr22)+
alphabetFrequency(Hsapiens$chrX)+alphabetFrequency(Hsapiens$chrY)
nucl=as.vector(Tbl_frq_esp_Nucl)
names(nucl)=names(Tbl_frq_esp_Nucl)
s1=sum(as.numeric(nucl)[1:4])
nucl_relat=((nucl/s1)*100)[1:4]
```

Análise dos padrões de frequência de cada SNV.

```
list.chr.AC = lapply(list.var.chr, function(z){get.var(z,'A<->C')})
list.chr.AG = lapply(list.var.chr, function(z){get.var(z,'A<->G')})
list.chr.AT = lapply(list.var.chr, function(z){get.var(z,'A<->T')})
list.chr.CG = lapply(list.var.chr, function(z){get.var(z,'C<->G')})
list.chr.CT = lapply(list.var.chr, function(z){get.var(z,'C<->T')})
list.chr.GT = lapply(list.var.chr, function(z){get.var(z,'G<->T')})

list.pattern_1_AC= lapply(list.chr.AC, function(z) {neighborhood.pattern(z, 1, 20) })
ALLpattern_1_AC=Reduce('+', list.pattern_1_AC)
list.pattern_1_AG= lapply(list.chr.AG, function(z) {neighborhood.pattern(z, 1, 20) })
ALLpattern_1_AG=Reduce('+', list.pattern_1_AG)
list.pattern_1_AT= lapply(list.chr.AT, function(z) {neighborhood.pattern(z, 1, 20) })
ALLpattern_1_AT=Reduce('+', list.pattern_1_AT)
list.pattern_1_CG= lapply(list.chr.CG, function(z) {neighborhood.pattern(z, 1, 20) })
ALLpattern_1_CG=Reduce('+', list.pattern_1_CG)
list.pattern_1_CT= lapply(list.chr.CT, function(z) {neighborhood.pattern(z, 1, 20) })
ALLpattern_1_CT=Reduce('+', list.pattern_1_CT)
list.pattern_1_GT= lapply(list.chr.GT, function(z) {neighborhood.pattern(z, 1, 20) })
ALLpattern_1_GT=Reduce('+', list.pattern_1_GT)

list.ALLvar.nucl=list(ALLpattern_1_AC, ALLpattern_1_AG, ALLpattern_1_AT,
ALLpattern_1_CG,
ALLpattern_1_CT, ALLpattern_1_GT )
list.ALLvar.nucl.fr=lapply(list.ALLvar.nucl, function(z){prop.table(z, 1)})

vector.var=c('A<->C','A<->G','A<->T','C<->G','C<->T','G<->T')
df.vector.var=as.vector(rep(vector.var,each= nrow(ALLpattern_1_AC)*4^1))
list.data.var.nucl.fr= lapply(list.ALLvar.nucl.fr, melt)
frq_exp_nucl=c(0.2952696, 0.204455, 0.2045758, 0.2956996 )
vector.frq_exp_nucl=as.vector(rep(frq_exp_nucl, each= nrow(ALLpattern_1_AC)))
list.data.var.nucl.fr_F=lapply(list.data.var.nucl.fr, function(z){cbind(z,
vector.frq_exp_nucl)})
data.ALL.nucl= as.data.frame(cbind(rbind(list.data.var.nucl.fr_F[[1]],
list.data.var.nucl.fr_F[[2]], list.data.var.nucl.fr_F[[3]],
list.data.var.nucl.fr_F[[4]],list.data.var.nucl.fr_F[[5]],
list.data.var.nucl.fr_F[[6]]), df.vector.var))
data.ALL.nucl_F=cbind(data.ALL.nucl, data.ALL.nucl$value -
data.ALL.nucl$vector.frq_exp_nucl)
names(data.ALL.nucl_F)=c('Position', 'Nucleotide','obs','exp', 'var', 'bias')

geomline.pattern_AC_nucl=ggplot(data.ALL.nucl_F[which(data.ALL.nucl_F$var
=="A<->C"),],
aes(Position, bias, group=Nucleotide, shape=Nucleotide, color=Nucleotide)) +
geom_line(size=1.25) + geom_point(size=2)+
xlab("Posição")+ylab("Viés")+labs(group="Nucleótido")+
ggtitle("Variação␣A<->C")+ scale_color_brewer(palette="Set3")+scale_y_continuous
(limits=c(-0.1, 0.15), breaks = seq(-0.1, 0.15, by=0.01))+
scale_x_continuous(breaks = seq(-20, 20, by=1) ) + theme(panel.background =
element_rect(fill = "white", color = "black", size = 1.1)) +
geom_vline(xintercept = 0 , color="grey85", linetype="dashed", size=0.7) +
geom_hline( yintercept=0, color="grey60", size=0.7)
print(geomline.pattern_AC_nucl)
geomline.pattern_AG_nucl=ggplot(data.ALL.nucl_F[which(data.ALL.nucl_F$var
=="A<->G"),],
aes(Position, bias, group=Nucleotide, shape=Nucleotide, color=Nucleotide)) +
geom_line(size=1.25) + geom_point(size=2)+
xlab("Posição")+ylab("Viés")+labs(group="Nucleótido")+
ggtitle("Variação␣A<->G")+ scale_color_brewer(palette="Set3") +scale_y_continuous
(limits=c(-0.1, 0.15), breaks = seq(-0.1, 0.15, by=0.01))+
scale_x_continuous(breaks=seq(-20,20,by=1))+theme(panel.background=element_rect
(fill = "white", color = "black", size = 1.1)) +
geom_vline(xintercept = 0 , color="grey85", linetype="dashed", size=0.7) + geom_hline
(yintercept=0, color="grey60", size=0.7)
print(geomline.pattern_AG_nucl)
```

```
geomline.pattern_AT_nucl= ggplot(data.ALL.nucl_F[which(data.ALL.nucl_F$var =="A<->T"),],
aes(Position, bias, group=Nucleotide, shape=Nucleotide, color=Nucleotide)) +
geom_line(size=1.25) + geom_point(size=2)+
xlab("Posição")+ylab("Viés")+labs(group="Nucleótido")+
ggtitle("Variação A<->T")+ scale_color_brewer(palette="Set3") +scale_y_continuous
(limits=c(-0.1, 0.15), breaks = seq(-0.1, 0.15, by=0.01))+
scale_x_continuous(breaks= seq(-20,20, by=1))+ theme(panel.background=
element_rect
(fill = "white", color = "black", size = 1.1)) +
geom_vline(xintercept = 0 , color="grey85", linetype="dashed", size=0.7) +
geom_hline( yintercept=0, color="grey60", size=0.7)
print(geomline.pattern_AT_nucl)
geomline.pattern_CG_nucl= ggplot(data.ALL.nucl_F[which(data.ALL.nucl_F$var =="C<->G"),],
aes(Position, bias, group=Nucleotide, shape=Nucleotide, color=Nucleotide)) +
geom_line(size=1.25) + geom_point(size=2)+
xlab("Posição")+ylab("Viés")+labs(group="Nucleótido")+
ggtitle("Variação C<->G")+ scale_color_brewer(palette="Set3") +scale_y_continuous
limits=c(-0.1, 0.15), breaks = seq(-0.1, 0.15, by=0.01))+
scale_x_continuous(breaks= seq(-20, 20, by=1))+theme(panel.background=
lement_rect
(fill = "white", color = "black", size = 1.1)) +
geom_vline(xintercept = 0 , color="grey85", linetype="dashed", size=0.7) + geom_hline
( yintercept=0, color="grey60", size=0.7)
print(geomline.pattern_CG_nucl)
geomline.pattern_CT_nucl= ggplot(data.ALL.nucl_F[which(data.ALL.nucl_F$var =="C<->T"),],
aes(Position, bias, group=Nucleotide, shape=Nucleotide, color=Nucleotide)) +
geom_line(size=1.25) + geom_point(size=2)+
xlab("Posição")+ylab("Viés")+labs(group="Nucleótido")+
ggtitle("Variação C<->T")+ scale_color_brewer(palette="Set3") +scale_y_continuous
(limits=c(-0.1, 0.15), breaks = seq(-0.1, 0.15, by=0.01))+
scale_x_continuous(breaks=seq(-20,20,by=1))+theme(panel.background=
element_rect
(fill = "white", color = "black", size = 1.1)) +
geom_vline(xintercept = 0 , color="grey85", linetype="dashed", size=0.7) + geom_hline
( yintercept=0, color="grey60", size=0.7)
print(geomline.pattern_CT_nucl)
geomline.pattern_GT_nucl= ggplot(data.ALL.nucl_F[which(data.ALL.nucl_F$var =="G<->T"),],
aes(Position, bias, group=Nucleotide, shape=Nucleotide, color=Nucleotide)) +
geom_line(size=1.25) + geom_point(size=2)+
xlab("Posição")+ylab("Viés")+labs(group="Nucleótido")+
ggtitle("Variação G<->T")+ scale_color_brewer(palette="Set3") +scale_y_continuous
(limits=c(-0.1, 0.15), breaks = seq(-0.1, 0.15, by=0.01))+
scale_x_continuous(breaks=seq(-20,20,by=1))+theme(panel.background=
element_rect
(fill = "white", color = "black", size = 1.1)) +
geom_vline(xintercept = 0 , color="grey85", linetype="dashed", size=0.7) + geom_hline
( yintercept=0, color="grey60", size=0.7)
print(geomline.pattern_GT_nucl)

vector.var=c('A<->C','A<->G','A<->T','C<->G','C<->T','G<->T')
vector.nucl=c('A','C','G','T')
exp=rep(1/40, 40)
list.data.var.nucl= lapply(list.ALLvar.nucl, melt)
data.ALL.nuclGOF= as.data.frame(cbind(rbind(list.data.var.nucl[[1]],
list.data.var.nucl[[2]], list.data.var.nucl[[3]], list.data.var.nucl[[4]],
list.data.var.nucl[[5]], list.data.var.nucl[[6]]) , df.vector.var))
names(data.ALL.nuclGOF)=c('Position', 'Nucleotide','cont','var')

l_GOF=list()
for ( i in vector.var){l_GOF[[i]] = data.ALL.nuclGOF[which(data.ALL.nuclGOF$var == i
& data.ALL.nuclGOF$Position != 0 ),  ] }

listGOF_AC=list()
for (i in vector.nucl){t=chisq.test(l_GOF[[1]][which(l_GOF[[1]]$Nucleotide== i),
'cont'],p = exp)$statistic
listGOF_AC[[i]]=c(t, round(as.numeric(sqrt(t/length(exp))), 2))}
```

```
listGOF_AG=list()
for (i in vector.nucl){t=chisq.test(l_GOF[[2]][which(l_GOF[[2]]$Nucleotide == i),
'cont' ], p = exp)$statistic
listGOF_AG[[i]]=c(t, round(as.numeric(sqrt(t/length(exp))), 2))}
listGOF_AT=list()
for (i in vector.nucl){t=chisq.test(l_GOF[[3]][which(l_GOF[[3]]$Nucleotide == i),
'cont' ], p = exp)$statistic
listGOF_AT[[i]]=c(t, round(as.numeric(sqrt(t/length(exp))), 2))}
listGOF_CG=list()
for (i in vector.nucl){t=chisq.test(l_GOF[[4]][which(l_GOF[[4]]$Nucleotide == i),
'cont' ] , p = exp)$statistic
listGOF_CG[[i]]=c(t, round(as.numeric(sqrt(t/length(exp))), 2))}
listGOF_CT=list()
for (i in vector.nucl){t=chisq.test(l_GOF[[5]][which(l_GOF[[5]]$Nucleotide == i),
'cont' ] , p = exp)$statistic
listGOF_CT[[i]]=c(t, round(as.numeric(sqrt(t/length(exp))), 2))}
listGOF_GT=list()
for (i in vector.nucl){t=chisq.test(l_GOF[[6]][which(l_GOF[[6]]$Nucleotide == i),
'cont' ] , p = exp)$statistic
listGOF_GT[[i]]=c(t, round(as.numeric(sqrt(t/length(exp))), 2))}

list.pattern_2_AC= lapply(list.chr.AC, function(z){neighborhood.pattern(z,2,20)})
ALLpattern_2_AC=Reduce('+', list.pattern_2_AC)
list.pattern_2_AG= lapply(list.chr.AG, function(z){neighborhood.pattern(z,2,20)})
ALLpattern_2_AG=Reduce('+', list.pattern_2_AG)
list.pattern_2_AT= lapply(list.chr.AT, function(z){neighborhood.pattern(z,2,20)})
ALLpattern_2_AT=Reduce('+', list.pattern_2_AT)
list.pattern_2_CG= lapply(list.chr.CG, function(z){neighborhood.pattern(z,2,20)})
ALLpattern_2_CG=Reduce('+', list.pattern_2_CG)
list.pattern_2_CT= lapply(list.chr.CT, function(z){neighborhood.pattern(z,2,20)})
ALLpattern_2_CT=Reduce('+', list.pattern_2_CT)
list.pattern_2_GT= lapply(list.chr.GT, function(z){neighborhood.pattern(z,2,20)})
ALLpattern_2_GT=Reduce('+', list.pattern_2_GT)
list.ALLvar.dinucl=list(ALLpattern_2_AC, ALLpattern_2_AG, ALLpattern_2_AT,
ALLpattern_2_CG,ALLpattern_2_CT, ALLpattern_2_GT )
list.ALLvar.dinucl.fr=lapply(list.ALLvar.dinucl, function(z){prop.table(z, 1)})

vector.var2=c('A<->C','A<->G','A<->T','C<->G','C<->T','G<->T')
df.vector.var2=as.vector(rep(vector.var2,each= nrow(ALLpattern_2_AC)*4^2))
list.var.dinucl.fr= lapply(list.ALLvar.dinucl.fr, melt)
frq_exp_dinucl=c(0.097747210, 0.050339835, 0.069924165,0.077258419, 0.072535448 ,
0.052096925 ,0.009861512, 0.069961111, 0.059335043, 0.042660269, 0.052125953,
0.050454531 ,0.065651918, 0.059358019, 0.072664114, 0.098025528 )
vector.frq_exp_dinucl=as.vector(rep(frq_exp_dinucl, each= nrow(ALLpattern_2_AC)))
list.var.dinucl.fr_F=lapply(list.var.dinucl.fr, function(z)
{cbind(z,vector.frq_exp_dinucl)})
data.ALL.dinucl= as.data.frame(cbind(rbind(list.var.dinucl.fr_F[[1]],
list.var.dinucl.fr_F[[2]], list.var.dinucl.fr_F[[3]], list.var.dinucl.fr_F[[4]],
list.var.dinucl.fr_F[[5]],  list.var.dinucl.fr_F[[6]]) , df.vector.var2))
data.ALL.dinucl_F=cbind(data.ALL.dinucl, data.ALL.dinucl$value -
data.ALL.dinucl$vector.frq_exp_dinucl)
names(data.ALL.dinucl_F)=c('Position', 'Dinucleotide','obs','exp', 'var', 'bias')

colourCount2 = length(levels(data.ALL.dinucl_F$Dinucleotide))
getPalette2 = colorRampPalette(brewer.pal(10, "Set3"))
geomline.pattern_AC_dinucl= ggplot(data.ALL.dinucl_F[which
(data.ALL.dinucl_F$var =="A<->C"),],
aes(Position, bias, group=Dinucleotide, shape=Dinucleotide, color=Dinucleotide)) +
geom_line(size=0.825) + geom_point(size=1.3)+
xlab("Posição")+ylab("Viés")+labs(group="Dinucleótidos")+
ggtitle("Variação␣A<->C") + scale_fill_manual(values = getPalette2(colourCount2)) +
scale_y_continuous(limits=c(-0.05, 0.05), breaks = seq(-0.05, 0.05, by=0.01))+
scale_x_continuous(breaks = seq(-20, 20, by=1) ) +
theme(panel.background = element_rect(fill = "white", color = "black", size = 1.1)) +
geom_vline(xintercept = 0 , color="grey85", linetype="dashed", size=0.7) +
 geom_hline( yintercept=0, color="grey60", size=0.7)
```

```
print(geomline.pattern_AC_dinucl)

geomline.pattern_AG_dinucl=
ggplot(data.ALL.dinucl_F[which(data.ALL.dinucl_F$var =="A<->G"),],
aes(Position, bias, group=Dinucleotide, shape=Dinucleotide, color=Dinucleotide)) +
geom_line(size=0.825) + geom_point(size=1.3)+
xlab("Posição")+ylab("Viés")+labs(group="Dinucleótidos")+
ggtitle("Variação␣A<->G") + scale_fill_manual(values = getPalette2(colourCount2)) +
 scale_y_continuous(limits=c(-0.05, 0.05), breaks = seq(-0.05, 0.05, by=0.01))+
scale_x_continuous(breaks = seq(-20, 20, by=1) ) + theme(panel.background =
element_rect(fill = "white", color = "black", size = 1.1)) +
geom_vline(xintercept = 0 , color="grey85", linetype="dashed", size=0.7) +
 geom_hline( yintercept=0, color="grey60", size=0.7)
print(geomline.pattern_AG_dinucl)

geomline.pattern_AT_dinucl=
ggplot(data.ALL.dinucl_F[which(data.ALL.dinucl_F$var =="A<->T"),],
aes(Position, bias, group=Dinucleotide, shape=Dinucleotide, color=Dinucleotide)) +
geom_line(size=0.825) + geom_point(size=1.3)+
xlab("Posição")+ylab("Viés")+labs(group="Dinucleótidos")+
ggtitle("Variação␣A<->T") + scale_fill_manual(values = getPalette2(colourCount2)) +
 scale_y_continuous(limits=c(-0.05, 0.05), breaks = seq(-0.05, 0.05, by=0.01))+
scale_x_continuous(breaks = seq(-20, 20, by=1) ) + theme(panel.background =
element_rect(fill = "white", color = "black", size = 1.1)) +
geom_vline(xintercept = 0 , color="grey85", linetype="dashed", size=0.7) +
 geom_hline( yintercept=0, color="grey60", size=0.7)
print(geomline.pattern_AT_dinucl)

geomline.pattern_CG_dinucl=
ggplot(data.ALL.dinucl_F[which(data.ALL.dinucl_F$var =="C<->G"),],
aes(Position, bias, group=Dinucleotide, shape=Dinucleotide, color=Dinucleotide)) +
geom_line(size=0.825) + geom_point(size=1.3)+
xlab("Posição")+ylab("Viés")+labs(group="Dinucleótidos")+
ggtitle("Variação␣C<->G") + scale_fill_manual(values = getPalette2(colourCount2)) +
 scale_y_continuous(limits=c(-0.05, 0.05), breaks = seq(-0.05, 0.05, by=0.01))+
scale_x_continuous(breaks = seq(-20, 20, by=1) ) + theme(panel.background =
element_rect(fill = "white", color = "black", size = 1.1)) +
geom_vline(xintercept = 0 , color="grey85", linetype="dashed", size=0.7) +
geom_hline( yintercept=0, color="grey60", size=0.7)
print(geomline.pattern_CG_dinucl)

geomline.pattern_CT_dinucl=
ggplot(data.ALL.dinucl_F[which(data.ALL.dinucl_F$var =="C<->T"),],
aes(Position, bias, group=Dinucleotide, shape=Dinucleotide, color=Dinucleotide)) +
geom_line(size=0.825) + geom_point(size=1.3)+
xlab("Posição")+ylab("Viés")+labs(group="Dinucleótidos")+
ggtitle("Variação␣C<->T") + scale_fill_manual(values = getPalette2(colourCount2)) +
 scale_y_continuous(limits=c(-0.05, 0.05), breaks = seq(-0.05, 0.05, by=0.01))+
scale_x_continuous(breaks = seq(-20, 20, by=1) ) + theme(panel.background =
element_rect(fill = "white", color = "black", size = 1.1)) +
geom_vline(xintercept = 0 , color="grey85", linetype="dashed", size=0.7) +
geom_hline( yintercept=0, color="grey60", size=0.7)
print(geomline.pattern_CT_dinucl)

geomline.pattern_GT_dinucl=
ggplot(data.ALL.dinucl_F[which(data.ALL.dinucl_F$var =="G<->T"),],
 aes(Position, bias, group=Dinucleotide, shape=Dinucleotide, color=Dinucleotide)) +
geom_line(size=0.825) + geom_point(size=1.3)+
xlab("Posição")+ylab("Viés")+labs(group="Dinucleótidos")+
ggtitle("Variação␣G<->T") + scale_fill_manual(values = getPalette2(colourCount2)) +
 scale_y_continuous(limits=c(-0.05, 0.05), breaks = seq(-0.05, 0.05, by=0.01))+
scale_x_continuous(breaks = seq(-20, 20, by=1) ) + theme(panel.background =
element_rect(fill = "white", color = "black", size = 1.1)) +
geom_vline(xintercept = 0 , color="grey85", linetype="dashed", size=0.7) +
geom_hline( yintercept=0, color="grey60", size=0.7)
print(geomline.pattern_GT_dinucl)
```

```
vector.var=c('A<->C','A<->G','A<->T','C<->G','C<->T','G<->T')
vector.dinucl=c('AA','AC','AG','AT', 'CA', 'CC', 'CG', 'CT', 'GA',
'GC', 'GG', 'GT', 'TA', 'TC', 'TG', 'TT')
df.vector.var2=as.vector(rep(vector.var2,each= nrow(ALLpattern_2_AC)*4^2))
list.data.var.dinucl= lapply(list.ALLvar.dinucl, melt)
data.ALL.dinuclGOF= as.data.frame(cbind(rbind(list.data.var.dinucl[[1]],
list.data.var.dinucl[[2]], list.data.var.dinucl[[3]], list.data.var.dinucl[[4]],
list.data.var.dinucl[[5]], list.data.var.dinucl[[6]]) , df.vector.var2))
names(data.ALL.dinuclGOF)=c('Position', 'Dinucleotide','cont','var')

l2_GOF=list()
for ( i in vector.var) {l2_GOF[[i]] =
data.ALL.dinuclGOF[which(data.ALL.dinuclGOF$var == i &
data.ALL.dinuclGOF$Position != 0 ),  ] }
list2PhiAC=list()
for (i in vector.dinucl){list2PhiAC[[i]]=round(as.numeric(sqrt(chisq.test(l2_GOF[[1]]
[which(l2_GOF[[1]]$Dinucleotide == i),'cont'],p = exp)$statistic/length(exp))),2)}
list2PhiAG=list()
for (i in vector.dinucl){
list2PhiAG[[i]]=round(as.numeric(sqrt(chisq.test(l2_GOF[[2]]
[which(l2_GOF[[2]]$Dinucleotide == i ), 'cont' ] , p = exp)$statistic/length(exp))),2)}
list2PhiAT=list()
for (i in vector.dinucl){
list2PhiAT[[i]]=round(as.numeric(sqrt(chisq.test(l2_GOF[[3]]
[which(l2_GOF[[3]]$Dinucleotide == i ), 'cont' ] , p = exp)$statistic/length(exp))),2)}
list2PhiCG=list()
for (i in vector.dinucl){
list2PhiCG[[i]]=round(as.numeric(sqrt(chisq.test(l2_GOF[[4]]
[which(l2_GOF[[4]]$Dinucleotide == i ), 'cont' ] , p = exp)$statistic/length(exp))),2)}
list2PhiCT=list()
for (i in vector.dinucl){
list2PhiCT[[i]]=round(as.numeric(sqrt(chisq.test(l2_GOF[[5]]
[which(l2_GOF[[5]]$Dinucleotide == i ), 'cont' ] , p = exp)$statistic/length(exp))),2)}
list2PhiGT=list()
for (i in vector.dinucl){
list2PhiGT[[i]]=round(as.numeric(sqrt(chisq.test(l2_GOF[[6]]
[which(l2_GOF[[6]]$Dinucleotide == i ), 'cont' ] , p = exp)$statistic/length(exp))),2)}
Matrix2Phi=cbind(as.matrix(list2PhiAC), as.matrix(list2PhiAG), as.matrix(list2PhiAT),
as.matrix(list2PhiCG),as.matrix(list2PhiCT),as.matrix(list2PhiGT))

list.pattern_3_AC= lapply(list.chr.AC, function(z) {neighborhood.pattern(z, 3, 10) })
ALLpattern_3_AC=Reduce('+', list.pattern_3_AC)
list.pattern_3_AG= lapply(list.chr.AG, function(z) {neighborhood.pattern(z, 3, 10) })
ALLpattern_3_AG=Reduce('+', list.pattern_3_AG)
list.pattern_3_AT= lapply(list.chr.AT, function(z) {neighborhood.pattern(z, 3, 10) })
ALLpattern_3_AT=Reduce('+', list.pattern_3_AT)
list.pattern_3_CG= lapply(list.chr.CG, function(z) {neighborhood.pattern(z, 3, 10) })
ALLpattern_3_CG=Reduce('+', list.pattern_3_CG)
list.pattern_3_CT= lapply(list.chr.CT, function(z) {neighborhood.pattern(z, 3, 10) })
ALLpattern_3_CT=Reduce('+', list.pattern_3_CT)
list.pattern_3_GT= lapply(list.chr.GT, function(z) {neighborhood.pattern(z, 3, 10) })
ALLpattern_3_GT=Reduce('+', list.pattern_3_GT)
list.ALLvar.trinucl=list(ALLpattern_3_AC, ALLpattern_3_AG, ALLpattern_3_AT,
ALLpattern_3_CG, ALLpattern_3_CT, ALLpattern_3_GT )
list.ALLvar.trinucl.fr=lapply(list.ALLvar.trinucl, function(z){prop.table(z, 1)})
vector.var3=c('A<->C','A<->G','A<->T','C<->G','C<->T','G<->T')
df.vector.var3=as.vector(rep(vector.var3,each= nrow(ALLpattern_3_AC)*4^3))
list.var.trinucl.fr= lapply(list.ALLvar.trinucl.fr, melt)
frq_exp_trinucl=c(0.038356036, 0.014548233, 0.019932776, 0.024910174, 0.020131352,
0.011622403,0.002509919, 0.016076165, 0.022099860, 0.013977795, 0.017749306,
0.016097208,0.020605451, 0.013343207, 0.018364916, 0.024944852, 0.018905492,
0.015011558,0.020254100, 0.018364303, 0.018425634, 0.013141440, 0.002761142,
0.017768711, 0.002205036, 0.002379226, 0.002761146, 0.002516104, 0.012885941,
0.016833965,0.020269153, 0.019972055, 0.019703463, 0.009439399, 0.016834132,
0.013358056,0.014395122, 0.011901217, 0.002381242, 0.013982693, 0.015440744,
```

```
0.011895615,0.013153529, 0.011636065, 0.011346571, 0.009452444, 0.015048771,
0.014606748,0.020782223, 0.011340648, 0.012903162, 0.020625891, 0.019583345,
0.015431866,0.002209209, 0.022133545, 0.019589347, 0.014407638, 0.018461976,
0.020205158,0.020813959, 0.019728407, 0.018981279, 0.038501884)

vector.frq_exp_trinucl=as.vector(rep(frq_exp_trinucl, each= nrow(ALLpattern_3_AC)))
list.var.trinucl.fr_F=lapply(list.var.trinucl.fr, function(z)
{cbind(z,vector.frq_exp_trinucl)})
data.ALL.trinucl= as.data.frame(cbind(rbind(list.var.trinucl.fr_F[[1]],
list.var.trinucl.fr_F[[2]], list.var.trinucl.fr_F[[3]], list.var.trinucl.fr_F[[4]],
list.var.trinucl.fr_F[[5]], list.var.trinucl.fr_F[[6]]) , df.vector.var3))
data.ALL.trinucl_F=cbind(data.ALL.trinucl, data.ALL.trinucl$value -
data.ALL.trinucl$vector.frq_exp_trinucl)
names(data.ALL.trinucl_F)=c('Position', 'Trinucleotide','obs','exp', 'var', 'bias')

colourCount3 = length(levels(data.ALL.trinucl_F$Trinucleotide))
getPalette3 = colorRampPalette(brewer.pal(10, "Set3"))
geomline.pattern_AC_trinucl=
ggplot(data.ALL.trinucl_F[which(data.ALL.trinucl_F$var =="A<->C"),],
aes(Position, bias, group=Trinucleotide, shape=Trinucleotide, color=Trinucleotide)) +
geom_line(size=0.825) + geom_point(size=1.3)+
xlab("Posição")+ylab("Viés")+labs(group="Trinucleótidos")+
ggtitle("Variação␣A<->C") + scale_fill_manual(values = getPalette3(colourCount3)) +
 scale_y_continuous(limits=c(-0.02, 0.025), breaks = seq(-0.02, 0.025, by=0.01))+
scale_x_continuous(breaks = seq(-10, 10, by=1) ) + theme(panel.background =
element_rect(fill = "white", color = "black", size = 1.1)) +
geom_vline(xintercept = 0 , color="grey85", linetype="dashed", size=0.7) +
geom_hline( yintercept=0, color="grey60", size=0.7)
print(geomline.pattern_AC_trinucl)

geomline.pattern_AG_trinucl=
ggplot(data.ALL.trinucl_F[which(data.ALL.trinucl_F$var =="A<->G"),],
aes(Position, bias, group=Trinucleotide, shape=Trinucleotide, color=Trinucleotide)) +
geom_line(size=0.825) + geom_point(size=1.3)+
xlab("Posição")+ylab("Viés")+labs(group="Trinucleótidos")+
ggtitle("Variação␣A<->G") + scale_fill_manual(values = getPalette3(colourCount3)) +
 scale_y_continuous(limits=c(-0.02, 0.025), breaks = seq(-0.02, 0.025, by=0.01))+
scale_x_continuous(breaks = seq(-10, 10, by=1) ) + theme(panel.background =
 element_rect(fill = "white", color = "black", size = 1.1)) +
geom_vline(xintercept = 0 , color="grey85", linetype="dashed", size=0.7) +
geom_hline( yintercept=0, color="grey60", size=0.7)
print(geomline.pattern_AG_trinucl)

geomline.pattern_AT_trinucl=
ggplot(data.ALL.trinucl_F[which(data.ALL.trinucl_F$var =="A<->T"),],
aes(Position, bias, group=Trinucleotide, shape=Trinucleotide, color=Trinucleotide)) +
geom_line(size=0.825) + geom_point(size=1.3)+
xlab("Posição")+ylab("Viés")+labs(group="Trinucleótidos")+
ggtitle("Variação␣A<->T") + scale_fill_manual(values = getPalette3(colourCount3)) +
scale_y_continuous(limits=c(-0.02, 0.025), breaks = seq(-0.02, 0.025, by=0.01))+
scale_x_continuous(breaks = seq(-10, 10, by=1) ) + theme(panel.background =
element_rect(fill = "white", color = "black", size = 1.1)) +
geom_vline(xintercept = 0 , color="grey85", linetype="dashed", size=0.7) +
geom_hline( yintercept=0, color="grey60", size=0.7)
print(geomline.pattern_AT_trinucl)

geomline.pattern_CG_trinucl=
ggplot(data.ALL.trinucl_F[which(data.ALL.trinucl_F$var =="C<->G"),],
aes(Position, bias, group=Trinucleotide, shape=Trinucleotide, color=Trinucleotide)) +
geom_line(size=0.825) + geom_point(size=1.3)+
xlab("Posição")+ylab("Viés")+labs(group="Trinucleótidos")+
ggtitle("Variação␣C<->G") + scale_fill_manual(values = getPalette3(colourCount3)) +
scale_y_continuous(limits=c(-0.02, 0.025), breaks = seq(-0.02, 0.025, by=0.01))+
scale_x_continuous(breaks = seq(-10, 10, by=1) ) + theme(panel.background =
element_rect(fill = "white", color = "black", size = 1.1)) +
geom_vline(xintercept = 0 , color="grey85", linetype="dashed", size=0.7) +
```

```
geom_hline( yintercept=0, color="grey60", size=0.7)
print(geomline.pattern_CG_trinucl)

geomline.pattern_CT_trinucl=
ggplot(data.ALL.trinucl_F[which(data.ALL.trinucl_F$var =="C<->T"),],
aes(Position, bias, group=Trinucleotide, shape=Trinucleotide, color=Trinucleotide)) +
geom_line(size=0.825) + geom_point(size=1.3)+
xlab("Posição")+ylab("Viés")+labs(group="Trinucleótidos")+
ggtitle("Variação␣C<->T") + scale_fill_manual(values = getPalette3(colourCount3)) +
scale_y_continuous(limits=c(-0.02, 0.025), breaks = seq(-0.02, 0.025, by=0.01))+
scale_x_continuous(breaks = seq(-10, 10, by=1) ) + theme(panel.background =
element_rect(fill = "white", color = "black", size = 1.1)) +
geom_vline(xintercept = 0 , color="grey85", linetype="dashed", size=0.7) +
geom_hline( yintercept=0, color="grey60", size=0.7)
print(geomline.pattern_CT_trinucl)

geomline.pattern_GT_trinucl=
ggplot(data.ALL.trinucl_F[which(data.ALL.trinucl_F$var =="G<->T"),],
aes(Position, bias, group=Trinucleotide, shape=Trinucleotide, color=Trinucleotide)) +
geom_line(size=0.825) + geom_point(size=1.3)+
xlab("Posição")+ylab("Viés")+labs(group="Trinucleótidos")+
ggtitle("Variation␣G<->T") + scale_fill_manual(values = getPalette3(colourCount3)) +
scale_y_continuous(limits=c(-0.02, 0.025), breaks = seq(-0.02, 0.025, by=0.01))+
scale_x_continuous(breaks = seq(-10, 10, by=1) ) + theme(panel.background =
element_rect(fill = "white", color = "black", size = 1.1)) +
geom_vline(xintercept = 0 , color="grey85", linetype="dashed", size=0.7) +
geom_hline( yintercept=0, color="grey60", size=0.7)
print(geomline.pattern_GT_trinucl)

list.ALLtrans.nucl=list( ALLpattern_1_AG, ALLpattern_1_CT)
list.ALLtransv.nucl=list(ALLpattern_1_AC, ALLpattern_1_AT,
ALLpattern_1_CG,ALLpattern_1_GT )
ALLpattern_1_trans=Reduce('+', list.ALLtrans.nucl)
ALLpattern_1_transv=Reduce('+', list.ALLtransv.nucl)
list.ALLpattern_1=list(ALLpattern_1_trans, ALLpattern_1_transv)
list.ALLpattern_1.fr=lapply(list.ALLpattern_1, function(z){prop.table(z, 1)})

TabAT1TRANS=cbind(list.ALLpattern_1[[1]][22:41,1],
apply(t(list.ALLpattern_1[[1]]),1,rev)[22:41,4], c(-1:-20))
colnames(TabAT1TRANS)=c("A", "T", "Pos")
prop.table(TabAT1TRANS[,1:2], 1)
TabCG1TRANS=cbind(list.ALLpattern_1[[1]][22:41,2],
apply(t(list.ALLpattern_1[[1]]),1,rev)[22:41,3], c(-1:-20))
colnames(TabCG1TRANS)=c("C", "G", "Pos")
TabAT1TRANSV=cbind(list.ALLpattern_1[[2]][22:41,1],
apply(t(list.ALLpattern_1[[2]]),1,rev)[22:41,4], c(-1:-20))
colnames(TabAT1TRANSV)=c("A", "T", "Pos")
TabCG1TRANSV=cbind(list.ALLpattern_1[[2]][22:41,2],
apply(t(list.ALLpattern_1[[2]]),1,rev)[22:41,3], c(-1:-20))
colnames(Tab~CG1TRANSV)=c("C", "G", "Pos")
summary(assocstats(TabAT1TRANS[,1:2]))
summary(assocstats(TabCG1TRANS[,1:2]))
summary(assocstats(TabAT1TRANSV[,1:2]))
summary(assocstats(TabCG1TRANSV[,1:2]))
chisq.test(TabCG1TRANS[,1:2])
Phi1=Phi(TabAT1TRANS[,1:2])
Phi2=Phi(TabCG1TRANS[,1:2])
Phi3=Phi(TabAT1TRANSV[,1:2])
Phi4=Phi(TabCG1TRANSV[,1:2])

Resi_ajAT1TRANS=chisq.test(TabAT1TRANS[,1:2])$stdres
Resi_ajCG1TRANS=chisq.test(TabCG1TRANS[,1:2])$stdres
Resi_ajAT1TRANSV=chisq.test(TabAT1TRANSV[,1:2])$stdres
Resi_ajCG1TRANSV=chisq.test(TabCG1TRANSV[,1:2])$stdres
heatmap.2(Resi_ajAT1TRANS,col=brewer.pal(11,"RdYlGn"),trace="none")
heatmap.2(Resi_ajCG1TRANS,col=brewer.pal(11,"RdYlGn"),trace="none")
```

```
heatmap.2(Resi_ajAT1TRANSV,col=brewer.pal(11,"RdYlGn"),trace="none")
heatmap.2(Resi_ajCG1TRANSV,col=brewer.pal(11,"RdYlGn"),trace="none")

list.ALLtrans.dinucl=list( ALLpattern_2_AG, ALLpattern_2_CT)
list.ALLtransv.dinucl=list(ALLpattern_2_AC,ALLpattern_2_AT,ALLpattern_2_CG,
ALLpattern_2_GT)
ALLpattern_2_trans=Reduce('+', list.ALLtrans.dinucl)
ALLpattern_2_transv=Reduce('+', list.ALLtransv.dinucl)
list.ALLpattern_2=list(ALLpattern_2_trans, ALLpattern_2_transv)
list.ALLpattern_2.fr=lapply(list.ALLpattern_2, function(z){prop.table(z, 1)})
vector.group=c('Transition','Transversion')
df.vector.group=as.vector(rep(vector.group,each= nrow(ALLpattern_2_trans)*4^2))
list.data.ALLpattern_2.fr= lapply(list.ALLpattern_2.fr, melt)

vector.frq_exp_dinucl=as.vector(rep(frq_exp_dinucl, each= nrow(ALLpattern_2_trans)))
list.data.ALLpattern_2.fr_F=lapply(list.data.ALLpattern_2.fr,
function(z){cbind(z,vector.frq_exp_dinucl)})

data.ALLpattern_2= as.data.frame(cbind(rbind(list.data.ALLpattern_2.fr_F[[1]],
list.data.ALLpattern_2.fr_F[[2]]) , df.vector.group))

data.ALLpattern_2_F=cbind(data.ALLpattern_2, data.ALLpattern_2$value -
data.ALLpattern_2$vector.frq_exp_dinucl) names(data.ALLpattern_2_F)=c('Position',
'Dinucleotide','obs','exp', 'group', 'bias')
head(list.ALLpattern_2[[1]])

TabAATTTRANS=cbind(list.ALLpattern_2[[1]][22:41,1],
apply(t(list.ALLpattern_2[[1]]),1,rev)[22:41,16], c(-1:-20))
colnames(TabAATTTRANS)=c("AA", "TT", "Pos")
assocstats(TabAATTTRANS[,1:2])
Resi_ajTRANS=chisq.test(TabAATTTRANS[,1:2])$stdres
heatmap.2(Resi_ajTRANS,col=brewer.pal(11,"RdYlGn"),trace="none")
TabACGTTRANS=cbind(list.ALLpattern_2[[1]][22:41,2],
apply(t(list.ALLpattern_2[[1]]),1,rev)[22:41,12], c(-1:-20))
colnames(TabAATTTRANS)=c("AC", "GT", "Pos")
assocstats(TabACGTTRANS[,1:2])

TabAGCTTRANS=cbind(list.ALLpattern_2[[1]][22:41,3],
apply(t(list.ALLpattern_2[[1]]),1,rev)[22:41,8], c(-1:-20))
colnames(TabAATTTRANS)=c("AG", "CT", "Pos")
assocstats(TabAGCTTRANS[,1:2])
Resi_ajTRANS=chisq.test(TabAGCTTRANS[,1:2])$stdres
heatmap.2(Resi_ajTRANS,col=brewer.pal(11,"RdYlGn"),trace="none")

TabATATTRANS=cbind(list.ALLpattern_2[[1]][22:41,4],
apply(t(list.ALLpattern_2[[1]]),1,rev)[22:41,4], c(-1:-20))
colnames(TabATATTRANS)=c("AT", "AT", "Pos")
assocstats(TabATATTRANS[,1:2])

TabCATGTRANS=cbind(list.ALLpattern_2[[1]][22:41,5],
apply(t(list.ALLpattern_2[[1]]),1,rev)[22:41,15], c(-1:-20))
colnames(TabCATGTRANS)=c("CA", "TG", "Pos")
assocstats(TabCATGTRANS[,1:2])
Resi_ajTRANS=chisq.test(TabCATGTRANS[,1:2])$stdres
heatmap.2(Resi_ajTRANS,col=brewer.pal(11,"RdYlGn"),trace="none")

TabCCGGTRANS=cbind(list.ALLpattern_2[[1]][22:41,6],
apply(t(list.ALLpattern_2[[1]]),1,rev)[22:41,11], c(-1:-20))
colnames(TabCCGGTRANS)=c("CC", "GG", "Pos")
assocstats(TabCCGGTRANS[,1:2])
Resi_ajTRANS=chisq.test(TabCCGGTRANS[,1:2])$stdres
heatmap.2(Resi_ajTRANS,col=brewer.pal(11,"RdYlGn"),trace="none")

TabCGCGTRANS=cbind(list.ALLpattern_2[[1]][22:41,7],
apply(t(list.ALLpattern_2[[1]]),1,rev)[22:41,7], c(-1:-20))
colnames(TabCGCGTRANS)=c("CG", "CG", "Pos")
```

```
assocstats(TabCGCGTRANS[,1:2])

TabGATCTRANS=cbind(list.ALLpattern_2[[1]][22:41,9],
apply(t(list.ALLpattern_2[[1]]),1,rev)[22:41,14], c(-1:-20))
colnames(TabGATCTRANS)=c("GA", "TC", "Pos")
assocstats(TabGATCTRANS[,1:2])

TabGCGCTRANS=cbind(list.ALLpattern_2[[1]][22:41,10],
apply(t(list.ALLpattern_2[[1]]),1,rev)[22:41,10], c(-1:-20))
colnames(TabGATCTRANS)=c("GC", "TC", "Pos")
assocstats(TabGCGCTRANS[,1:2])
Resi_ajTRANS=chisq.test(TabGCGCTRANS[,1:2])$stdres
heatmap.2(Resi_ajTRANS,col=brewer.pal(11,"RdYlGn"),trace="none")

TabTATATRANS=cbind(list.ALLpattern_2[[1]][22:41,13],
apply(t(list.ALLpattern_2[[1]]),1,rev)[22:41,13], c(-1:-20))
assocstats(TabTATATRANS[,1:2])

Resi_ajAATTTRANS=chisq.test(TabAATTTRANS[,1:2])$stdres
heatmap.2(Resi_ajAATTTRANS,col=brewer.pal(11,"RdYlGn"),trace="none")

Resi_ajCCGGTRANS=chisq.test(TabCCGGTRANS[,1:2])$stdres
heatmap.2(Resi_ajCCGGTRANS,col=brewer.pal(11,"RdYlGn"),trace="none")

TabAATTTRANSV=cbind(list.ALLpattern_2[[2]][22:41,1],
apply(t(list.ALLpattern_2[[2]]),1,rev)[22:41,16], c(-1:-20))
assocstats(TabAATTTRANSV[,1:2])
TabACGTTRANSV=cbind(list.ALLpattern_2[[2]][22:41,2],
apply(t(list.ALLpattern_2[[2]]),1,rev)[22:41,12], c(-1:-20))
assocstats(TabACGTTRANSV[,1:2])
Resi_ajACGTTRANSV=chisq.test(TabACGTTRANSV[,1:2])$stdres
heatmap.2(Resi_ajACGTTRANSV,col=brewer.pal(11,"RdYlGn"),trace="none")

TabAGCTTRANSV=cbind(list.ALLpattern_2[[2]][22:41,3],
apply(t(list.ALLpattern_2[[2]]),1,rev)[22:41,8], c(-1:-20))
assocstats(TabAGCTTRANSV[,1:2])

TabATATTRANSV=cbind(list.ALLpattern_2[[2]][22:41,4],
apply(t(list.ALLpattern_2[[2]]),1,rev)[22:41,4], c(-1:-20))
assocstats(TabATATTRANSV[,1:2])
Resi_ajATATTRANSV=chisq.test(TabATATTRANSV[,1:2])$stdres
heatmap.2(Resi_ajATATTRANSV,col=brewer.pal(11,"RdYlGn"),trace="none")

TabCATGTRANSV=cbind(list.ALLpattern_2[[2]][22:41,5],
apply(t(list.ALLpattern_2[[2]]),1,rev)[22:41,15], c(-1:-20))
assocstats(TabCATGTRANSV[,1:2])

TabCCGGTRANSV=cbind(list.ALLpattern_2[[2]][22:41,6],
apply(t(list.ALLpattern_2[[2]]),1,rev)[22:41,11], c(-1:-20))
assocstats(TabCCGGTRANSV[,1:2])

TabCGCGTRANSV=cbind(list.ALLpattern_2[[2]][22:41,7],
apply(t(list.ALLpattern_2[[2]]),1,rev)[22:41,7], c(-1:-20))
assocstats(TabCGCGTRANSV[,1:2])

TabGATCTRANSV=cbind(list.ALLpattern_2[[2]][22:41,9],
apply(t(list.ALLpattern_2[[2]]),1,rev)[22:41,14], c(-1:-20))
assocstats(TabGATCTRANSV[,1:2])

TabGCGCTRANSV=cbind(list.ALLpattern_2[[2]][22:41,10],
apply(t(list.ALLpattern_2[[2]]),1,rev)[22:41,10], c(-1:-20))
assocstats(TabGCGCTRANSV[,1:2])

TabTATATRANSV=cbind(list.ALLpattern_2[[2]][22:41,13],
apply(t(list.ALLpattern_2[[2]]),1,rev)[22:41,13], c(-1:-20))
assocstats(TabTATATRANSV[,1:2])
```