UNIVERSIDADE FEDERAL DE MINAS GERAIS
Luciana Werneck Zuccherato

Estrutura populacional e diversidade de variações em número de cópias (CNVs) de genes do sistema imune em populações nativas da América do Sul

BELO HORIZONTE
2012

Luciana Werneck Zuccherato

Estrutura populacional e diversidade de variações em número de cópias (CNVs) de genes do sistema imune em populações nativas da América do Sul

Tese apresentada ao Programa de Pós-graduação em Genética do Departamento de Biologia Geral do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutor em Genética
Orientador: Prof. Dr. Eduardo Martín Tarazona-Santos

Belo Horizonte
Departamento de Biologia Geral
Instituto de Ciências Biológicas
Universidade Federal de Minas Gerais
Novembro de 2012

Agradecimentos

Gostaria prestar meus sinceros agradecimentos:

Ao professor Eduardo Tarazona pela orientação e atenção durante todas as conquistas científicas desde o meu ingresso no LDGH.

Ao professor Edward Hollox por ter me aceitado carinhosamente em seu laboratório e por todos os votos de confiança. Aos amigos ingleses, escoceses, australianos, italianos, alemães, gregos, árabes e brasileiros que me ensinaram que viver fora do país é uma experiência enriquecedora.

Aos colegas e amigos presentes e ausentes do LDGH: Maria Clara, Wagner, Maíra, Ferdi, Camila, Andrea, Nelson, Rodrigo, Moara, Juliana, Roxana, Fernanda, Marília, Giordano, Raquel, Laélia, Silvana Matheus e Latife pelo carinho e companheirismo.

À Cláudia Benedetto, por ter me transformado de técnica em cientista, pelo exemplo de dedicação e carinho pela Ciência e pelo mundo.

Aos professores da Pós-graduação em Genética pelos importantes ensinamentos que contribuíram para a minha formação e pela amizade e carinho recebidos. A todos os colegas do Departamento de Biologia Geral, que sempre me trataram com carinho e simpatia. Em especial aos do LBEM e Genepop, que sempre me fizeram sentir parte do laboratório. A todos os meus colegas de coração do ICB por todas as alegrias que me trouxeram no dia-a-dia e por nutrirem por mim o mesmo carinho que eu sinto por eles.

À Dani e a Lane, pelos conselhos essenciais nas minhas escolhas.

Aos meus pais, irmãos e sobrinhos por sempre me mostrarem o quanto a vida é linda! Por sempre me apoiarem nas minhas escolhas e sempre me lembrarem do quão especial eu sou. Aos meus amigos Mirinha, Elisa, Danusa e Sandra por sempre terem ficado ao meu lado nos momentos de alegrias e tristezas.

Enfim, a todos que plantaram uma semente de carinho em mim e me transformaram sempre numa pessoa melhor! Obrigada!

Índice

Índice de Figuras

Índice de Tabelas

Lista de Abreviaturas e Símbolos

Encontram-se abaixo as definições das abreviaturas e símbolos que foram citados no texto:

aCGH- *array Comparative Genomic Hybridization* (Arranjo de hibridização genômica comparativa)

AMOVA- *Analysis of Molecular Variance* (Análise de variância molecular)

AR- Artrite reumatoide

CEPH- *Centre d'etude de Polymorphismes Humaines*

CGH- *Comparative Genomic Hybridization* (Hibridização genômica comparativa)

CNV- *Copy Number Variation* (Variação em número de cópia)

DA- Doença de Alzheimer

DL- Desequilíbrio de ligação

DNA- *Deoxyribonucleic acid* (Ácido desoxirribonucleico)

DP- Doença de Parkinson

EM- *Expectation-Maximization* (Maximização de esperança)

FCGRs- *Fc Gamma Receptors* (Receptores Fc gama)

FISH- *Fluorescent in situ Hybridization* (Hibridização fluorescente *in situ*)

FoSTeS- *Fork Stalling and Template Switching* (Interrupção da forquilha e troca de molde)

HAART- *Highly Active Antiretroviral Therapy* (Terapia antirretroviral fortemente ativa)

HGDP- *Human Genome Diversity Panel* (Painel de Diversidade Genômica Humana)

HNA1- *Human Neutrophil Antigen-1* (Antígeno neutrófilo humano 1)

Ia- Informatividade de ancestralidade

IBD- *Identity by Descend* (Identidade por descendência)

IgG- Imunoglobulina gama

INDELs- Polimorfismos de inserção-deleção

LCLs- Linhagens celulares de linfoblastos

LCR- *Low Copy Repeat* (Repetição de baixo número de cópias)

LES- Lúpus Eritematoso Sistêmico

LINE- *Long Interspersed Nuclear Element* (Elemento nuclear longo intercalado)

LTR- *Long Terminal Repeat* (Repetição longa terminal)

MAPH- *Multiplex Amplification Probe Hybridization*

MDS- *Multidimensional Scaling* (Escalonamento multidimensional)

MIAs- Marcadores Informativos de Ancestralidade

ML- *Maximum Likelihood* (Máxima verossimilhança)

MLPA- *Multiplex Ligation Dependent Probe Amplification*

NAHR- *Non Allelic Homologous Recombination* (Recombinação homóloga não alélica)

NK- *Natural Killer*

NHEJ- *Nonhomologous End Joining* (Junção de extremidades não homólogas)

OMIM- *Online Mendelian Inheritance in Man*

pb- Pares de bases

PCR- *Polymerase Chain Reaction* (Reação em cadeia da polimerase)

PRT- *Paralogue Ratio Test*

REDVR- *Restriction Enzyme Digest Variant Ratio*

RNA- *Ribonucleic acid* (Ácido ribonucleico)

SINE- *Short Interspersed Nuclear Element* (Elemento nuclear intercalado curto)

SNP- *Single Nucleotide Polymorphism* (Polimorfismo de nucleotídeo único)

SVA- elemento hominídeo derivado das repetições SINE-R, VNTR e *Alu*

DT1- Diabetes Tipo 1

TPRT- *Target Site-Primed Reverse Transcription*

VNTR- *Variable Number Tandem Repeat* (Repetição em *tandem* de número variável)

Resumo

Regiões variáveis em número de cópias (CNVRs) são formalmente definidas como segmentos de DNA que variam de um kilobase a vários megabases de tamanho e apresentam variações em número de cópias em relação a uma sequência referência, cujo valor usual de cópias é 2. Essas regiões representam uma parte significativa da variação genética humana, onde os graus de diversidade global e os padrões de estrutura genética se assemelham ao observado em análises de polimorfismos de nucleotídeo único (*Single Nucleotide Polymorphisms*, SNPs). Podem estar presentes em regiões genômicas que envolvem genes sensíveis a dosagem, cujos impactos funcionais são responsáveis pelas associações entre variantes de número de cópias (CNVs) raras e comuns e uma variedade de doenças genéticas complexas.

No presente trabalho nos propusemos a estimar a diversidade de genes que apresentam variação de número de cópias no genoma de populações nativas, cuja amostragem é sempre sub-representada nos grandes estudos genéticos e limitada às populações disponíveis no painel comercial HGDP-CEPH. Analisamos a distribuição de frequência de CNVs multicópias do *cluster* de beta-defensinas, genes *CCL3L1/CCL4L1*, *FCGR3A*, *FCGR3B* e *FCGR2C* e as variantes alélicas *FCGR3B*-HNA1a/1b, *FCGR2C*-Q57X e Clado I e II do gene *DEFB103,* em um grande conjunto de amostras autóctones de nativos americanos peruanos das populações de Ashaninka, Monte Carmelo, Shimaa e Quéchua, utilizando o técnica de *Paralogue Ratio Test*. Os resultados foram comparados com dados disponíveis na literatura para se investigar a existência de distribuições diferenciais de diplótipos em nativos sul-americanos, e acrescentar novas informações sobre a diversidade genética de CNVs do continente. Os resultados mostraram que os ameríndios tendem a apresentar uma maior frequência de eventos de deleções no gene *FCGR3B*, um aumento da frequência do alótipo *FCGR3B*-HNA1a, assim como de duplicações do número de cópias da região *CCL3L1/CCL4L1*. Especialmente, a população Shimaa mostra uma maior frequência do diplótipo 7 cópias na região de beta-defensinas.

Estes resultados representam possíveis impactos funcionais dos genes envolvidos na resposta imunológica destas populações apresentadas no contexto evolutivo da diversidade genômica global, com futuras implicações para estudos epidemiológicos de suscetibilidade a doenças autoimunes e infecciosas.

Abstract

Copy number variable regions (CNVRs) are formally defined as a segment of DNA ranging from one kilobase to several megabases in size and with variable copy number in comparison to the usual copy number of two of a reference genome. These regions represent a significant part of human genetic variation and reveal degrees of diversity and overall patterns of genetic structure comparable to single nucleotide polymorphism (SNP) diversity. Their presence in genomic regions harboring dosage-sensitive genes may have functional impacts that account for associations between rare and common copy number variation (CNV) and a variety of complex genetic diseases.

In the present study we aimed to estimate the diversity of genes that show CNV in Native American populations whose sampling is always underrepresented in genetic studies and limited to populations from the HGDP-CEPH commercial panel. We analyzed the frequency distribution of the multicopy immune system beta-defensin region, *CCL3L1/CCL4L1*, *FCGR3A, FCGR3B* and *FCGR2C* genes, and the allelic variants *FCGR3B*-HNA1a/1b, *FCGR2C*-Q57X and *DEFB103*-CladeI/II in a large sample set of autochthonous Peruvian Native Americans Ashaninka, Monte Carmelo and Shimaa populations using the Paralogue Ratio Test (PRT) technique. The results were compared to published data to investigate the existence of differential diplotype distributions in Native Americans from South Americans and brought up to date information concerning their CNV diversity. Amerindians tend to display an increased frequency of deletion events in the *FCGR3B* gene and of the *FCGR3B* HNA1a allotype, and duplication of *CCL3L1/CCL4L1* copy number. Especially, the Shimaa population shows a higher frequency of the seven copies diplotype at the beta-defensin region.

These results posit possible impacts to expression profiles of the involved immune genes in the presented population in the context of evolutionary global genomic diversity, with implications for future epidemiological studies on susceptibility to autoimmune and infectious diseases.