Pós-Graduação em Ciência da Computação

“Identificação de novos genes e SNPs relacionados ao Mal de Parkinson e doenças relacionadas através de GWAS”

Por

Rebecca Cristina Linhares de Carvalho

Dissertação de Mestrado

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE/2015

Universidade Federal de Pernambuco
Centro de Informática
Pós-graduação em Ciência da Computação

Rebecca Cristina Linhares de Carvalho

IDENTIFICAÇÃO DE NOVOS GENES E SNPS RELACIONADOS AO MAL DE PARKINSON E DOENÇAS RELACIONADAS ATRAVÉS DE GWAS

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientador: *Katia Silva Guimarães*

RECIFE

2015

Catalogação na fonte
Bibliotecária Jane Souto Maior, CRB4-571

C331i Carvalho, Rebecca Cristina Linhares de
Identificação de novos genes e SNPs relacionados ao mal de Parkinson e doenças relacionadas através de GWAS / Rebecca Cristina Linhares de Carvalho. – Recife: O Autor, 2015.
73 f.: il. fig., tab.

Orientador: Katia Silva Guimarães.
Dissertação (Mestrado) – Pernambuco. CIn, Ciência da computação, 2015.
Inclui referências e apêndice.

1. Ciência da Computação. 2. Biologia computacional. 3. Bioinformática. 4. Programas estatísticos. I. Guimarães, Katia Silva (orientadora). II. Título.

004 CDD (23. ed.) UFPE- MEI 2015-153

Dissertação de Mestrado apresentada por **Rebecca Cristina Linhares de Carvalho** à Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, sob o título **"Identificação de novos genes e SNPs relacionados ao Mal de Parkinson e doenças relacionadas, através de GWAS",** orientada pela **Profa. Katia Silva Guimarães** aprovada pela Banca Examinadora formada pelos professores:

_____

Profa. Renata Maria Cardoso Rodrigues de Souza
Centro de Informática / UFPE

_____

Profa. Jeane Cecília Bezerra de Melo
Departamento de Estatística e Informática / UFRPE

_____

Profa. Katia Silva Guimarães
Centro de Informática / UFPE

Visto e permitida a impressão.
Recife, 25 de fevereiro de 2015.

_____

Profa. Edna Natividade da Silva Barros
Coordenador da Pós-Graduação em Ciência da Computação do
Centro de Informática da Universidade Federal de Pernambuco.

Eu dedico essa dissertação a toda a minha família, a Guelryson, meus amigos e a professora Katia que me deu o apoio necessário para chegar até aqui.

Agradecimentos

"Deem graças ao Senhor, clamem pelo seu nome,
divulguem entre as nações o que ele tem feito". 1 Crônicas 16:8

À minha família pelo apoio e cuidado.

À tia Ana Maria Aquino, pela amizade e carinho. Eternas saudades.

Ao meu namorado e amigo Guelryson Eninng, pelo carinho, apoio, incentivo, companheirismo e paciência em muitas "fases do mestrado".

Aos meus irmãos, amigos e colegas por todo acolhimento, incentivo, ajuda, delicadeza, sensibilidade e alegria.

À Renata, por estar pronta a me ouvir, sempre gentil e presente.

Ao IBCIn pelas novas e preciosas amizades, pelas palavras que alimentam o espírito e a todo apoio dado.

À Profa. PhD. Katia Guimarães Silva, pela confiança, disponibilidade e objetividade. Pela oportunidade de trabalhar ao seu lado, pela condução deste meu trabalho e por ser a maior incentivadora na superação dos meus limites.

À Fundação de Amparo a Ciência e Tecnologia (FACEPE), pelo apoio financeiro dado a este trabalho.

Ao grupo de estudos Parkinson's Progression Markers Initiative (PPMI) pelos dados genéticos cedidos, que foram de suma importância para o nosso trabalho.

A coisa mais indispensável a um homem é reconhecer o uso que deve fazer do seu próprio conhecimento.

—PLATÃO

Resumo

Parkinsonismo é uma síndrome neurológica em que os neurônios que normalmente produzem o hormônio chamado dopamina se deterioram, causando a perda de controle progressivo do movimento (ex. bradicinesia, a rigidez muscular, o temor de repouso e os reflexos posturais prejudicados). Um mal com sintomas semelhantes ao parkinsonismo é a síndrome Scans Without Evidence of Dopaminergic Deficits (SWEDDs), na qual os pacientes não apresentam evidências de déficit de dopamina. As causas de parkinsonismo primário como a doença de Parkinson (DP), bem como SWEDDs, não são completamente conhecidos. Os estudos de associações no genoma completo (*Genome-wide association studies* - GWAS) têm proporcionado ganhos tangíveis para a compreensão da arquitetura genética de doenças complexas, trazendo contribuições consistentes e importantes para DP. GWASs foram realizados no passado com os dados de DP com resultados que influenciaram fortemente desenvolvimentos posteriores.

Nesse trabalho, nós desenvolvemos um estudo sobre fatores genéticos que possam contribuir para o entendimento da ocorrência da DP e SWEDDs. Para isso, nós usamos o conjunto de ferramentas de análise de associação envolvendo estudo de caso-controle do método PLINK para executar GWASs, a fim de identificar SNPs que estão associados à DP e SWEDDs. Para fixar o nível de significância dos nossos resultados, nós optamos por usar somente dados reais fornecidos por Parkinson's Progression Markers Initiative (PPMI). O PPMI é um consórcio internacional projetado para identificar biomarcadores de progressão da DP, tanto para melhorar a compreensão da etiologia da doença, como para fornecer ferramentas cruciais para aumentar a probabilidade de sucesso na elaboração de novos ensaios terapêuticos para DP.

Na análise de associação, feita com dados genótipos de três grupos de indivíduos (indivíduos saudáveis, indivíduos com DP e indivíduos com SWEDDs), recuperamos SNPs que mostram forte ligação com PD e SWEDDs, alguns deles já associados na literatura científica com a DP ou a outras doenças degenerativas. Mas, também encontramos cerca de 60 SNPs que não estão relatados na literatura, que mostram evidências de serem fortemente relacionadas com a propensão para DP ou SWEDDs. Estes resultados apresentam alvos promissores para futuros

estudos genômicos e podem contribuir para o entendimento da ocorrência da DP e SWEDDs. Curiosamente, embora SWEDDs seja uma doença clinicamente ligada a DP por uma série de sintomas comuns, os SNPs recuperados com as melhores classificações a partir do conjunto de dados DP não faziam parte do conjunto de dados SWEDDs, e vice-versa, o que sugere que esses dois conjuntos de marcadores poderiam ser mais cuidadosamente explorados nos estudos genômicos como SNPs comuns de interesse para as duas doenças.

Palavras-chave:

SNP. GWAS. Doença de Parkinson. SWEDDs. Parkinsonismo.

Abstract

Parkinsonism is a neurological disorder neurons that normally produce the hormone called dopamine deteriorate, causing progressive loss of movement control (e.g. bradykinesia, muscle rigidity, tremor at rest, and impaired postural reflexes). A syndrome with similar symptoms is Scans Without Evidence of Dopaminergic Deficits (SWEDDs), in which the patients not present evidence of dopaminergic deficits. The causes of primary parkinsonism, or Parkinson's disease (PD), as well as SWEDDs, are not completely known. Genome-wide association studies (GWASs) have provided substantial contribution to the understanding of the architecture of complex diseases, bringing consistent and important contributions to PD. GWASs have been performed in the past with PD data with results that strongly influenced later developments.

In this work, a study of genetic factors that contributes to the set of tools of association analysis for a better understanding of occurency of PD and SWEDDs. To that end, the set of tools of association analysis is involved in a study case for the PLINK method to execute GWASs with the objective of identifying SNPs which are associated to DP and SweDDs. To determine the level of significance of our results, only real data provided by Parkinson's Progression Markers Initiative (PPMI) was used. PPMI is an international consortium created to indetify biomarkers of DP progression, to better comprehend the etiology of this disease and to provide key tools to increase the probability of success in the development of new therapeutic trials for PD.

The association analysis, done with genotype data from three groups of individuals (healthy, affected by PD and affected by SWEDDs), SNPS that have shown strong connection to PD and SWEDDs were recovered, some of them are already linked in the scientific literature to PD or other degenerative diseases. But, we also have found about 60 SNPs that are not reported in the literature, which show evidence to be strongly related to the propensity to PD or SWEDDs. These results are promising targets for future genomic studies and may contribute to the understanding of the occurrence of PD and SWEDDs. Interestingly, although SWEDDs is a disorder clinically linked to PD by a series of common symptoms, the top ranked SNPs recovered

from the PD dataset were not part of the SWEDDs dataset and conversely, that suggests that those two sets of markers could be more carefully explored in the genomic studies as common SNPs of interest for the two diseases.

Keywords:

SNP. GWAS. Parkinson's disease. SWEDDs. Parkinsonism.

Lista de Figuras

2.1 No nucletídeo do DNA, o açúcar é uma desoxirribose ligada a um único grupo fosfato, e a base nitrogenada que podem ser de dois tipos: purinas (adenina (A) e citosina (C)) e pirimidinas (Guanina (G) e timina (T)). Fonte: Modificada de http://www.ufv.br/dbg/genetica/cap10.htm. 22

2.2 a) Fita dupla de DNA com pares de bases unidos por ligações de hidrogênio. b) Dupla-hélice de DNA. Fonte: Modificada de http://tdbio.blogspot.com.br. . . . 24

2.3 O genoma humano - 46 cromossomos, subdivididos em 22 pares de cromossomos autossômicos e um par de cromossomos sexuais (representado por dois cromossomos X em mulheres, e um cromossomo X e um Y em homens). Fonte: Modificada de http://upload.wikimedia.org 25

2.4 Representação de SNP, a molécula de DNA 1 difere da molécula de DNA 2 na mesma localização. Fonte: http://www.siriusgenomics.com 27

2.5 Ilustração de mulher com a doença de parkinson. Fonte: Modificada de http://body-disease.com 28

2.6 A substância negra pars compacta conforme visto no cérebro sem a doença de parkinson e com a doença de parkinson. Fonte: http://ampark.com.br 29

2.7 PLINK - Terminal de comando 36

4.1 Distribuição das três populações de indivíduos por idade. Fonte: Modificada de http://www.ppmi-info.org/ 47

4.2 Distribuição das três populações de indivíduos pelo gênero. Fonte: Modificada de http://www.ppmi-info.org/ 47

5.1 Resultados de GWAS no conjunto de dados DP/PPMI - O gráfico Manhattan plot mostra os resultados da análise de associação genética dos SNPs mais significativos, que ocorreram com maior frequência nos cromossomos 1, 2, 4, 6, 10, e 11. 53

5.2 Resultados de GWAS no conjunto de dados DP/PPMI - O gráfico Q-Q plot mostra resultados consistentes de associações, onde o desvio curvado entre as linhas X (esperado) e Y (observado) representa o número de associações verdadeiras entre milhares de SNPs não associados. 54
5.3 Resultados de GWAS no conjunto de dados SWEDDs/PPMI - O gráfico Manhattan plot mostra os resultados da análise de associação genética dos SNPs mais significativos, que ocorreram com maior frequência nos cromossomos 1, 6, e 16. 57
5.4 Resultados de GWAS no conjunto de dados SWEDDs/PPMI - O gráfico Q-Q plot mostra resultados consistentes de associações, onde o desvio curvado entre as linhas X (esperado) e Y (observado) representa o pequeno número de associações verdadeiras entre milhares de SNPs não associados. 58

Lista de Tabelas

2.1 Diferenciando SWEDDs e DP 30

2.2 Na tabela de distribuição de contagem os alelos A e a são igualmente comuns (50:50). 32

2.3 Após genotipagem na tabela de distribuição de contagem de alelos, há uma diferença nos dados, o alelo A é muito mais comum nos casos que nos controles. 33

2.4 Contagem das frequências. 34

2.5 Contagem das frequências da doença de Parkinson versus saudável 34

5.1 SNPs identificados no conjunto de dados DP/PPMI, ordenados por *p*-value de alta significância de associação. 50

5.2 Sítios de alta significância recuperados na análise do conjunto de dados DP/PPMI mencionados na literatura. 51

5.3 SNPs identificados no conjunto de dados SWEDD/PPMI, ordenados por *p*-value de alta significância de associação. 55

5.4 Sítios de alta significância recuperados na análise do conjunto de dados SWEDD/PPMI mencionados na literatura. 56

Lista de Acrônimos

CHISQ Teste estatístico Chi-Quadrado, do inglês *Chi-Square* 32

DNA Ácido desoxirribonucleico, do inglês *deoxyribonucleic acid* 21

DP Doença de Parkinson 17

dbSNP Banco de dados de SNPs, do inglês *Single Nucleotide Polymorphism Database* 43

GWAS Estudos de associações no genoma completo, do inglês *Genome Wide Association Study* 18

geno *Missing rate per SNP* 46

MAF Frequência do menor alelo, do inglês *Minor allele frequence* 46

mind *Missing rate per person* 46

HWE Equilíbrio de Hardy-Weinberg, do inglês *Hardy-Weinberg equilibrium* 46

NCBI Centro Nacional de Informação Biotecnológica, do inglês *National Center for Biotechnology Information* 43

NHGRI Instituto de Pesquisa Nacional do Genoma Humano, do inglês *National Human Genome Research Institute* 43

OR Teste estatístico Razão de Probabilidade, do inglês *Odds Ratio* 33

bp pares de base, do inglês *base pairs* 23

PPMI *Parkinson's Progression Markers Initiative* 19

RSID *Identificador de SNPs atribuído pelo NCBI* 44

SNP Polimorfismo de único nucleotídeo, do inglês *single nucleotide polymorphism* . 19

SWEDD Testes que não apresentam evidências de déficit de dopamina, do inglês *scans without evidence of dopaminergic deficit* 18

Sumário

1 Introdução **17**

2 Fundamentação Teórica **21**

2.1 Genética humana básica 21

2.1.1 Moléculas de DNA 22

2.1.2 Cromossomos 23

2.1.3 Polimorfismos Genéticos 24

2.2 Doença de Parkinson 27

2.3 SWEDDs 28

2.4 Estudos caso-controle 30

2.5 Fundamentos de Estatística 31

2.5.1 Teste Qui-Quadrado 31

2.5.2 Razão de chances 33

2.6 Método Computacional 35

2.7 Considerações Finais 37

3 Trabalhos Relacionados **38**

3.1 Análises de dados genéticos 38

3.2 Análises de associação 39

3.2.1 Análises de associação e a Doença de Parkinson 41

3.3 Considerações Finais 42

4 Conjuntos de Dados **43**

4.1 dbSNP 43

4.2 Parkinson's Progression Markers Initiative 44

4.3 Considerações Finais 47

5 Análises e Resultados **48**

5.1 Doença de Parkinson 48

5.2 SWEDDs 53

5.3 Considerações Finais 58

6 Discussão e Conclusão **59**

Referências **62**

Apêndice **67**

A Apêndice **68**

1

Introdução

Parkinsonismo refere-se a uma síndrome neurológica que abrange a bradicinesia, a rigidez muscular, o tremor de repouso e os reflexos posturais prejudicados, e que envolve um amplo diagnóstico diferencial (BOHLHALTER; KAGI, 2011). Segundo BOHLHALTER; KAGI (2011), este conjunto de sintomas é chamado de quatro sinais cardinais: a) Bradicinesia é a dificuldade de realizar movimentos automáticos em sequência e simultâneos, essenciais nas atividades diárias; b) A rigidez muscular ou rigidez, é uma manifestação motora do paciente, que apresenta o sentimento de resistência ao deslocar passivamente o seu membro; c) Tremor de repouso, apresenta-se na forma de tremores não intensos e pode ser interrompido por períodos, sendo observado geralmente quando o membro superior está em uma posição de repouso ou imóvel e diminui quando está em movimento. Este tipo de tremor ocorre com mais frequência nos membros superiores, podendo também ser observado nos membros inferiores e na mandíbula; d) Reflexos posturais prejudicados ou instabilidade postural, é uma anormalidade postural que apresenta como principal risco a queda do paciente. Parkinsonismo é classificado de acordo com as suas causas: a) Parkinsonismo primário, que abrange desordens neurodegenerativas de origem desconhecida ou de origem genética; e b) Parkinsonismo secundário, que deriva de outras causas como drogas, síndrome multi-infarto vascular, toxinas, contágio (e.g. HIV e neurosífilis) e traumas. Em particular, parkinsonismo induzido por droga é prevalente e pode ser sub-diagnosticada em pacientes idosos (ESPER; FACTOR, 2008). O tipo mais comum de parkinsonismo é a Doença de Parkinson (DP), correspondendo a cerca de 75% de todas as formas

deste distúrbio.

Com uma estimativa prevalente entre 1% e 2% em indivíduos com mais de 65 anos de idade, DP é uma das doenças neurodegenerativas mais comum relacionadas à idade , que afeta mais do que o número combinado de pessoas diagnosticadas com esclerose múltipla, distrofia muscular, e doença de Lou Gehrigs (HECKMAN et al., 2014). Estima-se que 4% das pessoas com doença de Parkinson são diagnosticados antes dos 50 anos (DISEASE FOUNDATION, 2015). Estudos estatísticos são muito raros no Brasil, porém um estudo epidemiológico realizado numa cidade em Minas Gerais encontrou uma prevalência de 3,3% de DP em pessoas com idade acima de 65 anos (BARBOSA et al., 2006). Ela ataca o sistema nervoso central, sendo caracterizada pela degeneração das células responsáveis pela produção de dopamina, chamada *substantia nigra pars compacta*, na região do cérebro. Há uma clara correlação entre a perda de dopamina e os sintomas motores da DP (por exemplo, movimentos e comportamentos involuntários) e, particularmente, com bradicinesia.

Outra doença neurológica cuja etiologia é desconhecida é o *Parkinson-mimics* ou Testes que não apresentam evidências de déficit de dopamina, do inglês *scans without evidence of dopaminergic deficit* (SWEDD), na qual os pacientes não apresentam evidências de déficit de dopamina, apesar de estar associada com vários sintomas de DP. Os pacientes com SWEDDs tem a presença de tremor de repouso assimétrico, e muitas vezes, uma proporção desses pacientes podem ter tremor distônico ou tremor associado com distonia (MIAN et al., 2011). No entanto, mesmo com essas características sem o sintoma da verdadeira bradicinesia, consequentemente não há parkinsonismo (BOHLHALTER; KAGI, 2011).

Apesar dos recentes avanços nas pesquisas, a causa da morte neuronal em DP é ainda desconhecida. Existe uma hipótese de que a doença é o resultado da interação entre fatores genéticos que predispõe um indivíduo ao desenvolvimento da doença e fatores ambientais que iniciam o processo neurodegenerativo. Esses fatores podem também estar ligados ao processo de envelhecimento, o que seria o terceiro fator determinante para o aparecimento da doença.

Os Estudos de associações no genoma completo, do inglês *Genome Wide Association Study* (GWAS) tem proporcionado ganhos tangíveis para determinação da arquitetura genética de diversas doenças complexas, cujas as manifestações dependem da exposição a múltiplos fatores

ambientais e sociais e a fatores genéticos. A compreensão da arquitetura genética de doenças complexas traz grandes perspectivas de avanços no que corresponde a melhor compreensão destas doenças, podendo ajudar na identificação de alvos específicos para a sua prevenção, reduzindo o risco de desenvolvimento.

A causalidade genética vem ganhando cada vez mais importância sob o entendimento DP, em particular, GWASs têm demonstrado consistentes contribuições genéticas, avaliadas através de estudos independentes pela comunidade científica (SATAKE et al., 2009; HAMZA et al., 2010; PANKRATZ et al., 2012; HAMZA et al., 2011; HILL-BURNS et al., 2013). Sanchez e colegas realizaram um GWAS, mostrando pela primeira vez um papel claro para variabilidade genética comum no risco de desenvolvimento da DP, e descreve uma possível heterogeneidade genética populacional específica relacionada com esta doença (SIMON-SANCHEZ et al., 2009). Do e colegas descobriram duas novas associações no genoma completo de significância próximos aos genes SCARB2 (rs6812193) e SREBF1 / RAI1 (rs11868035) (DO et al., 2011).

Outros autores têm usado GWASs para tentar identificar genes suscetíveis da DP anteriormente desconhecidos (HILL-BURNS et al., 2014), estabelecendo correlações através de redes genéticas envolvendo genes da DP previamente identificados que provaram serem informativos (ROBINSON, 2014). No entanto, o GWAS na DP não tem sido capaz ainda de identificar sinais de associação conclusivos. Nesse trabalho, nós desenvolvemos um estudo sobre fatores genéticos que possam contribuir para o entendimento da ocorrência da DP e SWEDDs. Para isso, nós usamos o conjunto de ferramentas de análise de associação envolvendo estudo de caso-controle do método PLINK para executar GWASs, a fim de identificar Polimorfismo de único nucleotídeo, do inglês *single nucleotide polymorphism* (SNP) que estão associados à DP e SWEDDs. Para fixar o nível de significância dos nossos resultados, nós optamos por usar somente dados reais fornecidos pelo banco de dados *Parkinson's Progression Markers Initiative* (PPMI). O PPMI é um consórcio internacional projetado para identificar biomarcadores de progressão da DP, tanto para melhorar a compreensão da etiologia da doença, como para fornecer ferramentas cruciais para aumentar a probabilidade de sucesso na elaboração de novos ensaios terapêuticos para DP.

Esta dissertação esta organizada de seguinte forma: No Capítulo 2, revisamos sobre os princípios do conhecimento da genética e sobre o Parkinsonismo. Incluímos também nesta

revisão, conceitos básicos de estatística. No Capítulo 3, descrevemos outros trabalhos que identificaram SNPs que estão associados à DP. No Capítulo 4, apresentamos a fonte de pesquisa dos nossos dados, e são descritos os conjuntos de dados reais que foram utilizados no estudo. No Capítulo 5, exibimos todas as nossas análises e resultados, e por fim, no Capítulo 5, apresentamos nossas discussões e conclusões.

2

Fundamentação Teórica

Este capítulo tem como objetivo descrever a teoria genética, estatística e computacional necessária para uma melhor compreensão deste trabalho e sua contextualização a respeito das áreas do conhecimento envolvidas.

O capítulo está organizado da seguinte forma: Na Seção 2.1, será explicada de forma clara e objetiva as informações essenciais sobre genética envolvidas no estudo. Na Seção 2.2, serão explicadas as informações essenciais sobre a doença de Parkinson e SWEDD. Na Seção 2.3, será explicada a teoria essencial sobre o estudo caso-controle envolvido no método para análise de associação. Na Seção 2.4, resume uma introdução a teoria estatística envolvida no estudo. Na Seção 2.5, será introduzido o contexto computacional envolvido no método para análise de associação.

2.1 Genética humana básica

Esta seção resume uma introdução relevante das informações essenciais sobre genética para uma melhor compreensão do estudo, descrevendo a molécula de Ácido desoxirribonucleico, do inglês *deoxyribonucleic acid* (DNA) e cromossomos, assim como é feita a descrição sobre polimorfismo genético e SNPs.

2.1.1 Moléculas de DNA

De acordo com a Teoria da Evolução, a terra tem cerca de 4,6 bilhões de anos, e as primeiras manifestações de vida iniciaram a cerca de 3.5 bilhões de anos atrás. As primeiras formas de vida eram muito simples, mas devido ao processo contínuo de evolução e diversificação nos últimos bilhões de anos, podemos hoje encontrar organismos muito complexos e muito simples. Os principais responsáveis pela química molecular desses seres vivos são moléculas chamadas de proteínas e ácidos nucleicos. Resumidamente, as proteínas são responsáveis pelo que um ser vivo é, e os ácidos nucleicos codificam a informação necessária para produzir proteínas, além de serem responsáveis por passar essa "receita" para gerações subsequentes (SETUBAL; MEIDANIS, 1997).

Um tipo de ácido nucleico é o ácido desoxirribonucleico (*deoxyribonucleic acid* - (DNA)). Uma molécula de DNA consiste em duas longas cadeias polipeptídicas compostas por quatro tipos de subunidades nucleotídicas. Cada uma dessas cadeias é conhecida como uma cadeia de DNA, ou fita de DNA (ALBERTS et al., 2010).

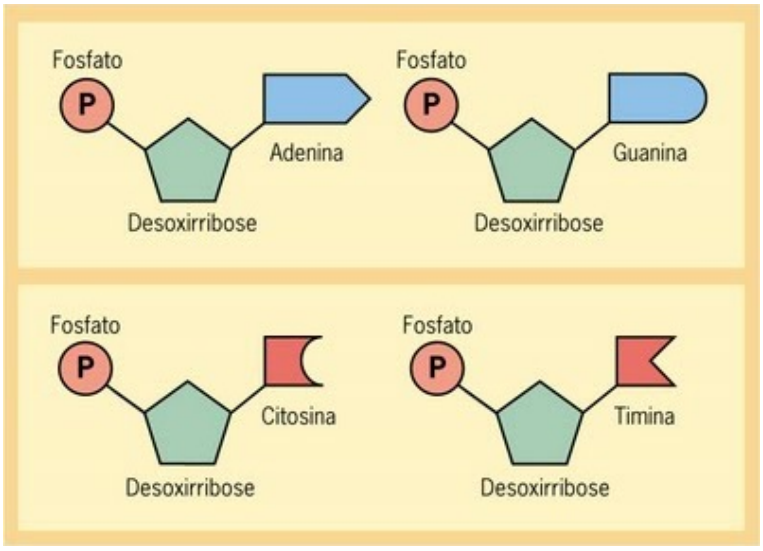

Figura 2.1: No nucletídeo do DNA, o açúcar é uma desoxirribose ligada a um único grupo fosfato, e a base nitrogenada que podem ser de dois tipos: purinas (adenina (A) e citosina (C)) e pirimidinas (Guanina (G) e timina (T)). Fonte: Modificada de http://www.ufv.br/dbg/genetica/cap10.htm.

Essas subunidades nucleotídicas ou nucleotídeos são formados por açúcares com cinco carbonos, aos quais um ou mais grupos de fosfato estão ligados, e uma base contendo nitrogênio. No nucletídeo do DNA, o açúcar é uma desoxirribose ligada a um único grupo fosfato, e a base

(Figura 2.1). Existem quatro tipos de base: adenina (A), citosina (C), Guanina (G) e timina (T). As bases A e G pertencem ao grupo de substâncias chamadas purinas, enquanto que C e T pertencem ao grupo das pirimidinas.

A famosa dupla-hélice da molécula de DNA descoberta por James Watson e Francis Crick em 1953, é decorrente das características químicas e estruturais de suas duas cadeias polinucleotídicas. Essas duas cadeias são mantidas unidas por ligações de hidrogênio entre as bases das duas fitas, todas as bases são voltadas para o interior da dupla-hélice, e o esqueleto de açúcar-fosfato encontra-se na região externa (Figura 2.2 a) (ALBERTS et al., 2010). Em ambos os casos, uma base do tipo purina forma par com uma base do tipo pirimidina, ou seja, A sempre forma par com T, e G com C. Assim, as bases A e T, são o complemento uma da outra, ou o par de base complementar. O mesmo ocorre com C e G, que são bases com pareamento complementar. Esses pares são conhecidos como pares de base, do inglês *base pairs* (bp) de Watson-Crick. Os membros de cada par de bases somente encaixam-se na dupla-hélice se as duas fitas da hélice estiverem na posição antiparalela, isto é, somente se a polaridade de uma fita estiver em orientação oposta à da outra fita (Figura 2.2 b) (ALBERTS et al., 2010). A consequência fundamental dessa estrutura é que ela possibilita inferir a sequência de nucleotídeos de uma fita dada a outra fita, que é exatamente complementar.

Em sua forma simples, um gene pode ser visualizado como um segmento de uma molécula de DNA, contendo a informação necessária para construir uma proteína ou uma molécula de RNA (SETUBAL; MEIDANIS, 1997). Nos seres humanos um gene pode ter cerca de 10.000 pb, o ponto preciso onde um gene começa e termina é reconhecido através de certos mecanismos celulares.

2.1.2 Cromossomos

O conjunto completo de cromossomos que situa-se dentro do núcleo celular é chamado de genoma humano (SETUBAL; MEIDANIS, 1997). A composição de genes no genoma humano, é especificada no DNA dos 46 cromossomos humanos (THOMPSON; MCINNES; WILLARD, 1993) (Figura 2.3).

Entende-se como cromossomo a estrutura compactada e organizada por proteínas estru-

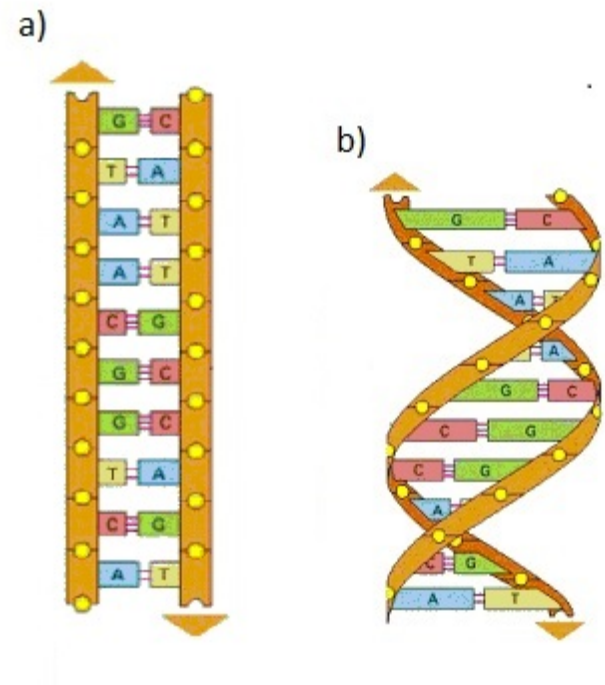

Figura 2.2: a) Fita dupla de DNA com pares de bases unidos por ligações de hidrogênio. b) Dupla-hélice de DNA. Fonte: Modificada de http://tdbio.blogspot.com.br.

turais de uma molécula de DNA dentro do núcleo celular. No núcleo de um célula eucariótica, os cromossomos aparecem em pares, e são chamados pares homólogos de cromossomos (SETUBAL; MEIDANIS, 1997). Os cromossomos do par são muito similares na estrutura, no entanto, não são totalmente idênticos. Um cromossomo é oriundo da mãe e o outro do pai, e são diferenciados devido à existência de variações ou alterações genéticas.

Nos seres humanos os 46 cromossomos normalmente contém 23 pares de cromossomos, subdivididos em 22 pares de cromossomos autossômicos e um par de cromossomos sexuais (representado por dois cromossomos X em mulheres, e um cromossomo X e um Y em homens) (OLAZAR, 2013).

2.1.3 Polimorfismos Genéticos

Quando comparamos no mesmo cromossomo a sequência genética de diferentes indivíduos da população, é notado que a sequência de pares de bases varia de indivíduo para indivíduo, mesmo apresentando em grande parte da sequência genética similaridade para todos os indivíduos. Na população humana, os indivíduos do mesmo sexo compartilham uma porcentagem

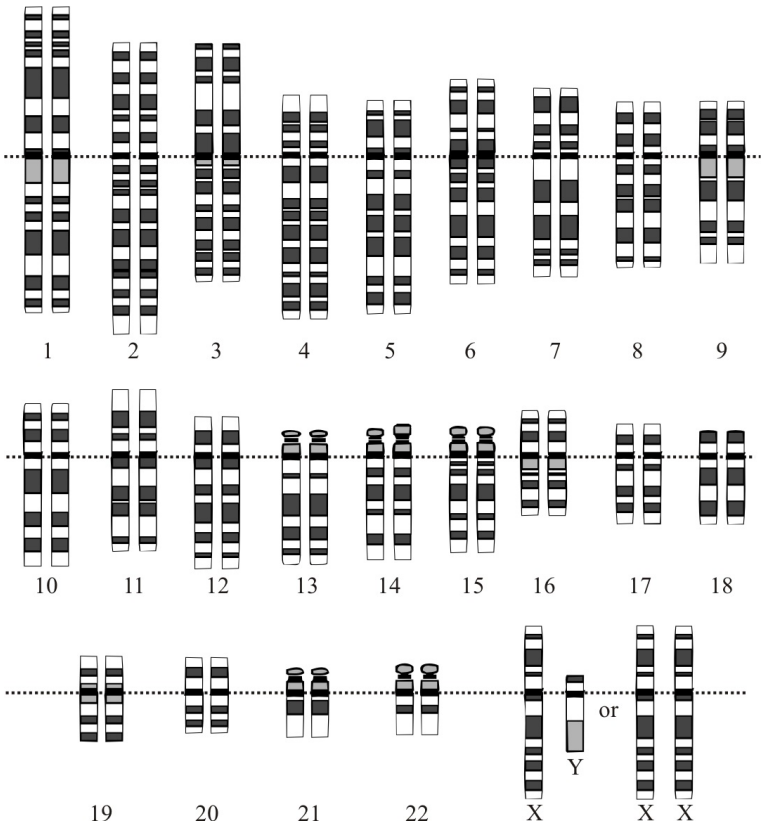

Figura 2.3: O genoma humano - 46 cromossomos, subdivididos em 22 pares de cromossomos autossômicos e um par de cromossomos sexuais (representado por dois cromossomos X em mulheres, e um cromossomo X e um Y em homens). Fonte: Modificada de http://upload.wikimedia.org

muito elevada de sua sequência de DNA, a semelhança entre o dna genômico de duas pessoas é cerca de 99% (OLAZAR, 2013).

A descrição da constituição genética de uma população conduz ao estudo das frequências relativas dos indivíduos com determinados genótipos. Assim, a constituição genética dos indivíduos da população humana, com relação aos genes, que ela transporta, é descrita pela relação das frequências gênicas ou alélicas. Atualmente aceita-se que um gene com frequência gênica entre 1% e 99% deve ser classificado como gene polimorfo, aquele com frequência inferior a 1% deve ser denominado gene idiomorfo, enquanto um gene com frequência superior a 99% deve ser classificado como gene monomorfo (BEIGUELMAN; EDITORA, 2008). Assim, por exemplo, se os alelos (formas alternativas do mesmo gene) *A*, *a* e *a1* de determinado loco (posição que o gene ocupa no cromossomo) tiverem frequências gênicas iguais, respectivamente, a 0,600, 0,395 e 0,005 diremos que os alelos *A* e *a* são polimorfos, enquanto o alelo *a1* será dito idiomorfo.

Por sua vez, as alterações genéticas nos caracteres que resultam de sítios (localizações especificas de genes) que incluem pelo menos dois alelos polimórficos são ditas polimorfismos genéticos ou sistemas genéticos polimórficos (BEIGUELMAN; EDITORA, 2008). Portanto, há polimorfismos genéticos quando as modificações de DNA produzem pelo menos duas formas alternativas de um mesmo gene. Essas pequenas variações no genoma por substituição, remoção ou inserção de informação genética contida no DNA fundamentam boa parte da variabilidade fenotípica interindividual.

O tipo mais comum de polimorfismo envolve variação genética em um par de base isoladamente, que aparecem como consequências de mutações. O polimorfismo de único nucleotídeo ou SNP (*single nucleotide polymorphisms*), são substituições de um único nucleotídeo de uma base de DNA (A, T, C, ou G) para outra que ocorrem em mais do que um por cento da população em geral (Figura 2.4) (LEARN.GENETICS, 2014). Por exemplo, um SNP pode substituir o nucleotídeo citosina (C) pelo nucleotídeo de timina (T) em uma determinada sequência de DNA.

A sequência do DNA apresenta diferença de 0,1% entre os indivíduos que se deve à presença dos polimorfismos. Os SNPs são os polimorfismos de DNA mais abundantes no genoma humano, e estão presentes numa frequência alélica mínima de 1 a 5% na população.

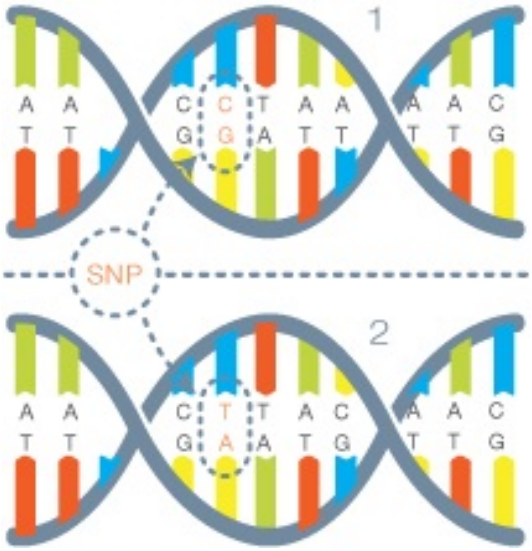

Figura 2.4: Representação de SNP, a molécula de DNA 1 difere da molécula de DNA 2 na mesma localização. Fonte: http://www.siriusgenomics.com

Eles ocorrem uma vez em cada 300 nucleotídeos em média, o que significa que existem cerca de 10 milhões de SNPs no genoma humano (NIH/NLM, 2014). Quando SNPs ocorrem dentro de um gene ou de uma região reguladora perto de um gene, pode desempenhar um papel mais direto em uma doença específica por afetar a função do gene. Dessa forma, podem ajudar aos cientistas a localizarem genes que estão associados com doenças, atuando como marcadores biológicos. SNPs também podem ser usados para rastrear a herança de genes relacionados a doenças entre membros de uma família (NIH/NLM, 2014).

2.2 Doença de Parkinson

Parkinsonismo é uma síndrome neurológica com quatro sinais cardinais: a bradicinesia, a rigidez muscular, o tremor de repouso e os reflexos posturais prejudicados (figura 2.5), cujo diagnóstico depende da presença da bradicinesia (BOHLHALTER; KAGI, 2011).

O tipo mais comum de parkinsonismo primário é a doença de Parkinson (DP), uma doença idiopática que corresponde a cerca de 75% de todas as formas de parkinsonismo (NINA, 2014). Doença de Parkinson é uma doença degenerativa e crônica do sistema nervoso central, caracterizada pela morte dos neurônios dopaminérgicos em duas partes do cérebro: na substância negra pars compacta e suas projeções para o estriado, causando vários déficits motores (Figura 2.6). Existe uma correlação clara entre a perda de dopamina e os sintomas motores de DP, e em

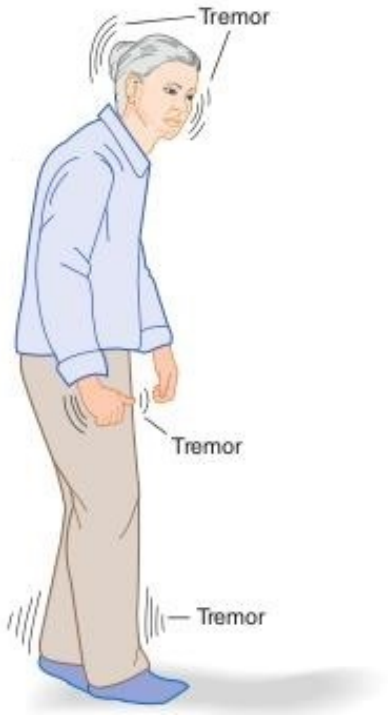

Figura 2.5: Ilustração de mulher com a doença de parkinson. Fonte: Modificada de http://body-disease.com

particular com a bradicinesia.

Como não existe nenhum exame complementar que confirme o diagnóstico, o conhecimento clínico e a perspicácia são as chaves para o diagnóstico. O critério diagnóstico requer, pelo menos, dois sintomas motores. E estes sintomas de distúrbios motores só tornam-se evidentes após a perda de, pelo menos, 60% dos neurônios na substância negra, resultando em uma drástica redução do conteúdo de dopamina no estriado de cerca de 80% inferior ao normal.

A DP tem ainda sua etiologia desconhecida, e afeta entre 100 e 200 pessoas por 100.000 com mais de 40 anos de idade. Não ocorre comumente em pessoas jovens com menos de 40 anos, e a incidência aumenta rapidamente após os 60 anos, sendo a média da idade de diagnóstico de 70,5 anos. A taxa de incidência é maior em homens (3:2) (MEDICINANET, 2014).

2.3 SWEDDs

Outra doença neurológica cuja etiologia é desconhecida é o *Parkinson-mimics* ou SWEDDs (*scans without evidence of dopaminergic deficit*) na qual os pacientes não apre-

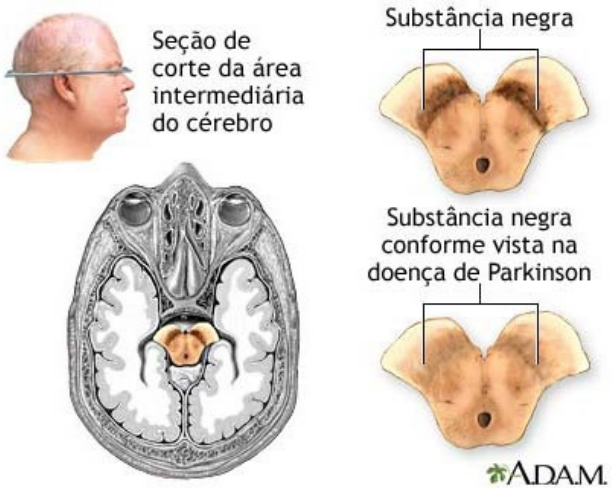

Figura 2.6: A substância negra pars compacta conforme visto no cérebro sem a doença de parkinson e com a doença de parkinson. Fonte: http://ampark.com.br

sentam evidências de déficit de dopamina, apesar de estar associada com vários sintomas de DP (BOHLHALTER; KAGI, 2011). A falta de déficit de dopamina tem sido explicada ou pela baixa sensibilidade as técnicas de exame de imagem ou pela ausência do verdadeiro parkinsonismo.

Os pacientes com SWEDDs tem a presença de tremor de repouso assimétrico, ou seja, o tremor geralmente aparece assimetricamente em apenas um lado. E muitas vezes, uma proporção desses pacientes podem ter tremor distônico ou tremor associado com distonia.

No entanto, mesmo com a presença desses sintomas característicos, se o sintoma da verdadeira bradicinesia está ausente, consequentemente não há parkinsonismo (BOHLHALTER; KAGI, 2011). A Tabela 2.1 resume algumas características clínicas que ajudam a discriminar DP e SWEDDs.

Sintoma	**DP ou SWEDDs**
Bradicinesia	DP
Características de tremor	
Tremor de repouso re-emergente	DP
Tremor em flexão-extesão do polegar	SWEDDs
Tremor da cabeça	DP
Predominantemente posicional	SWEDDs
Tremor distônico ou tremor associado com distonia	SWEDDs
Resposta adequada para o tratamento com L-Dopa	DP
Sintomas não motores, incluindo distúrbios do olfato	DP

Tabela 2.1: Diferenciando SWEDDs e DP

Os pacientes com SWEDDs não apresentam resposta clínica ao tratamento com L-Dopa (fármaco usado no tratamento da DP), apesar de ter características clínicas semelhantes aos da DP. Assim, estes pacientes parecem apresentar diferentes fisiopatologia, prognóstico e requisitos de tratamento. A proporção de pacientes com SWEDDs indevidamente diagnosticados com DP, que posteriormente, tem a imagem funcional dopaminérgica normal é de aproximadamente 10% (SCHWINGENSCHUH et al., 2010).

2.4 Estudos caso-controle

A Epidemiologia é a ciência que estuda os padrões da ocorrência de doenças em populações humanas e os fatores determinantes destes padrões (MENEZES, 2001). Aborda o processo saúde-doença em grupos de pessoas, que pode variar de pequenos grupos a populações inteiras. Suas aplicações variam desde a descrição das condições de saúde da população, da investigação dos fatores determinantes de doenças, da avaliação do impacto das ações até a avaliação da utilização dos serviços de saúde, contribuindo para a saúde da população.

O estudo caso-controle é um dos tipos de estudos epidemiológicos, e é o mais comum na literatura científica. É considerado por muitos autores o estudo mais poderoso e eficiente, garantindo robustez quando se estuda um grande número de SNPs (IOANNIDIS et al., 2001).Porém, de uma perspectiva epidemiológica, a principal desvantagem deste estudo é que às vezes levam ao surgimento de falsos positivos (CARDON; BELL, 2001). Os estudos mais comuns envolvem a análise de indivíduos independentes amostrados por uma população sem correlação com a família original (estudo de associação sobre base populacional), embora os indivíduos

possam ser amostrados por famílias (estudo de associação baseados em família).

O estudo é projetado para ajudar a determinar se uma exposição está associada com um resultado, isto é, doença ou condição de interesse. No estudo, primeiro são identificados os casos (indivíduos que já têm doença ou condição de interesse), grupo conhecido por ter o resultado, e os controles (indivíduos que não têm a doença ou sem a condição de interesse), grupo conhecido por ser livre de resultado. Em seguida, é feita uma revisão para saber quais indivíduos em cada grupo tiveram associados com o fator de risco em estudo (a exposição), comparando a frequência da exposição do grupo caso ao grupo controle. Assim, um estudo de caso-controle é sempre uma retrospectiva porque começa com um resultado, e então, segue com uma revisão para investigar as exposições.

Um estudo de caso-controle de associação genética compara a frequência de alelos ou genótipos por meio de marcadores genéticos, geralmente SNPs, em indivíduos de uma determinada população com e sem um determinado traço de doença, a fim de determinar se existe uma associação estatisticamente significativa entre o traço de doença e o marcador genético (CLARKE et al., 2011). Consequentemente, o aumento na frequência de um genótipo na comparação de casos com os controles indica que a presença desse genótipo pode aumentar o risco da doença.

2.5 Fundamentos de Estatística

Neste estudo, nos usamos o método PLINK que realiza análise estatística para execução da associação genética envolvendo estudo de caso-controle. Os testes de associação genéticos foram realizados com o teste Qui-Quadrado e a Razão de chances. Estes testes de associação serão descritos brevemente nas subseções a seguir.

2.5.1 Teste Qui-Quadrado

O propósito por trás da análise de associação é examinar cada SNP, um por um, testando para ver se existe uma diferença na frequência dos alelos observados do caso em comparação com controle. Se esta diferença é estatisticamente significativa, então esse alelo pode ser dito

associado com o fenótipo (WALTER; DIVISION, 2014).

O método utilizado para testar estatisticamente as frequências é o Teste estatístico Chi-Quadrado, do inglês *Chi-Square* (CHISQ). O teste qui-quadrado resume a diferença entre as frequências alélicas observadas em um SNP e as frequências alélicas esperadas (WALTER; DIVISION, 2014). Por exemplo, executando manualmente o teste de alelos em um único sítio, considerando um grupo com 100 indivíduos, divididos em 50 casos e 50 controles. Para um determinado SNP, assumimos que os alelos A e a são igualmente comuns na população (50:50). Se não existe associação entre o alelo presente e o fenótipo, nós deveríamos esperar a seguinte distribuição de contagem de alelos (Tabela 2.2):

	Alelo a	**Alelo A**
Case	50 [Ea]	50 [EA]
Control	50 [Ea]	50 [EA]

Tabela 2.2: Na tabela de distribuição de contagem os alelos A e a são igualmente comuns (50:50).

Após o grupo ser genotipado, podemos observar na tabela de contingência (Tabela 2.3) uma distribuição diferente das frequências alélicas para este SNP.

	Alelo a	**Alelo A**
Case	25 [Oa]	75 [OA]
Control	75 [Oa]	25 [OA]

Tabela 2.3: Após genotipagem na tabela de distribuição de contagem de alelos, há uma diferença nos dados, o alelo A é muito mais comum nos casos que nos controles.

Para determinar se essas duas distribuições foram significativamente diferentes, é realizado o teste qui-quadrado:

n = 4 (existem quatro possíveis combinações de alelo)

X^2 = (Oa-Ea)^2/Ea + (Oa-Ea)^2/Ea + (OA-EA)^2/EA + (OA-EA)^2/EA

= (25-50)^2/50 + (75-50)^2/50 + (75-50)^2/50 + (25-50)^2/50

= 625/50 + 625/50 + 625/50 + 625/50

= 12.5 + 12.5 + 12.5 + 12.5

= 50

Após, para determinar a significância deste valor é calculado o *p*-value do resultado do qui-quadrado. O *p*-value é uma medida de quão fortemente associados os valores do SNP estão com os valores do fenótipo. Desta forma, o *p*-value <0,0001 indica que a probabilidade de se obter estas frequências é muito baixa (<0,01%), e portanto a importância da significância da associação é alta (WALTER; DIVISION, 2014).

2.5.2 Razão de chances

Em um estudo de caso-controle, a força de uma associação é medida pelo Teste estatístico Razão de Probabilidade, do inglês *Odds Ratio* (OR). Neste tipo de estudo, a OR de interesse é a probabilidade da doença (a probabilidade de que a doença está presente em comparação com a probabilidade de que ela está ausente) em indivíduos expostos versus indivíduos não expostos, sempre que a prevalência da doença estudada nos indivíduos não expostos for igual ou menor que 5% (CLARKE et al., 2011).

Considerando a tabela de contingência 2x2 (Tabela 2.4):

O valor OR pode ser calculado pela seguinte equação:

		Casos	Controles
Fator de risco (exposição)	Sim	a	b
	Não	c	d

Tabela 2.4: Contagem das frequências.

$$OR = \frac{a \times d}{b \times c}$$

A OR quantifica a associação entre a exposição e a doença e varia entre 0 a infinito. O valor entre 0 e 1 indica não associação entre a exposição e a doença, e o valor de 1 a mais infinito indica associação entre a exposição e a doença. Quanto mais distante de 1, mais forte é a associação.

Por exemplo, considerando o alelo A em um determinado SNP na tabela de contingência 2x2 (Tabela 2.5), onde a população de origem dos casos é de indivíduos com a doença de Parkinson e os controles representam a população que não tem indivíduos com a doença de Parkinson e que não tem parentes de sangue ou de primeiro grau com a doença de Parkinson, já que esse tipo de doença pode está ligada a hereditariedade.

		Casos	Controles
Possui o alelo A	Sim	(a) 100	(b) 33
	Não	(c) 55	(d) 57

Tabela 2.5: Contagem das frequências da doença de Parkinson versus saudável

O valor calculado da OR para este exemplo será:

$$OR = \frac{100 \times 57}{33 \times 55} = 3,140$$

Nesse exemplo o valor do OR é 3,140, e como este valor varia entre 1 a mais infinito, podemos que afirmar que o alelo A em um determinado SNP está associado com a doença de Parkinson.

Os dados sob estudo são amostrais, assim, deve ser considerado que as estimativas observadas podem refletir meras flutuações amostrais do verdadeiro efeito da exposição à doença (OLAZAR, 2013). O verdadeiro efeito nunca será conhecido, mas pode-se dispor de uma “boa” estimativa dele quando se tem uma amostra representativa da população de referência.

2.6 Método Computacional

Nesse trabalho, nós usamos o método PLINK para executar GWASs. O método PLINK (PURCELL et al., 2007; PURCELL, 2014), é um conjunto de ferramentas cujo o foco é realizar análise de dados genótipo e fenótipo. PLINK possui a vantagem de ser software livre de código aberto, além de ser uma ferramenta de análise genética projetada para lidar com grandes conjuntos de dados que contém centenas de milhares de marcadores genotipados de milhares de indivíduos que podem ser rapidamente manipulados e analisados na sua totalidade (PURCELL et al., 2007).

PLINK abrange cinco domínios principais de função: gerência de dados, estatísticas de resumo, estratificação da população, análise de associação e estimativa de identidade por descendência. Nesse trabalho, nos concentramos na análise de associação no contexto de base populacional. Em particular, no método básico para associação da doença envolvendo estudo de caso-controle. No PLINK, o teste de associação para traço de doença é baseado na comparação das frequências alélicas entre dois grupos de indivíduos (casos e controles) (PLINK, 2014a).

O método PLINK foi instalado e configurado no ambiente Linux. O PLINK é um aplicativo sem interface gráfica, então os comandos de teste de associação básico de caso-controle foram executados no terminal de comando Linux (Figura 2.7). Cada comando PLINK gera um arquivo de log, que fornece informações úteis sobre o conjunto de dados e o comando que acabou de ser executado. Os arquivos de log gerados com os comandos usados nesse estudo podem ser conferidos no Apêndice A. Além do arquivo de log, a maioria dos comandos PLINK geram também arquivos adicionais contendo resultados. Estes têm diferentes extensões de arquivos específicos para cada comando. Por exemplo, no caso do teste de associação básico de caso-controle executado nesse estudo, o arquivo gerado foi o $pdprocessed.assoc$.

A fim de tornar a análise mais confiável, vários critérios foram aplicados ao teste de associação básico de caso-controle (PLINK, 2014b) para definir se um SNP ou um indivíduo no conjunto de dados seria incluído nas análises, tais como:

- **A frequência do alelo**: exclui automaticamente SNPs com base em um limiar para a menor frequência alélica (*Minor Allele Frequency* - (MAF)).

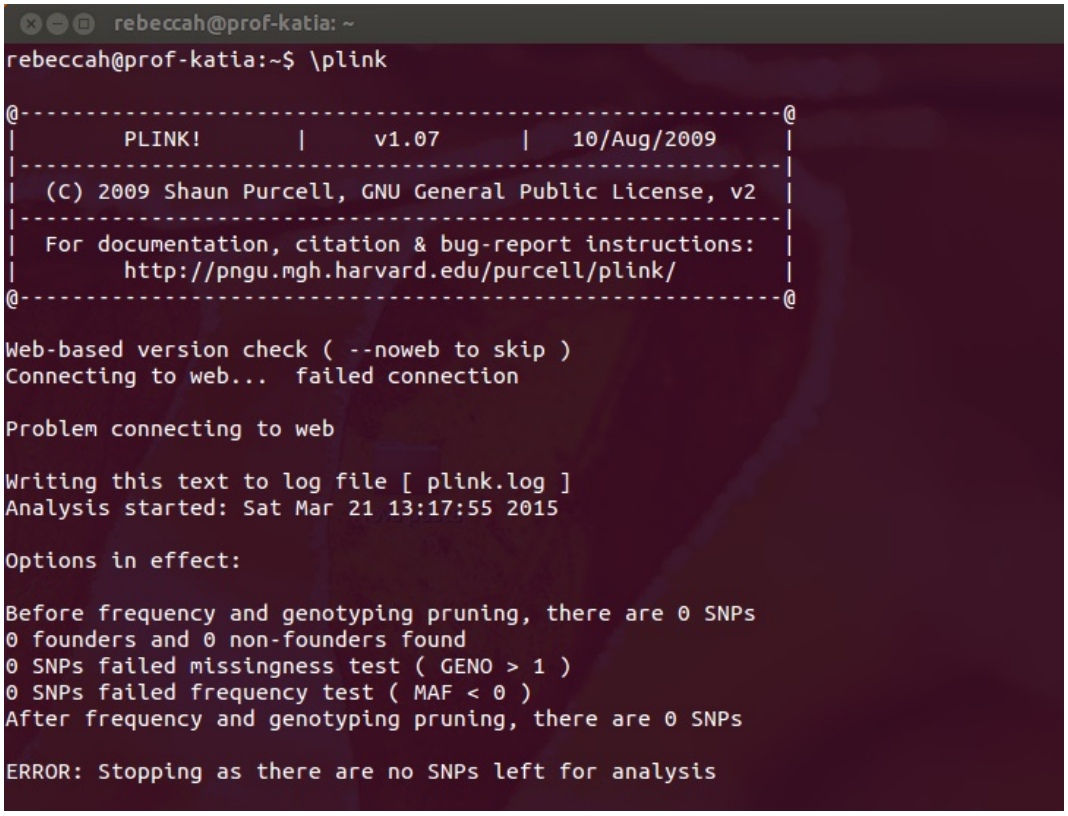

Figura 2.7: PLINK - Terminal de comando

- **Missing rate per SNP**: exclui SNPs com base na taxa de genótipos faltantes.
- **Missing rate per person**: exclui indivíduos com muitos dados de genótipos faltantes.
- **Equilíbrio de Hardy-Weinberg (HWE)**: exclui marcadores que falham no teste de Hardy-Weinberg em um limite de significância especificado.

Os resultados da análise foram classificados pelo *p*-value, que representa a importância do SNP no conjunto de dados utilizado. Então, após classificar os resultados por *p*-value, os SNPs com valor de *p*-value < $10e^{-4}$ foram investigados na literatura científica. Genes foram associados aos SNPs selecionados por meio de busca em bases de dados (NCBI-Gene, OMIM, Pdgene, ScanDB, etc) especializadas e confiáveis, disponíveis publicamente no NCBI, MPI, etc. (NCBI, 2014a; PDGENE, 2014; COX et al., 2014). Genes associados a SNPs selecionados também foram investigados, uma vez que foram estudados mais amplamente pela comunidade científica biológica de SNPs. Vários dos SNPs selecionados (ou os genes onde eles estão localizados) foram associados na literatura com DP, patologias neurológicas, ou outros tipos de doenças, validando os resultados.

2.7 Considerações Finais

Este capítulo apresentou os principais conceitos que envolvem este estudo. Primeiramente, foi apresentada as informações essenciais sobre genética, descrevendo a molécula de DNA e cromossomos, assim como polimorfismo genético e SNPs. Subsequentemente, as informações essenciais sobre a doença de Parkinson e SWEDD foram introduzidas. Após, foram apresentadas a teoria essencial sobre o estudo caso-controle envolvido no método para análise de associação, e a teoria sobre o teste Qui-Quadrado e a Razão de chances envolvidos na análise estatística. Por fim, o contexto computacional envolvido na análise de associação foi introduzido, apresentando o método PLINK.

3

Trabalhos Relacionados

O trabalho desenvolvido nesta pesquisa envolve a análises de dados genéticos. Dessa forma, neste capítulo será descrita brevemente a análises de dados genéticos, e em seguida será explicada as principais informações sobre análises de associação e a suscetibilidade da doença de Parkinson associada a genes/SNPs.

Este capítulo está organizado da seguinte forma: A Seção 3.1, introduz resumidamente as análises de dados genéticos. Na Seção 3.2, serão introduzidas as informações sobre análises de associação e sobre identificação de genes/SNPs de suscetibilidade a DP.

3.1 Análises de dados genéticos

Bioinformática é o ramo sinérgico entre Computação, Matemática e Biologia Molecular que contribui com modelos, análises estatísticas, algoritmos e sistemas de computação, entre outras contribuições teóricas e práticas à área de Biologia Molecular (CLOTE; R.BACKOFEN, 2000). Tem como objetivo realizar análises de dados biológicos, como sequências de bases de DNA e genes, predizer a estrutura e função de diversas macromoléculas (MOUNT, 2004). É um ramo do conhecimento relativamente recente, surgiu da necessidade do uso de ferramentas computacionais para análises de dados genéticos, originado com o projeto genoma, na década de 1990, enfatizando o desenvolvimento de ferramentas para realizar o armazenamento e manipulação dos dados biológicos gerados durante um projeto de sequenciamento. Assim, é especialmente

dedicada aos vários e complexos problemas oferecidos pela biologia molecular.

Em 1990, o Projeto Genoma Humano foi lançado e, quase 14 anos depois, a sequência completa do genoma humano foi disponibilizada (CONSORTIUM et al., 2004), a um custo estimado de US $2,7 bilhões. Desde então, os dados genéticos tem sido recolhidos a uma taxa continuamente crescente. O rápido crescimento de ferramentas genômicas, tais como matrizes de genotipagem de alelos de SNPs e geração de sequenciamento de DNA, tem produzido uma quantidade sem precedentes de informações sobre os genótipos dos indivíduos em muitas espécies (KARCHIN, 2009). Rastrear SNPs funcionais é um dos principais desafios da genética moderna, e um novo ramo da biologia computacional surgiu para apoiar esse esforço. Há aproximadamente uma década surgiram as primeiras técnicas computacionais com o intuito de prever o impacto biológico de SNPs, como a previsão de *Non-synonymous SNP* (nsSNPs) em genes associados a doenças humanas (SUNYAEV et al., 1999) e estudos de suscetibilidade a doenças hereditárias em humanos associadas a SNPs (WANG; MOULT, 2001). Na previsão de nsSNPs em genes associados a doenças humanas, novos nsSNPs foram identificados por análise computacional de um banco de dados público, permitindo avaliar o nível de abrangência do conhecimento de nsSNPs estudados e de genes de relevância médica. No estudo de suscetibilidade a doenças hereditárias em humanos associadas a SNPs foi analisado o efeito de um conjunto de mutações causadoras de doenças que derivam de SNPs, através de estudos de mutagénese *in vitro*, juntamente com o contexto estrutural de proteína de cada mutação.

Posteriormente, uma variedade de métodos foi introduzida, como o servidor Web para prever o efeito de um nsSNPs sobre a estrutura e função da proteína *PolyPhen* (RAMENSKY; BORK; SUNYAEV, 2002), o *SNP Function Portal*, um banco de dados web para explorar a função de SNP (WANG et al., 2006), e o software *Protein Analysis THrough Evolutionary Relationships* (PANTHER) cujo objetivo é inferir as funções de genes com base em suas relações evolutivas (MI et al., 2007).

3.2 Análises de associação

A identificação de SNPs que estão associados com risco do desenvolvimento de doença complexa é um objetivo importante dos estudos modernos de genética. Estes conhecimentos

podem ser utilizados tanto para a compreensão dos mecanismos fundamentais das doenças complexas como para gerar perfis de risco individuais que são úteis no contexto da saúde pública.

A metodologia para estudos de associações no genoma completo (*genome-wide association studies* - GWAS) foi muito aguardada, pois proporcionaria uma abordagem eficaz e imparcial para fazer grandes progressos em nossa compreensão das bases genética de doenças. Consequentemente, os GWAS surgiram como ferramentas populares para identificar variantes genéticas que estão associadas com risco de doenças. Tendo como princípio básico fundamental a hipótese doença comum/variante comum. De acordo com a hipótese doença comum/variante comum, as doenças genéticas complexas ocorrem quando muitos polimorfismos, cada qual com um efeito modesto e baixa penetrância, são herdados (KUMAR, 2011).

Os grandes avanços na tecnologia de genotipagem permitiram uma rápida triagem no genoma completo de variantes comuns em grandes populações e inauguraram uma nova era na investigação da base genética de doenças complexas. Até certo ponto, GWAS têm revolucionado a genética na maneira de lidar com as doenças. Um número crescente de GWAS tem sido publicado, tornando-se rapidamente uma metodologia padrão e uma ferramenta valiosa na compreensão genética das doenças complexas. Isso ocorreu por causa de dois principais fatos: primeiro, GWAS não necessitam de conhecimento prévio da biologia da doença e, portanto, pode ser usado de um modo semelhante no estudo de qualquer doença; segundo, por causa da natureza de GWAS, se fazem necessárias grandes quantidades de amostras a serem testadas simultaneamente, consequentemente, não há limitação por tamanho de conjunto de dados.

Um GWAS típico é composto por uma fase de descoberta, na qual um conjunto inicial de sítios de susceptibilidade promissores são identificados, seguido por uma etapa de validação, na qual os SNPs identificados na fase inicial de descoberta são replicados em um grupo de estudo separado (WU et al., 2010). Logo, a abordagem padrão para analisar GWAS na fase de descoberta envolve a análise individual de SNPs e a classificação dos SNPs com base no seu *p*-value individual, seguida pela etapa a validação.

O PLINK é o método mais utilizado em GWAS, estando em evidência pelas 3.624 citações em artigos científicos, quando procurado na biblioteca de medicina Pubmed Central [http://www.ncbi.nlm.nih.gov/pubmed, em 14 de janeiro de 2015].

3.2.1 Análises de associação e a Doença de Parkinson

A primeira identificação de genes de suscetibilidade a DP ocorreu em 1997, quando uma mutação foi identificada no gene α-synuclein localizado no cromossomo 4, que codifica para uma proteína pré-sináptica (POLYMEROPOULOS et al., 1997). Polymeropoulos e colegas descobriram uma mutação *missense* (mutação que resulta da substituição de um par de bases por outro) no gene SNCA ou α-synuclein causadora da DP Mendeliana. As formas mendelianas da DP representam apenas 10-20% dos casos da doença (SPATARO et al., 2014). A DP Mendeliana tem sido ligada a mutação patogênica nos seguintes genes:

- α-synuclein (SNCA/PARK1) localizado no cromossomo 4 (POLYMEROPOULOS et al., 1997; SINGLETON et al., 2003).
- Parkinson Disease Protein 2 (PARK2/PARKIN) localizado no cromossomo 6 (KITADA et al., 1998).
- PTEN induced putative kinase 1 (PINK1/PARK6) localizado no cromossomo 1 (VALENTE et al., 2004).
- dj-1 protein (DJ1/PARK7) localizado no cromossomo 1 (BONIFATI et al., 2003).
- Leucine-rich repeat kinase 2 (LRRK2/PARK8) localizado no cromossomo 12 (ZIMPRICH et al., 2004; PAISAN-RUIZ et al., 2004).
- ATPase type 13A2 (ATP13A2/PARK9) localizado no cromossomo 1 (RAMIREZ et al., 2006).
- vacuolar protein sorting 35 homolog (VPS35/PARK17) localizado no cromossomo 16 (VILARINO-GUELL et al., 2011).

A forma idiopática da DP envolve complexas interações entre o genoma e exposições ambientais (HAMZA et al., 2011; HILL-BURNS et al., 2013). A DP Idiopática e DP Mendeliana tem sido associadas aos genes glucosidase, beta, acid (GBA), LRRK2 e human leukocyte antigen (HLA). O HLA está mais fortemente associado com a DP Idiopática, ao contrário dos outros dois genes (HAMZA et al., 2010).

Importantes resultados dos estudos de GWAS tem sido obtidos para se compreender o mecanismo molecular da DP, como a descoberta de numerosos loci de susceptibilidade sem separar os subtipos da doença (SIMON-SANCHEZ et al., 2009; HAMZA et al., 2010, 2011; HILL-BURNS et al., 2014). Satake e colegas (SATAKE et al., 2009) identificaram novos sítios de suscetibilidade para DP, PARK16 e BST1, e também detectaram fortes associações em SNCA e LRRK2. Em Do e colegas (DO et al., 2011) foram descobertas duas novas associações significativas no genoma completo com DP, SNP rs6812193, próximo do gene SCARB2, e SNP rs11868035, próximo do gene SREBF1/RAI1. Panktratz e colegas (PANKRATZ et al., 2012) identificaram um novo sítio de suscetibilidade para DP, RIT2, e confirmaram a associação de vários genes conhecidos a DP. Em Hill-Burns e colegas (HILL-BURNS et al., 2013) foi identifica um novo gene associado a DP através da interação com o efeito do tabagismo/nicotina, sugerindo que o gene SV2C desempenha um papel na patogénese da DP. Ayuso e colegas (AYUSO et al., 2014) identificaram dois polimorfismos que modificam a atividade de transcrição do gene, nomeado VNTR (GT)n e o SNP rs2071746, que estão fortemente associados com risco da DP. Em grande parte destes estudos publicados o método PLINK foi aplicado para realizar as análise estatísticas. Dessa forma, nesse trabalho nós aplicamos também o método PLINK para realizar as análise estatísticas com o intuito de identificar várias regiões genômicas que apresentaram fortes sinais de seleção e associação, incluindo SNPs/genes candidatos a associação com DP e SWEDDs.

3.3 Considerações Finais

Este capítulo apresentou uma breve descrição sobre a análises de dados genéticos e os principais trabalhos correlatos ao estudo proposto. Primeiramente, foram introduzidas resumidamente as análises de dados genéticos. Após, foram explicadas as principais informações sobre a metodologia GWAS e foram apresentados vários trabalhos que objetivam a análises de associação para a identificação de genes de suscetibilidade a DP.

4

Conjuntos de Dados

O trabalho desenvolvido nesta pesquisa utilizou um conjunto de dados genéticos. Dessa forma, neste capítulo serão descritos brevemente o banco de dados de SNPs usado para pesquisar as informações dos SNPs identificados. Subsequentemente, também serão descritos os conjuntos de dados reais de caso e controles fornecidos por PPMI.

O capítulo está organizado da seguinte forma: Na Seção 4.1, será apresentada as informações essenciais sobre o banco de dados de SNPs. Na Seção 4.2, resume uma introdução sobre o PPMI e apresenta as informações essenciais sobre os conjuntos de dados reais utilizados na análise de associação.

4.1 dbSNP

Existe atualmente um grande interesse na descoberta de SNP, uma vez que um catálogo denso de SNPs é esperado para facilitar grande escala de estudos na genética de associação, no mapeamento genético e na biologia evolutiva. Para atender a essa necessidade e complementar a base de dados de sequências genéticas GenBank, o Centro Nacional de Informação Biotecnológica, do inglês *National Center for Biotechnology Information* (NCBI), em colaboração com o Instituto de Pesquisa Nacional do Genoma Humano, do inglês *National Human Genome Research Institute* (NHGRI), criou, em setembro de 1998,Banco de dados de SNPs, do inglês *Single Nucleotide Polymorphism Database* (dbSNP) (http://www.ncbi.nlm.nih.gov/SNP) (SHERRY

et al., 2001). O dbSNP foi criado para atuar como uma única base de dados que contém toda a variação genética identificada, podendo incluir uma ampla gama de polimorfismos moleculares (substituições de SNPs, pequenas inserções/remoções, etc).

O banco de dados dbSNP tem servido como um repositório público central para a variação genética e, tal como acontece com todos os recursos de dados do NCBI, os dados dentro dbSNP estão disponíveis livremente e em uma variedade de formas (SHERRY et al., 2001). Embora a maioria das submissões sejam para o Homo sapiens, dbSNP tem submissões que incluem Mus musculus, Oryza sativa, e muitas outras espécies, e, em geral, o banco de dados pode aceitar informações de qualquer espécie e de qualquer parte de um genoma particular.

Nos estudos GWAS, os resultados das associações de SNPs a traços, geralmente são relatados por *Identificador de SNPs atribuído pelo NCBI* (RSID). Padronizados pelo NCBI (NCBI, 2014b,c), os identificadores rsID começam com o prefixo rs e são seguidos de uma sequência numérica, por exemplo rs41286661. O identificador rsID é utilizado por pesquisadores e bancos de dados para se referir a SNPs específicos. Nesse estudo, utilizamos a ferramenta de busca do banco de dados dbSNP para pesquisar as informações dos SNPs identificados e relacionar com facilidade os nossos resultados com os resultados dos estudos já realizados por pesquisadores da comunidade científica.

4.2 Parkinson's Progression Markers Initiative

Parkinson´s Progression Markers Iniative (PPMI) é um estudo clínico observacional para verificar marcadores de progressão na doença de Parkinson (PPMI, 2014). O projeto envolve pesquisadores e cerca de 35 centros clínicos nos EUA, Europa, Israel e Austrália, com o propósito de desenvolver uma avaliação detalhada de dados genômicos de grupos de interesse. O estudo PPMI tornou-se possível graças à colaboração dos pesquisadores, dos participantes, parceiros de financiamento e ao patrocínio da Fundação Michael J. Fox para Pesquisa de Parkinson (MJFF).

O estudo foi projetado para estabelecer um conjunto abrangente de imagens e de dados de bioamostra usados para definir biomarcadores da progressão DP. Um biomarcador pode ser qualquer característica física mensurável associada com a presença da doença (marcador de diagnóstico), ou qualquer característica que muda ao longo do tempo de uma forma que pode

ser ligada à progressão da doença (marcador de progressão) (PPMI, 2014). O objetivo final é a definição destes biomarcadores a partir da avaliação exaustiva do grupo de interesse significativo usando imagens avançadas, amostragens biológicas e avaliações clínicas e comportamentais. Para que os biomarcadores de progressão da doença de Parkinson identificados possam ser usados em estudos terapêuticos.

Os dados genéticos foram obtidos utilizando o equipamento *Illumina Immunochip* em 523 amostras de DNA extraídas de sangue coletado de acordo com o *PPMI Research Biomarkers Laboratory Manual*. (Para mais detalhes, ver (HERNANDEZ, 2011)). O *Immunochip* foi *Illumina Infinium iSelect HD* (ILLUMINA, 2014), que é capaz de obter informações de até 196.524 polimorfismos (718 pequenas inserções/remoções, 195.806 SNPs).

Todos os indivíduos incluídos no estudo foram recrutados a partir de cerca de 24 centros clínicos, e deram consentimento informado. Os dados e amostras adquiridas de participantes do estudo permitiram o desenvolvimento de um banco de dados de Parkinson abrangente e biorepositório, que está atualmente disponível para a comunidade científica realizar investigações. Os grupos de interesse estudados pelo PPMI são os indivíduos com DP, os indivíduos saudáveis, indivíduos com SWEDDs, os indivíduos com doença de Parkinson Prodromal, entre outros grupos.

Os dados incluídos na nossa avaliação tiveram um total de 195.806 SNPs de 523 participantes. Nesse trabalho todos os dados genéticos foram obtidos do portal do próprio PPMI, em 30 de janeiro de 2014. Utilizamos três grupos de indivíduos do estudo PPMI:

- **Indivíduos com DP (doença Parkinson)**, grupo com 325 indivíduos homem ou mulher com 30 anos ou mais de idade, com diagnóstico de DP por dois anos ou menos, e que não estão tomando medicamentos para DP;
- **Indivíduos do grupo controle ou saudáveis**, grupo com 152 indivíduos sem DP com 30 anos ou mais, e que não tem parentes de sangue ou de primeiro grau com DP;
- **Indivíduos com exames sem evidência de um déficit dopaminérgico (SWEDD)**, grupo com 46 indivíduos homem ou mulher com 60 anos ou mais de idade, diagnosticados com DP em que o teste de diagnóstico DaTscans não mostram evidências de

um déficit dopaminérgico, mas apresentam mutação nos genes LRRK2 e SCNA.

Segundo o portal do PPMI (PPMI, 2014), há 423 indivíduos do grupo DP, 196 indivíduos do grupo controle e 64 indivíduos do grupo SWEDD, mas devido a indisponibilidade do arquivo contendo os rsIDs padrão, optamos pelo Immunochip SNP Data. No Immunochip SNP Data, há 325 indivíduos do grupo DP, 152 indivíduos do grupo controle e 46 indivíduos do grupo SWEDD, totalizando 523 indivíduos, sendo 343 do gênero masculino e 180 do gênero feminino.

Para o controle de qualidade dos dados, foram adicionados ao teste de associação caso-controle filtros de controle de qualidade para garantir maior precisão nos resultados. Os SNPs de todos os pacientes considerados foram removidos de acordo com os seguintes critérios de controle de qualidade: (1) Frequência do menor alelo, do inglês *Minor allele frequence* (MAF) < 0,01; (2) Equilíbrio de Hardy-Weinberg, do inglês *Hardy-Weinberg equilibrium* (HWE) p <= 0,001; (3) *Missing rate per SNP* (geno) > 0,1; e (4) *Missing rate per person* (mind) > 0.1. Nesse trabalho, o teste de associação caso-controle foi conduzido em dois grupos:

- O grupo de associação da doença de Parkinson com 477 indivíduos (325 indivíduos com DP e 152 indivíduos controle).
- O grupo de associação da síndrome SWEDD com 198 indivíduos (46 indivíduos com DP e 152 indivíduos controle).

138.652 SNPs passaram no controle de qualidade do grupo de associação da DP e 141.411 SNPs passaram no controle de qualidade do grupo de associação SWEDD. Em ambos os grupos o total de indivíduos após o controle de qualidade permaneceu o mesmo.

No gráfico (Figura 4.1) é apresentada a distribuição das três populações de indivíduos (Saudável, DP e SWEDDs) estudadas levando em consideração a idade. Nos três grupos estudados, a menor idade entre os indivíduos é de 30 anos e a maior é de 83 anos. Em relação as duas populações caso, o gráfico aponta que a DP é muito frequente nos indivíduos com idade entre 60 e 70 anos, e observa-se que a síndrome SWEDDs teve maior frequência nos indivíduos entre 64 e 66 anos de idade. Observa-se a partir do gráfico que a frequência de idade nos indivíduos da DP está dentro da média de idade de diagnóstico (75 anos). Pode-se observar também a mesma frequência na síndrome SWEDDs, sugerindo uma característica similar a DP.

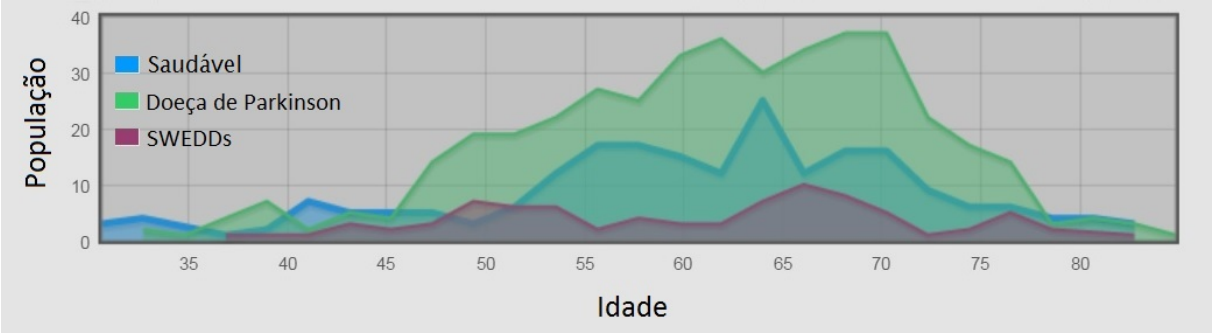

Figura 4.1: Distribuição das três populações de indivíduos por idade. Fonte: Modificada de http://www.ppmi-info.org/

O gráfico (Figura 4.2) apresenta a distribuição das três populações de indivíduos (Saudável, DP e SWEDDs) estudadas levando em consideração o gênero. Em relação às duas populações caso, o gráfico aponta que a DP e a síndrome SWEDDs são mais frequentes nos indivíduos do sexo masculino. Um aspecto percebido a partir do gráfico é a taxa de incidência da DP maior em homens, o que ocorre comumente nesta doença. Observa-se a mesma incidência na síndrome SWEDDs, sugerindo mais uma característica similar a DP.

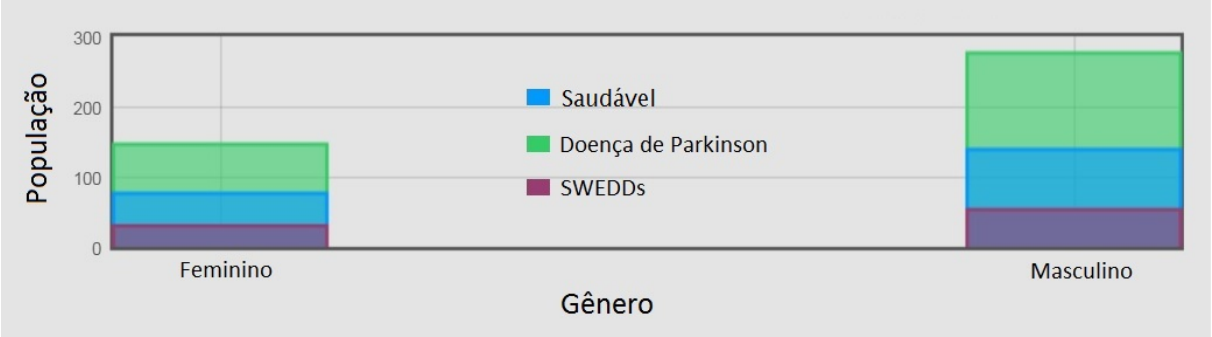

Figura 4.2: Distribuição das três populações de indivíduos pelo gênero. Fonte: Modificada de http://www.ppmi-info.org/

4.3 Considerações Finais

Este capítulo apresentou brevemente o banco de dados de SNPs e os conjuntos de dados reais que foram utilizados neste trabalho. Em resumo, foram descritas as informações sobre o banco de dados de SNPs e o padrão de identificação de SNPs. Também foi apresentada uma introdução sobre o PPMI, as informações sobre os dados genéticos obtidos e descrevemos o controle de qualidade adicionado a estes dados.

5

Análises e Resultados

Nesse trabalho, nós usamos o método PLINK para executar GWASs. Nos concentramos na análise de associação no contexto de base populacional. No método básico para associação da doença envolvendo estudo de caso-controle, foram adicionados filtros de controle para garantir maior precisão nos resultados. Os dados genéticos usados nos nossos testes foram fornecidos pelo Consórcio PPMI. Em particular, nós utilizamos três grupos de indivíduos do estudo PPMI: indivíduos saudáveis, indivíduos com DP e indivíduos com SWEDDs. O teste de associação caso-controle foi conduzido em dois grupos, no grupo de associação da DP e no grupo de associação da síndrome SWEDD. Dessa forma, neste capítulo serão descritos os testes de análise de associação realizados e os resultados obtidos.

O capítulo está organizado da seguinte forma: Na Seção 5.1 serão descritos os testes realizados e apresentados os resultados obtidos no grupo de associação da DP. Na Seção 5.2 serão descritos os testes realizados e apresentados os resultados obtidos no grupo de associação da síndrome SWEDD.

5.1 Doença de Parkinson

O grupo de associação da DP, incluiu 477 indivíduos (325 casos e 152 controles). Nesse conjunto de dados genéticos foi usado o método PLINK para executar o teste de associação caso-controle e os filtros de controle. Após a execução do teste, os resultados foram classificados

por *p*-values, para identificar quais SNPs são mais fortemente associados com a DP. Os 58 SNPs selecionados com *p*-values de alta significância de associação são apresentados na Tabela 5.1. Nós notamos que todos eles têm *p*-values de 8.043e-05 ou melhor. Destes, oito SNPs foram encontrados na literatura científica como estando associados significativamente com o parkinsonismo, e outros três SNPs encontrados foram associados a outras doenças neurodegenerativas (Tabela 5.2). Nós notamos também que cinco SNPs têm valores OR maiores que 1, afirmando a associação dos SNPs rs41286661, rs3822019, rs2290402 e rs56039006 com a DP, e identificando uma nova associação significativa com a DP, SNP rs1397596 no gene OR10A6 (Tabela 5.1, Tabela 5.2).

Chr	Gene	SNP	Alelles	CHISQ	*p*-value	OR
2	DYTN	rs16838587	T/C	24.34	8.057e-07	0.03449
2	DYTN	rs6750799	A/C	24.18	8.753e-07	0.0347
2	DYTN	rs16838593	C/T	22.18	2.487e-06	0.03749
6	EYA4	rs212805	A/G	20.66	5.489e-06	0.3074
2	DYTN	rs13429846	A/G	20.02	7.675e-06	0.04104
4	DGKQ	rs41286661	T/C	19.52	9.971e-06	2.534
2	DYTN	rs6760991	C/T	19.48	1.015e-05	0.1208
23		rs7054078	T/G	19.45	1.032e-05	0.2034
3		rs1512525	T/C	18.98	1.324e-05	0.0751
11	DLG2	rs2200204	T/C	18.83	1.427e-05	0.4185
10		rs2653508	G/A	18.14	2.052e-05	0.1044
4	TMEM175	rs3822019	T/C	18.06	2.144e-05	2.392
1	GLIS1	rs2478979	T/C	17.81	2.439e-05	0.1283
2	LTBP1	rs150655	G/C	17.81	2.439e-05	0.1283
10		rs1199071	T/C	17.81	2.439e-05	0.1283
17	BCAS3	rs16944668	C/T	17.81	2.443e-05	0.04544
4	TMEM175	rs2290402	T/C	17.8	2.456e-05	2.376
17	BCAS3	rs7218256	C/T	17.75	2.521e-05	0.04558
5		rs458006	T/G	17.73	2.549e-05	0.4803
1	SPSB1	rs1294054	A/C	17.58	2.757e-05	0.2501
6	EYA4	rs211619	A/G	17.51	2.862e-05	0.3259
10		rs1769007	C/T	17.51	2.864e-05	0.1494
10	IPMK	rs2790243	C/T	17.51	2.864e-05	0.1494
10	IPMK	rs2790232	T/C	17.43	2.976e-05	0.1498
11		rs10770163	A/G	17.39	3.051e-05	0.1307
10	CISD1	rs2225987	G/A	17	3.747e-05	0.1527
3	LPP	rs3732911	T/G	16.57	4.692e-05	0.3744
19	CRTC1	rs12973480	A/C	16.5	4.873e-05	0.4023
2	AL121656.5	rs150698	T/C	16.4	5.13e-05	0.1121
3		rs17078685	C/T	16.4	5.13e-05	0.1121
1	PCNXL2	rs2477858	G/A	16.32	5.354e-05	0.5691
4	DGKQ	rs56039006	G/C	16.31	5.372e-05	2.275
10		rs73286371	C/T	16.31	5.384e-05	0.3129
3		rs11929453	T/C	16.26	5.514e-05	0.1128
6	HLA-DQB1	rs9275425	A/C	16.25	5.561e-05	0.5422
⋮	⋮	⋮	⋮	⋮	⋮	⋮
11	OR10A6	rs1397596	G/A	15.63	7.693e-05	2.165
10		rs16851	C/T	15.61	7.765e-05	0.1606
10	IPMK	rs2590346	G/A	15.61	7.765e-05	0.1606
10	IPMK	rs2590347	C/G	15.61	7.765e-05	0.1606
10		rs1856565	G/A	15.55	8.043e-05	0.1611

Tabela 5.1: SNPs identificados no conjunto de dados DP/PPMI, ordenados por *p*-value de alta significância de associação.

Sítios de alta significância com associação confirmada para a doença de Parkinson						
Chr	Gene	SNP	Alelles	CHISQ	*p*-value	OR
4	DGKQ	rs41286661	T/C	19.52	9.971e-06	2.534
4	TMEM175	rs3822019	T/C	18.06	2.144e-05	2.392
1	GLIS1	rs2478979	T/C	17.81	2.439e-05	0.1283
4	TMEM175	rs2290402	T/C	17.8	2.456e-05	2.376
4	DGKQ	rs56039006	G/C	16.31	5.372e-05	2.275
6	HLA-DQB1	rs9275425	A/C	16.25	5.561e-05	0.5422
6	HLA-DQB1	rs9275428	G/A	15.4	8.692e-05	0.5539
6	HLA-DQB1	rs9275371	C/T	15.18	9.781e-05	0.555
Sítios de alta significância associados a doenças neurodegenerativas						
Chr	Gene	SNP	Alelles	CHISQ	*p*-value	OR
10	CISD1	rs2225987	G/A	17	3.747e-05	0.1527
11	OR10A6	rs1397596	G/A	15.63	7.693e-05	2.165
10	CISD1	rs35093457	T/C	15.47	8.367e-05	0.1412

Tabela 5.2: Sítios de alta significância recuperados na análise do conjunto de dados DP/PPMI mencionados na literatura.

Os oito SNPs recuperados e os quatro genes associados com DP são os seguintes:

- SNP rs41286661 com o sinal mais forte (OR = 2.534, P = 9.971e-06), e rs56039006, o qual também atingiu alta significância na análise geral do estudo, ambos no gene DGKQ (diacylglycerol kinase, theta 110kDa) (SIMON-SANCHEZ et al., 2011);
- SNPs rs3822019 e rs2290402 no gene TMEM175 (transmembrane protein 175) (DO et al., 2011; GENECARDS, 2014a);
- rs2478979 SNP no gene GLIS1 (GLIS family zinc finger 1) (SONG et al., 2012);
- rs9275425 SNPs, rs9275428 e rs9275371 no gene HLA-DQB1 (major histocompatibility complex, class II, DQ beta 1) (LAMPE et al., 2003; SAIKI et al., 2010); destes, o SNP rs9275371 SNP já foi pesquisado na literatura científica sendo associado com a artrite reumatóide (CHANDA et al., 2009), e o SNP rs9275390 associados à esclerose sistêmica (GORLOVA et al., 2011).

Os três SNPs recuperados e os dois genes associados a doenças neurodegenerativas são os seguintes:

- SNPs rs2225987 e rs35093457 no gene CISD1(CDGSH iron sulfur domain 1), associado às doenças síndrome de Wolfram 2 e síndrome de Wolfram (SHU; TSAI; CHI, 2003; CONLAN et al., 2009),

- SNP rs1397596 no gene OR10A6 (olfactory receptor, family 10, subfamily A, member 6), associado à doença neuronite (GENECARDS, 2014b).

Outros SNPs descobertos com susceptibilidade significativa com a DP e doenças neurodegenerativas não eram conhecidos anteriormente na literatura científica, no entanto mostram *p*-values de alta significância de associação (Tabela 5.1, Figura 5.1). Na análise foram identificados vinte e quatro SNPs relacionados com doze genes. Destes, os SNPs mais fortemente associados na análise foram:

- SNPs rs16838587, rs6750799 e rs16838593, que encontram-se no gene codificador da proteína DYTN (dystrotelin) localizado no cromossomo 2. O gene DYTN é expresso no sistema nervoso central (JIN et al., 2007).

- SNP rs 212805, que encontra-se no gene codificados da proteína EYA4 (transcriptional coactivator and phosphatase 4), localizado no cromossomo 6 (GENECARDS, 2014c).

Tendo em vista que esses SNPs e genes são novos na literatura, eles constituem alvos promissores para futuros estudos genômicos da DP.

O gráfico Manhattan plot do *P*-valor (Figura 5.1) dá uma visão geral dos resultados baseados em genes de testes para associação com DP. O eixo horizontal representa a posição genômica e o eixo vertical correspondente a -Log_{10} (*P*-valor), que representa a força de associação para DP, baseado no conjunto de dados DP/PPMI.

O pico no cromossomo 1 é correspondente ao gene GLIS1, e o *top hit* para esta região foi o SNP rs2478979 (OR = 0,1283, *p*-value = 2.439e-05). No cromossomo 4, o pico corresponde aos genes DGKQ e TMEM175, e o *top hit* para estas regiões foram rs41286661 (OR = 2.534, *p*-value = 9.971e-06) e rs3822019 (OR = 2,392, *p*-value = 2.144e-05), respectivamente. No cromossomo 6, o pico corresponde ao gene HLA-DQB1 e *top hit* para esta região é rs9275425 SNP (OR = 0,5422, *p*-value = 5.561e-05).

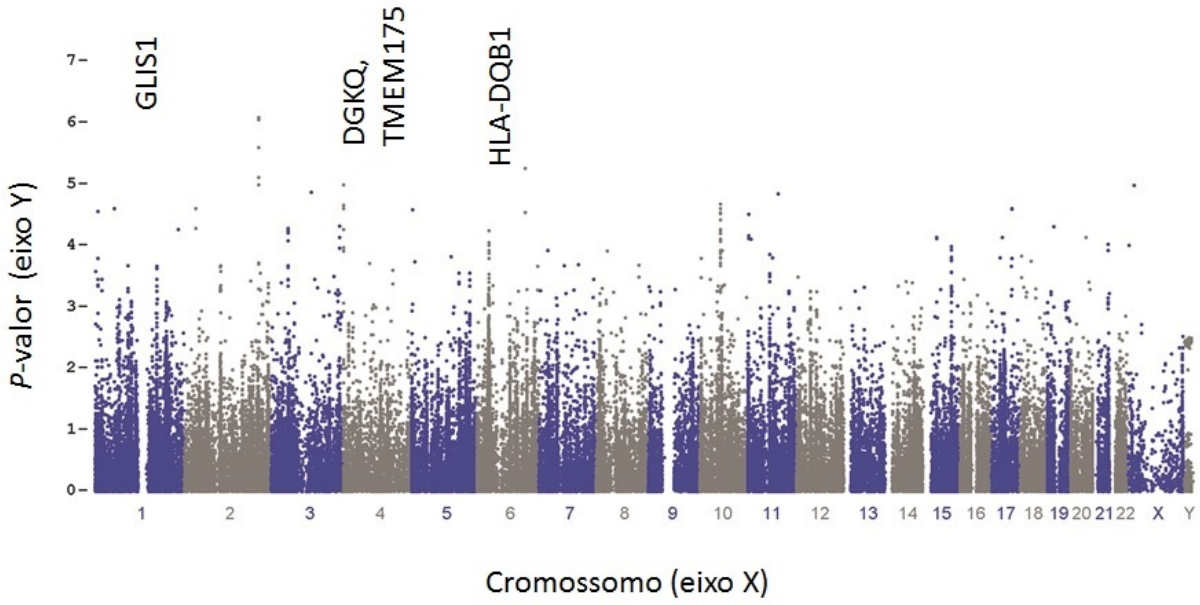

Figura 5.1: Resultados de GWAS no conjunto de dados DP/PPMI - O gráfico Manhattan plot mostra os resultados da análise de associação genética dos SNPs mais significativos, que ocorreram com maior frequência nos cromossomos 1, 2, 4, 6, 10, e 11.

Evidência de resultados consistentes é fornecida pelo gráfico quantil-quantil plot (*quantile-quantile plot* (Q-Q plot)) na (Figura 5.2), gerados através dos valores de qui-quadrado esperados em comparação com os valores de qui-quadrado observados.

O gráfico QQ plots apresenta a distribuição esperada do teste de associação estatístico (eixo X) através de milhões de SNPs em comparação com os valores observados (eixo-y). O desvio entre as linhas X e Y revela uma diferença consistente entre casos e controles, mostrando nitidamente as associações validadas. No gráfico Q-Q plot para DP os pontos se moveram curvando no final, representando o número de associações verdadeiras entre milhares de SNPs não associados.

5.2 SWEDDs

O grupo de associação da síndrome SWEDD, incluiu 198 indivíduos (46 casos e 152 controles). Nesse conjunto de dados genéticos foi usado o método PLINK para executar o teste de associação caso-controle e os filtros de controle. Após a axecução do teste, os resultados foram classificados por *p*-values, para identificar quais SNPs são mais fortemente associados com

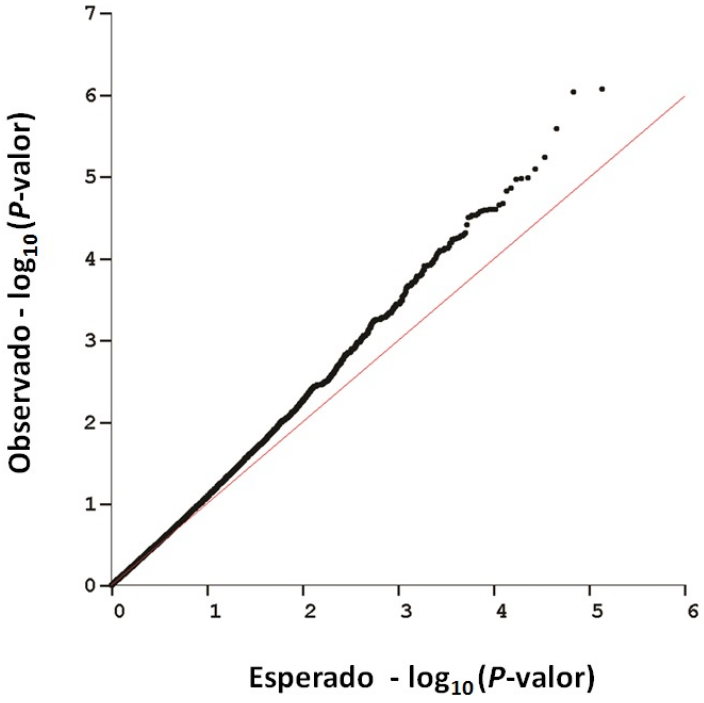

Figura 5.2: Resultados de GWAS no conjunto de dados DP/PPMI - O gráfico Q-Q plot mostra resultados consistentes de associações, onde o desvio curvado entre as linhas X (esperado) e Y (observado) representa o número de associações verdadeiras entre milhares de SNPs não associados.

a síndrome SWEDD. Os 30 SNPs selecionados com *p*-values de alta significância de associação são apresentados na Tabela 5.3. Nós notamos que todos eles têm *p*-values de 0.0001574 ou melhor. Dentre esses, cinco SNPs e três genes foram encontrados na literatura científica como estando associados significativamente com doenças neurodegenerativas, e outros sete SNPs encontrados foram associados a outros tipos de doenças (Tabela 5.4). Nós notamos também que todos os 30 SNPs têm valores OR maiores que 1, e como a força de uma associação é medida pela OR, podemos afirmar estes SNPs têm forte associação com a síndrome SWEDD (Tabela 5.3).

Os cinco SNPs recuperados e os três genes associados a doenças neurodegenerativas são os seguintes:

- SNP rs117438509 no gene CDH1(cadherin 1, type 1, E-cadherin (epithelial)), que atingiu alta significância na análise (OR = NA, P =1.219e-06) e associado com a doença de Alzheimer (SILVA et al., 2008; AULIA; TANG, 2006),
- SNPs rs5815615 e rs794423 no gene CLEC16A (C-type lectin domain family 16,

Chr	Gene	SNP	Alelles	CHISQ	*p*-value	OR
16	CDH1	rs117438509	G/A	23.55	1.219e-06	NA
1	LOC100506023	rs61828279	A/C	20.35	6.446e-06	6.984
6		rs961918	G/A	20.03	7.612e-06	3.056
16	CLEC16A	rs5815615	I/D	19.77	8.718e-06	7.4
16	CLEC16A	rs794423	A/C	19.37	1.077e-05	8.121
1		rs114797969	T/C	16.73	4.303e-05	NA
1	OLFML3	rs115198127	C/A	16.7	4.37e-05	2.875
10		rs11818028	A/C	16.64	4.509e-05	13.38
5	PAM	rs13172364	G/A	16.49	4.887e-05	7.293
5	PAM	rs59280720	A/T	16.49	4.887e-05	7.293
1	OLFML3	rs7364	A/C	16.43	5.049e-05	2.852
1	HIPK1	rs12127377	G/A	16.37	5.208e-05	2.841
1	HIPK1	rs11102707	A/G	16.37	5.208e-05	2.841
1	HIPK1	rs11102709	C/G	16.37	5.208e-05	2.841
1	HIPK1	rs1553450	G/A	16.37	5.208e-05	2.841
12	MYL6	rs78564374	A/G	15.95	6.517e-05	8.133
1	HIPK1	rs11102708	A/G	15.12	0.0001008	2.776
16		rs193758	G/C	14.88	0.0001145	6.208
5	PAM	rs1423136	A/G	14.57	0.0001347	3.892
1		rs878129	T/C	14.46	0.0001429	2.614
5	PAM	rs3909477	G/T	14.42	0.0001466	6.057
11	LOC101928962	rs117479043	G/C	14.42	0.0001466	6.057
5	PAM	rs35574814	A/G	14.42	0.0001466	6.057
5	PAM	rs17154763	G/A	14.42	0.0001466	6.057
2		rs115919155	A/G	14.42	0.0001466	6.057
5	PAM	rs7705636	A/T	14.42	0.0001466	6.057
5	PAM	rs13153062	A/G	14.42	0.0001466	6.057
5	PAM	rs7734438	T/A	14.42	0.0001466	6.057
2		rs2882970	C/T	14.36	0.000151	2.554
5	PAM	rs6872846	C/T	14.28	0.0001574	6.016

Tabela 5.3: SNPs identificados no conjunto de dados SWEDD/PPMI, ordenados por *p*-value de alta significância de associação.

member A), associado com as doenças de deficiência na Cadeia Alfa do Receptor IL-7 (interleucina-7) e esclerose múltipla (NISCHWITZ et al., 2011),

- SNPs rs13172364 e rs59280720 no gene PAM (peptidylglycine alpha-amidating monooxygenase), associado com a síndrome de pós-poliomielite e a doença de Menkes (STEVESON et al., 2003).

Sítios de alta significância associados a doenças neurodegenerativas						
Chr	Gene	SNP	Alelles	CHISQ	*p*-value	OR
16	CDH1	rs117438509	G/A	23.55	1.219e-06	NA
16	CLEC16A	rs5815615	I/D	19.77	8.718e-06	7.4
16	CLEC16A	rs794423	A/C	19.37	1.077e-05	8.121
5	PAM	rs13172364	G/A	16.49	4.887e-05	7.293
5	PAM	rs59280720	A/T	16.49	4.887e-05	7.293
Sítios de alta significância associados a outras doenças						
Chr	Gene	SNP	Alelles	CHISQ	*p*-value	OR
1	OLFML3	rs115198127	C/A	16.7	4.37e-05	2.875
1	OLFML3	rs7364	A/C	16.43	5.049e-05	2.852
1	HIPK1	rs12127377	G/A	16.37	5.208e-05	2.841
1	HIPK1	rs11102707	A/G	16.37	5.208e-05	2.841
1	HIPK1	rs11102709	C/G	16.37	5.208e-05	2.841
1	HIPK1	rs1553450	G/A	16.37	5.208e-05	2.841
12	MYL6	rs78564374	A/G	15.95	6.517e-05	8.133

Tabela 5.4: Sítios de alta significância recuperados na análise do conjunto de dados SWEDD/PPMI mencionados na literatura.

Os sete SNPs recuperados e os três genes associados outros tipos de doenças são os seguintes:

- SNPs rs115198127 e rs7364 no gene OLFML3 (olfactomedin-like 3), associado com as doenças glaucoma e sinusite (GENECARDS, 2014d),
- SNP rs12127377, rs11102707, rs11102709 e rs1553450 no gene HIPK1 (homeodomain interacting protein kinase 1), associado com as doenças síndrome de down e hipóxia (KIMURA et al., 2007), e
- SNP rs78564374 no gene MYL6 (myosin, light chain 6, alkali, smooth muscle and non-muscle), associado com as doenças síndrome lynch e rabdomiossarcoma (GENECARDS, 2014e).

A Figura 5.3 mostra o gráfico de Manhattan plot of the *P*-valor para o conjunto de dados SWEDD/PPMI. O pico no cromossomo 5 corresponde ao gene CDH1, e o *top hit* para esta região é rs117438509 (OR = NA, *p*-value = 1.219e-06). No cromossomo 16, o pico corresponde aos genes CLEC16A e PAM, e os *top hit* para estas regiões são SNPs rs5815615 (OR = 7,4, *p*-value = 8.718e-06) e rs13172364 (OR = 7,293, *p*-valor 4.887e-05), respectivamente.

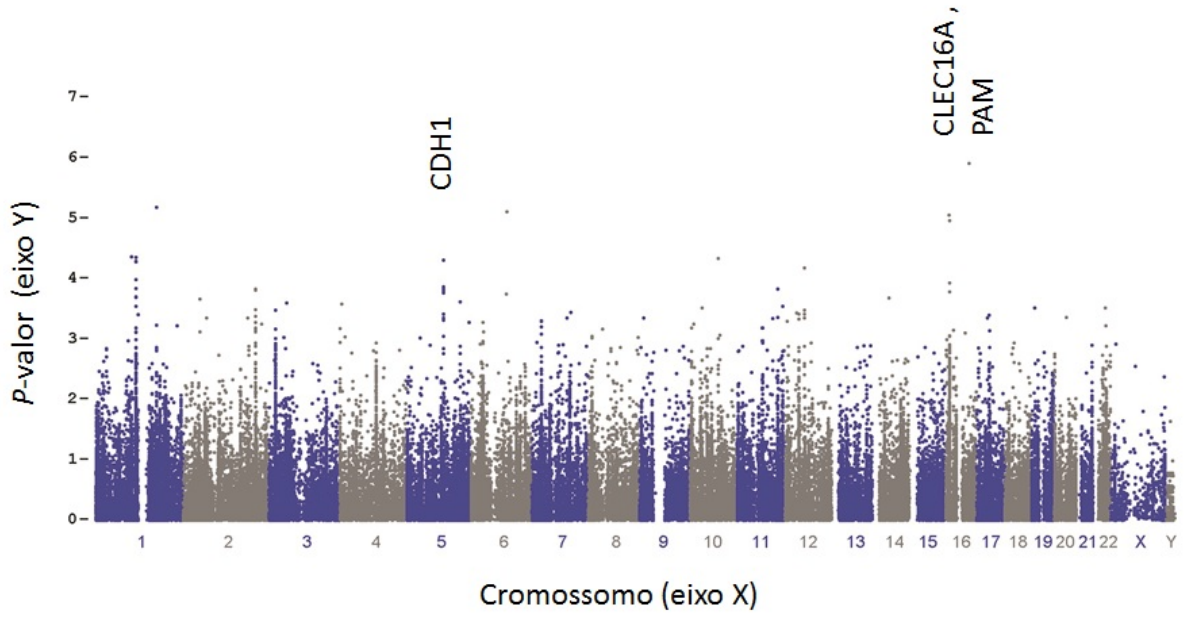

Figura 5.3: Resultados de GWAS no conjunto de dados SWEDDs/PPMI - O gráfico Manhattan plot mostra os resultados da análise de associação genética dos SNPs mais significativos, que ocorreram com maior frequência nos cromossomos 1, 6, e 16.

A Figura 5.4 mostra o gráfico Q-Q plot para SWEDD, onde o desvio entre as linhas X e Y revela uma diferença consistente entre casos e controles, mostrando nitidamente as associações validadas. Neste gráfico os pontos moveram-se levemente curvados no final, representando o pequeno número de associações verdadeiras entre milhares de SNPs não associados.

Uma observação muito importante é que nenhum dos SNPs selecionados com *p*-values de alta significância de associação no conjunto de dados DP (Tabela 5.1) está presente no conjunto de dados SWEDD. Por outro lado, os SNPs selecionados com *p*-values de alta significância de associação no conjunto de dados SWEDD (Tabela 5.3) também não estão presentes no conjunto de dados DP, o que é curioso, tendo em conta a estreita interseção dos sintomas clínicos destas duas doenças. Uma investigação mais profunda do efeito dessas mutações poderia lançar alguma luz sobre as causas dessas síndromes ou de algum sintoma específico.

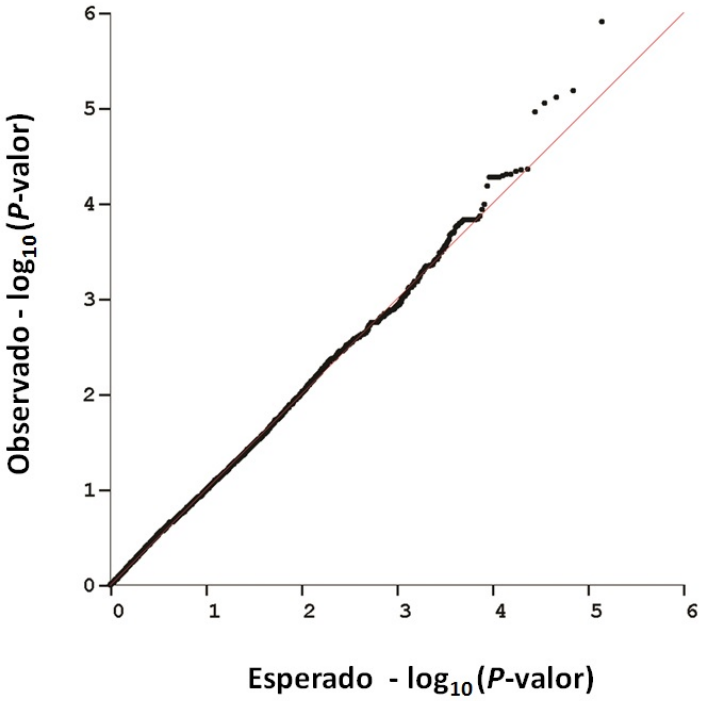

Figura 5.4: Resultados de GWAS no conjunto de dados SWEDDs/PPMI - O gráfico Q-Q plot mostra resultados consistentes de associações, onde o desvio curvado entre as linhas X (esperado) e Y (observado) representa o pequeno número de associações verdadeiras entre milhares de SNPs não associados.

5.3 Considerações Finais

Este capítulo apresentou os resultados obtidos na realização dos testes de análise de associação conduzido em dois grupos, no grupo de associação da DP e no grupo de associação da síndrome SWEDD. No primeiro grupo 58 SNPs foram recuperados com *p*-values de alta significância de associação. Destes, oito SNPs estão associados significativamente com o parkinsonismo, e três SNPs estão associados a outras doenças neurodegenerativas, e outros 47 SNPs descobertos, são novos na literatura. No segundo grupo 30 SNPs foram recuperados com *p*-values de alta significância de associação e com valores OR que sugerem forte associação com a síndrome SWEDD. Dentre esses, cinco SNPs estão associados significativamente com doenças neurodegenerativas e outros sete SNPs encontrados, estão associados a outros tipos de doenças. Assim, através dos estudos aqui apresentados, foi possível contribuir com cerca de 50 SNPs que constituem alvos promissores para futuros estudos genômicos da DP e da síndrome SWEDD.

6

Discussão e Conclusão

Nessa dissertação nós desenvolvemos um estudo sobre fatores genéticos que possam contribuir para o entendimento da ocorrência da DP e SWEDDs. Tendo como objetivo principal a análise de associação para, a partir de dados genéticos reais fornecidos pelo *PPMI Consortium*, procurar identificar SNPs que estão associados à DP e SWEDDs. Para isso, nós usamos o conjunto de ferramentas de análise de associação envolvendo estudo de caso-controle do método PLINK para executar GWASs.

Para a análise de associação nós utilizamos dados genótipos de três grupos de indivíduos do estudo PPMI: indivíduos saudáveis, indivíduos com DP e indivíduos com SWEDDs. Em nosso estudo, foram utilizados apenas os dados genéticos que passaram pelos critérios de qualidade relacionados com frequência alélica, *missing rates* e equilíbrio de Hardy-Weinberg para garantir maior precisão nos resultados, e foram considerados como relevantes somente os SNPs selecionados com estrito *p*-values de alta significância de associação. Para cada grupo analisado, O teste de associação caso-controle foi conduzido em dois grupos, no grupo de associação da DP (indivíduos com DP e indivíduos controle) e no grupo de associação da síndrome SWEDD (indivíduos com SWEDD e indivíduos controle).

No primeiro grupo nós identificamos 58 SNPs com *p*-values de alta significância de associação com a DP. Oito desses SNPs encontram-se em quatro genes (DGKQ, TMEM175, GLIS1 e HLA-DQB1), que estão confirmados por estudos anteriores na literatura científica como genes associados significativamente com a doença de Parkinson. Três outros SNPs foram ligados

a outras doenças neurodegenerativas, mas não diretamente a DP. E outros 47 SNPs descobertos, são novos na literatura. Nós identificamos também que cinco SNPs têm valores OR maiores que 1, afirmando a associação dos SNPs rs41286661, rs3822019, rs2290402 e rs56039006 com a DP, e identificando uma nova associação significativa com a DP, SNP rs1397596 no gene OR10A6.

No segundo grupo, 30 SNPs foram recuperados com *p*-values de alta significância de associação e com valores OR que sugerem forte associação com a síndrome SWEDD. Encontramos evidências empíricas na literatura associando cinco desses SNPs com doenças neurodegenerativas. Outros sete SNPs também são mencionados na literatura em relação a outros tipos de doenças, mas não com SWEDDs, e outros 18 SNPs descobertos são novos na literatura.

Os nossos resultados apontam para 65 novos SNPs que mostram uma evidência muito forte de associação com a PD e SWEDDs, os quais representam alvos promissores para futuros estudos genômicos e podem contribuir para o entendimento da ocorrência da DP e SWEDDs. Em particular, os resultados obtidos do grupo de associação da síndrome SWEDD apontaram para 30 novos SNPs com *p*-values de alta significância de associação e com valores OR que sugerem forte associação com a síndrome. Vale destacar que na literatura científica existem pouquíssimos estudos sobre a síndrome SWEDD, em específico os estudos de análise de associação são bastante escassos, sugerindo que estes SNPs deveriam ser mais explorados.

Comparando os conjuntos de SNPs recuperados com alta significância para DP com o conjunto de SNPs recuperado com alta significância para SWEDD, notamos que eles não apresentaram sobreposição. Verificou-se que nenhum desses SNPs selecionados foram comuns aos dois conjuntos de dados, embora não se possa afirmar que, se eles estivessem presentes em ambos os conjuntos, eles seriam selecionados. Numa visão mais ampla, através do gráfico Manhattan plot, nós também observamos que a propagação dos SNPs relevantes estão concentrados em diferentes cromossomos nos dois conjuntos de dados de doenças, que podem ou não estar associados à sobreposição de SNP nos conjuntos de dados. Essas observações indicam que uma investigação mais profunda dos SNPs no método Plink poderia lançar alguma luz sobre a estreita interseção dos sintomas clínicos destas duas doenças.

Como resultado deste trabalho de pesquisa, um artigo científico intitulado: “Confirmed and Novel Parkinsonism-related SNPs revealed from GWAS analysis” foi redigido e submetido

para publicação, e está em processo de revisão. O artigo foi submetido com o objetivo de exibir as análises e os resultados obtidos usando o método PLINK nos dados providos pelo PPMI.

Em trabalhos futuros, observamos que a função de análise de associação do método PLINK abrange também a regressão logística. A regressão logística testa as diferenças na frequência alélica entre dois grupos de indivíduos, possibilitando a condução de novas análises no conjunto de dados PPMI. Além da ferramenta PLINK o teste de regressão logística também pode ser realizado usando a ferramenta ProbABEL (AULCHENKO et al., 2007). Assim além de aplicar um novo teste ao conjunto de dados com a possibilidade de obter novos resultados, poderemos fazer uma análise comparativa destes resultados.

Referências

ALBERTS, B. et al. **Biologia molecular da célula**. [S.l.]: Artmed, 2010.

AULCHENKO, Y. S. et al. GenABEL: an r library for genome-wide association analysis. **Bioinformatics**, [S.l.], v.23, n.10, p.1294–1296, 2007.

AULIA, S.; TANG, B. L. Cdh1-APC/C, cyclin B-Cdc2, and Alzheimer's disease pathology. **Biochemical and Biophysical Research Communications**, [S.l.], v.339, n.1, p.1–6, 2006.

AYUSO, P. et al. An association study between Heme oxygenase-1 genetic variants and Parkinson's disease. **Frontiers in cellular neuroscience**, [S.l.], v.8, 2014.

BARBOSA, M. T. et al. Parkinsonism and Parkinson's disease in the elderly: a community-based survey in brazil (the bambuí study). **Movement Disorders**, [S.l.], v.21, n.6, p.800–808, 2006.

BEIGUELMAN, B.; EDITORA, S. Genética de populações humanas. **Ribeirão Preto: SBG**, [S.l.], p.483–488, 2008.

BOHLHALTER, S.; KAGI, G. Parkinsonism: heterogeneity of a common neurological syndrome. **Swiss Medical Weekly**, [S.l.], v.141, n.w13293, 2011.

BONIFATI, V. et al. Mutations in the DJ-1 gene associated with autosomal recessive early-onset Parkinsonism. **science**, [S.l.], v.299, n.5604, p.256–259, 2003.

CARDON, L. R.; BELL, J. I. Association study designs for complex diseases. **Nature Reviews Genetics**, [S.l.], v.2, n.2, p.91–99, 2001.

CHANDA, P. et al. A two-stage search strategy for detecting multiple loci associated with rheumatoid arthritis. **BMC Proceedings**, [S.l.], v.3, n.Suppl 7, p.S72, 2009.

CLARKE, G. M. et al. Basic statistical analysis in genetic case-control studies. **Nature protocols**, [S.l.], v.6, n.2, p.121–133, 2011.

CLOTE, P.; R.BACKOFEN. **Computational molecular biology**: an introduction. [S.l.]: John Wiley, 2000.

CONLAN, A. R. et al. Crystal Structure of Miner1: the redox-active 2fe-2s protein causative in Wolfram syndrome 2. **Journal of Molecular Biology**, [S.l.], v.392, n.1, p.143–153, 2009.

CONSORTIUM, I. H. G. S. et al. Finishing the euchromatic sequence of the human genome. **Nature**, [S.l.], v.431, n.7011, p.931–945, 2004.

COX, N. et al. **SCAN**: snp and cnv annotation database. 2014.

DISEASE FOUNDATION Parkinson's. **Statistics on Parkinson's**. 2015.

DO, C. B. et al. Web-Based Genome-Wide Association Study Identifies Two Novel Loci and a Substantial Genetic Component for Parkinson's Disease. **PLoS Genetics**, [S.l.], v.7, n.6, p.e1002141, 2011.

ESPER, C. D.; FACTOR, S. A. Failure of recognition of drug-induced Parkinsonism in the elderly. **Movement Disorders**, [S.l.], v.23, n.3, p.401–404, 2008.

GENECARDS. **GeneCards**: the human gene compendium - TMEM175 gene. 2014.

GENECARDS. **GeneCards**: the human gene compendium - OR10A6 gene. 2014.

GENECARDS. **GeneCards**: the human gene compendium - EYA4 gene. 2014.

GENECARDS. **GeneCards**: the human gene compendium - OLFML3 gene. 2014.

GENECARDS. **GeneCards**: the human gene compendium - MYL6 gene. 2014.

GORLOVA, O. et al. Identification of novel genetic markers associated with clinical phenotypes of systemic sclerosis through a genome-wide association strategy. **PLoS genetics**, [S.l.], v.7, n.7, p.e1002178, 2011.

HAMZA, T. H. et al. Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. **Nature Genetics**, [S.l.], v.42, n.9, p.781–785, 2010.

HAMZA, T. H. et al. Genome-wide gene-environment study identifies glutamate receptor gene GRIN2A as a Parkinson's disease modifier gene via interaction with coffee. **PLoS genetics**, [S.l.], v.7, n.8, p.e1002237, 2011.

HECKMAN, M. G. et al. The protective effect of LRRK2 p.R1398H on risk of Parkinson's disease is independent of MAPT and SNCA variants. **Neurobiology of Aging**, [S.l.], v.35, n.1, p.266.e5–266.e14, 2014.

HERNANDEZ, D. G. **Immunochip genotyping on DNA samples from PPMI**. [S.l.]: National Institute on Aging, Laboratory of Neurogenetics, 2011.

HILL-BURNS, E. et al. Identification of a novel Parkinson's disease locus via stratified genome-wide association study. **BMC Genomics**, [S.l.], v.15, n.1, p.118, 2014.

HILL-BURNS, E. M. et al. A genetic basis for the variable effect of smoking/nicotine on Parkinson's disease. **The Pharmacogenomics Journal**, [S.l.], v.13, n.6, p.530–537, 2013.

ILLUMINA. **Custom Genotyping Arrays**. 2014.

IOANNIDIS, J. P. et al. Replication validity of genetic association studies. **Nature genetics**, [S.l.], v.29, n.3, p.306–309, 2001.

JIN, H. et al. The dystrotelin, dystrophin and dystrobrevin superfamily: new paralogues and old isoforms. **BMC Genomics**, [S.l.], v.8, n.1, 2007.

KARCHIN, R. Next generation tools for the annotation of human SNPs. **Briefings in bioinformatics**, [S.l.], v.10, n.1, p.35–52, 2009.

KIMURA, R. et al. The DYRK1A gene, encoded in chromosome 21 Down syndrome critical region, bridges between β-amyloid production and tau phosphorylation in Alzheimer disease. **Human Molecular Genetics**, [S.l.], v.16, n.1, p.15–23, 2007.

KITADA, T. et al. Mutations in the parkin gene cause autosomal recessive juvenile Parkinsonism. **Nature**, [S.l.], v.392, n.6676, p.605–608, 1998.

KUMAR, V. **Robbins and Cotran Patologia-Bases Patológicas das Doenças**. [S.l.]: Elsevier Brasil, 2011.

LAMPE, J. et al. HLA Typing and Parkinson's Disease. **European Neurology**, [S.l.], v.50, n.2, p.64–68, 2003.

LEARN.GENETICS. **Making SNPs Make Sense**. 2014.

MEDICINANET. **Distúrbios do movimento**. 2014.

MENEZES, A. M. Noções básicas de epidemiologia. **Epidemiologia das doenças respiratórias**, [S.l.], v.1, 2001.

MI, H. et al. PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. **Nucleic acids research**, [S.l.], v.35, n.suppl 1, p.D247–D252, 2007.

MIAN, O. S. et al. Gait in SWEDDs patients: comparison with Parkinson's disease patients and healthy controls. **Movement Disorders**, [S.l.], v.26, n.7, p.1266–1273, 2011.

MOUNT, D. W. Sequence and genome analysis. **Bioinformatics: Cold Spring Harbour Laboratory Press: Cold Spring Harbour**, [S.l.], v.2, 2004.

NCBI. **Home - Gene - NCBI**. 2014.

NCBI. **National Center for Biotechnology Information**: clustered refsnps (rs) and other data computed in house. 2014.

NCBI. **The NCBI Handbook**: chapter 5 - the single nucleotide polymorphism database (dbsnp) of nucleotide sequence variation. 2014.

NIH/NLM. **National Library of Medicine, Genetics Home Reference**: single nucleotide polymorphisms (snps). 2014.

NINA, S. A. **Parkinsonismo**. 2014.

NISCHWITZ, S. et al. More CLEC16A gene variants associated with multiple sclerosis. **Acta Neurologica Scandinavica**, [S.l.], v.123, n.6, p.400–406, 2011.

OLAZAR, M. R. R. **UMA METODOLOGIA PARA A DESCOBERTA DE MARCADORES GENÉTICOS EM ESTUDOS DE ASSOCIAÇÂO**. 2013. Tese (Doutorado em Ciência da Computação) — Universidade Federal do Rio de Janeiro.

PAISAN-RUIZ, C. et al. Cloning of the gene containing mutations that cause PARK8-linked Parkinson's disease. **Neuron**, [S.l.], v.44, n.4, p.595–600, 2004.

PANKRATZ, N. et al. Meta-analysis of Parkinson's Disease: identification of a novel locus, rit2. **Annals of neurology**, [S.l.], v.71, n.3, p.370–384, 2012.

PDGENE. **PDGene database**. `http://www.pdgene.org/`.

PLINK. **Association analysis**. 2014.

PLINK. **Inclusion thresholds**. 2014.

POLYMEROPOULOS, M. H. et al. Mutation in the α-synuclein gene identified in families with Parkinson's disease. **science**, [S.l.], v.276, n.5321, p.2045–2047, 1997.

PPMI. **Parkinson's Progression Markers Initiative**. 2014.

PURCELL, S. **PLINK 1.07**. 2014.

PURCELL, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. **The American Journal of Human Genetics**, [S.l.], v.81, n.3, p.559 – 575, 2007.

RAMENSKY, V.; BORK, P.; SUNYAEV, S. Human non-synonymous SNPs: server and survey. **Nucleic acids research**, [S.l.], v.30, n.17, p.3894–3900, 2002.

RAMIREZ, A. et al. Hereditary Parkinsonism with dementia is caused by mutations in ATP13A2, encoding a lysosomal type 5 P-type ATPase. **Nature genetics**, [S.l.], v.38, n.10, p.1184–1191, 2006.

ROBINSON, M. O. **Parkinson's Disease Gene Networks**. [S.l.]: Molquant, Inc., 2014.

SAIKI, M. et al. Association of the human leucocyte antigen region with susceptibility to Parkinson's disease. **Journal of Neurology, Neurosurgery & Psychiatry**, [S.l.], v.81, n.8, p.890–891, 2010.

SATAKE, W. et al. Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. **Nature genetics**, [S.l.], v.41, n.12, p.1303–1307, 2009.

SCHWINGENSCHUH, P. et al. Distinguishing SWEDDs patients with asymmetric resting tremor from Parkinson's disease: a clinical and electrophysiological study. **Movement disorders**, [S.l.], v.25, n.5, p.560–569, 2010.

SETUBAL, J. C.; MEIDANIS, J. **Introduction to computational molecular biology**. [S.l.]: PWS Pub., 1997.

SHERRY, S. T. et al. dbSNP: the ncbi database of genetic variation. **Nucleic acids research**, [S.l.], v.29, n.1, p.308–311, 2001.

SHU, S.-G.; TSAI, C.-R.; CHI, C.-S. Wolfram syndrome: phenotype and novel mutation in two taiwanese siblings. **Journal of the Formosan Medical Association**, [S.l.], v.102, n.11, p.808–811, 2003.

SILVA, P. N. O. et al. Promoter Methylation Analysis of SIRT3, SMARCA5, HTERT and CDH1 Genes in Aging and Alzheimer's Disease. **Journal of Alzheimer's Disease**, [S.l.], v.13, n.2, p.173–176, 2008.

SIMON-SANCHEZ, J. et al. Genome-wide association study reveals genetic risk underlying Parkinson's disease. **Nature genetics**, [S.l.], v.41, n.12, p.1308–1312, 2009.

SIMON-SANCHEZ, J. et al. Genome-wide association study confirms extant PD risk loci among the Dutch. **European Journal Human Genetics**, [S.l.], v.19, n.6, p.655–661, 2011.

SINGLETON, A. B. et al. α-Synuclein locus triplication causes Parkinson's disease. **Science**, [S.l.], v.302, n.5646, p.841–841, 2003.

SONG, W. et al. GLIS1 rs797906: an increased risk factor for late-onset Parkinson's disease in the han chinese population. **European Neurology**, [S.l.], v.68, n.2, p.89–92, 2012.

SPATARO, N. et al. Mendelian genes for Parkinson's disease contribute to the sporadic forms of the disease. **Human molecular genetics**, [S.l.], p.ddu616, 2014.

STEVESON, T. C. et al. Menkes Protein Contributes to the Function of Peptidylglycine α-Amidating Monooxygenase. **Endocrinology**, [S.l.], v.144, n.1, p.188–200, 2003.

SUNYAEV, S. et al. Prediction of nonsynonymous single nucleotide polymorphisms in human disease-associated genes. **Journal of molecular medicine**, [S.l.], v.77, n.11, p.754–760, 1999.

THOMPSON, M. W.; MCINNES, R. R.; WILLARD, H. F. **Thompson e Thompson genética médica**. [S.l.]: Guanabara Koogan, 1993.

VALENTE, E. M. et al. Hereditary early-onset Parkinson's disease caused by mutations in PINK1. **Science**, [S.l.], v.304, n.5674, p.1158–1160, 2004.

VILARINO-GUELL, C. et al. VPS35 mutations in Parkinson's disease. **The American Journal of Human Genetics**, [S.l.], v.89, n.1, p.162–167, 2011.

WALTER; DIVISION, E. H. I. B. **GWAS Tutorial with PLINK and Haploview**: understanding the statistics. 2014.

WANG, P. et al. SNP Function Portal: a web database for exploring the function implication of snp alleles. **Bioinformatics**, [S.l.], v.22, n.14, p.e523–e529, 2006.

WANG, Z.; MOULT, J. SNPs, protein structure, and disease. **Human mutation**, [S.l.], v.17, n.4, p.263–270, 2001.

WU, M. C. et al. Powerful SNP-set analysis for case-control genome-wide association studies. **The American Journal of Human Genetics**, [S.l.], v.86, n.6, p.929–942, 2010.

ZIMPRICH, A. et al. Mutations in LRRK2 cause autosomal-dominant Parkinsonism with pleomorphic pathology. **Neuron**, [S.l.], v.44, n.4, p.601–607, 2004.

Apêndice

A

Apêndice

Arquivos de log gerados na execução do comando PLINK

Arquivo de log do grupo de associação da DP

PLINK! | v1.07 | 10/Aug/2009

(C) 2009 Shaun Purcell, GNU General Public License, v2

For documentation, citation & bug-report instructions:

http://pngu.mgh.harvard.edu/purcell/plink/

Skipping web check... [–noweb]

Writing this text to log file [../Dados/PPMI/Results/PARK-ALL/plink.log]

Analysis started: Wed Jun 11 14:44:46 2014

Options in effect:

–script plink_script.txt

–keep ../Dados/PPMI/Data/populations_healthy_pd.txt

–out ../Dados/PPMI/Results/PARK-ALL/plink

–bfile ../Dados/PPMI/Data/IMMUNO

–pheno ../Dados/PPMI/Phenotypes/pheno_parkinson.txt

–hwe 0.001

–geno 0.1

–mind 0.1

–maf 0.01

–assoc

–noweb

–silent

** For gPLINK compatibility, do not use '.' in –out **

Reading map (extended format) from [../Dados/PPMI/Data/IMMUNO.bim]

196524 markers to be included from [../Dados/PPMI/Data/IMMUNO.bim]

Reading pedigree information from [../Dados/PPMI/Data/IMMUNO.fam]

523 individuals read from [../Dados/PPMI/Data/IMMUNO.fam]

0 individuals with nonmissing phenotypes

Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)

Missing phenotype value is also -9

0 cases, 0 controls and 523 missing

343 males, 180 females, and 0 of unspecified sex

Reading genotype bitfile from [../Dados/PPMI/Data/IMMUNO.bed]

Detected that binary PED file is v1.00 SNP-major mode

Reading alternate phenotype from [../Dados/PPMI/Phenotypes/pheno_parkinson.txt]

523 individuals with non-missing alternate phenotype

Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)

Missing phenotype value is also -9

371 cases, 152 controls and 0 missing

Reading individuals to keep [../PPMI/Data/populations_healthy_pd.txt] ... 477 read

46 individuals removed with –keep option

Before frequency and genotyping pruning, there are 196524 SNPs

477 founders and 0 non-founders found

0 of 477 individuals removed for low genotyping (MIND > 0.1)

61538 heterozygous haploid genotypes; set to missing

Writing list of heterozygous haploid genotypes to [../PPMI/Results/PARK-ALL/plink.hh]

1043 markers to be excluded based on HWE test (p <= 0.001)

1145 markers failed HWE test in cases

1043 markers failed HWE test in controls

Total genotyping rate in remaining individuals is 0.956461

10936 SNPs failed missingness test (GENO > 0.1)

53861 SNPs failed frequency test (MAF < 0.01)

After frequency and genotyping pruning, there are 138652 SNPs

After filtering, 325 cases, 152 controls and 0 missing

After filtering, 314 males, 163 females, and 0 of unspecified sex

Writing main association results to [../Dados/PPMI/Results/PARK-ALL/plink.assoc]

Analysis finished: Wed Jun 11 14:44:55 2014

Arquivo de log do grupo de associação da síndrome SWEDD

PLINK! | v1.07 | 10/Aug/2009

(C) 2009 Shaun Purcell, GNU General Public License, v2

For documentation, citation & bug-report instructions:

http://pngu.mgh.harvard.edu/purcell/plink/

Skipping web check... [–noweb]

Writing this text to log file [../Dados/PPMI/Results/SWEDD-ALL/plink.log]

Analysis started: Wed Jun 11 14:45:21 2014

Options in effect:

–script plink_script.txt

–keep ../Dados/PPMI/Data/populations_healthy_swedd.txt

–out ../Dados/PPMI/Results/SWEDD-ALL/plink

–bfile ../Dados/PPMI/Data/IMMUNO

–pheno ../Dados/PPMI/Phenotypes/pheno_parkinson.txt

–hwe 0.001

–geno 0.1

–mind 0.1

–maf 0.01

–assoc

–noweb

–silent

** For gPLINK compatibility, do not use '.' in –out **

Reading map (extended format) from [../Dados/PPMI/Data/IMMUNO.bim]

196524 markers to be included from [../Dados/PPMI/Data/IMMUNO.bim]

Reading pedigree information from [../Dados/PPMI/Data/IMMUNO.fam]

523 individuals read from [../Dados/PPMI/Data/IMMUNO.fam]

0 individuals with nonmissing phenotypes

Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)

Missing phenotype value is also -9

0 cases, 0 controls and 523 missing

343 males, 180 females, and 0 of unspecified sex

Reading genotype bitfile from [../Dados/PPMI/Data/IMMUNO.bed]

Detected that binary PED file is v1.00 SNP-major mode

Reading alternate phenotype from [../Dados/PPMI/Phenotypes/pheno_parkinson.txt]

523 individuals with non-missing alternate phenotype

Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)

Missing phenotype value is also -9

371 cases, 152 controls and 0 missing

Reading individuals to keep [../PPMI/Data/populations_healthy_swedd.txt] ... 198 read

325 individuals removed with –keep option

Before frequency and genotyping pruning, there are 196524 SNPs

198 founders and 0 non-founders found

0 of 198 individuals removed for low genotyping (MIND > 0.1)

25514 heterozygous haploid genotypes; set to missing

Writing list of heterozygous haploid genotypes to [../PPMI/Results/SWEDD-ALL/plink.hh]

1043 markers to be excluded based on HWE test (p <= 0.001)

120 markers failed HWE test in cases

1043 markers failed HWE test in controls

Total genotyping rate in remaining individuals is 0.957167

10786 SNPs failed missingness test (GENO > 0.1)

51164 SNPs failed frequency test (MAF < 0.01)

After frequency and genotyping pruning, there are 141411 SNPs

After filtering, 46 cases, 152 controls and 0 missing

After filtering, 130 males, 68 females, and 0 of unspecified sex

Writing main association results to [../Dados/PPMI/Results/SWEDD-ALL/plink.assoc]

Analysis finished: Wed Jun 11 14:45:26 2014