IDENTIFICAÇÃO DE VARIAÇÕES NO NÚMERO DE CÓPIAS DE REGIÕES DO GENOMA (CNVs) EM BOVINOS DA RAÇA CANCHIM

[1]FERNANDA C. P. **PEREIRA**; [2]FABIANA B. **MOKRY**; [3]LUCIANA C. A. **REGITANO;**
[4]POLIANA F. **GIACHETTO**

Nº 12.604

RESUMO

A tecnologia de genotipagem de marcadores do tipo SNP, baseada no emprego de *chips* de DNA de alta densidade, tornou possível a detecção de variações no número de cópias (CNVs, do inglês Copy Number Variation) de regiões de um genoma. Essas alterações, resultantes de duplicações ou deleções de trechos do genoma, podem contribuir para a variação fenotípica observada entre indivíduos, incluindo humanos, animais e plantas. Em animais de produção, estudos recentes reportaram a associação de CNVs com níveis de proteína e gordura no leite, vida útil de rebanhos leiteiros e susceptibilidade a nematóides em bovinos Angus. Assim, com base nessa importância, este trabalho apresenta a construção de um *pipeline* de bioinformática para identificação e análise de CNVs a partir de dados de genotipagem utilizando *chips* de DNA de alta densidade. Para a execução do *pipeline,* que se baseia no uso da ferramenta PennCNV para a detecção de CNVs, e em programas e *scripts in house*, desenvolvidos nas linguagens Perl, Shell AWK e Split, foram utilizados dados da genotipagem de bovinos Canchim com o *chip* BovineHD BeadChip (Illumina®). Os programas foram desenvolvidos para realizar a conversão entre formatos de arquivos submetidos à análise, e também para a configuração, classificação e apresentação dos resultados gerados pelo *pipeline*. Nesta última etapa, as informações relacionadas aos CNVs identificados poderão ser visualizadas em uma planilha, além de gráficos e também em um navegador, na forma de um Genome Browser. Isso permite visualizar o *locus* genômico de cada CNV e sua relação com outros elementos genéticos, como genes, regiões regulatórias e micro RNAs, dentre outras. Os próximos passos do trabalho envolvem a integração do *pipeline* desenvolvido com uma plataforma Web de bioinformática, o Galaxy, para que a ferramenta seja amplamente disponibilizada para a comunidade científica, ampliando sua utilização e tornando possível seu aperfeiçoamento pelos usuários.

[1]Bolsista PIBIC/CNPq: graduação em Sistemas de Informação, PUCC, Campinas – SP. fernandacpp@cnptia.embrapa.br
[2]Colaboradora: Pós-doutoranda, UFSCar, São Carlos – SP.
[3]Colaboradora: Pesquisadora, Embrapa Pecuária Sudeste, São Carlos – SP.
[4]Orientadora: Pesquisadora, Embrapa Informática Agropecuária, Campinas – SP.

ABSTRACT

The development of SNP genotyping technology, like that ones based on high density DNA chips, made possible to detect copy number variations (CNVs) in genomic regions. Such variations can be result of partial genomic duplications or deletions, which can drive phenotypic variation observed among individuals, including humans, animals, and plants. Furthermore, CNVs are being widely used as molecular markers, for population genetics studies or for disease diagnostics. In livestock, some recent studies reported CNVs associated with fat and protein levels in milk, the average lifespan of dairy cattle, and nematode susceptibility in Angus cattle. In view of the importance of this issue, this project describes the development of a bioinformatics pipeline for CNV detection and analysis, using data from DNA chips genotyping. The pipeline is based on the PennCNV tool for CNV detection, and programs or in-house scripts, developed in Perl language and Shell AWK, respectively. By using the data from Canchim cattle genotyped with the BovineHD BeadChip these animals, such programs and scripts are used for conversion between different file formats, as well as for configuration, classification and presentation of pipeline's generated results. For this last stage, the proposed pipeline provides a sheet with details about identified CNVs, graphics and a Genome Browser navigation tool. This browser allows the user to visualize the CNV genomic *loci* and its association with other genetic elements, such as genes, micro RNAs and other regulatory regions. Future steps involves the integration of our pipeline with the Galaxy bioinformatics plataform, to let it widely available for scientific community, encompassing its usability and improvement.

INTRODUÇÃO

CNVs (do inglês, Copy Number Variation) referem-se a variações estruturais que produzem alterações no número de cópias de uma região genômica (comparado a um genoma referência), conceitualmente com tamanho variando de 1kb a vários Mb (Henrichsen et al., 2009). Essas alterações no número de cópias podem ser resultantes de duplicações ou deleções de trechos do genoma, e atualmente acredita-se em 4 mecanismos prováveis de geração de CNVs: a recombinação homóloga não alélica, a junção de extremidades não homólogas, ocorrência de um erro durante a replicação, denominado *fork stalling and template switching* e a retrotransposição (Clop et al., 2012). Inicialmente identificadas em humanos, sabe-se hoje que as CNVs contribuem em grande parte para a variação fenotípica observada entre os indivíduos (Sebat et al., 2004), além de serem associadas a uma série de doenças como

Alzheimer (Rovelet-Lecrux et al., 2006), Parkinson (Simon-Sanchez et al., 2008) e autismo (Sebat et al., 2007), dentre outras.

Dada a importância associada as CNVs em humanos, muitos estudos têm sido conduzidos desde sua identificação. Por outro lado, estudos buscando a identificação de CNVs em animais de produção ainda são escassos. No entanto, trabalhos recentes publicados por Seroussi et al. (2010) e Liu et al. (2011) reportaram a associação de CNVs identificadas no genoma de bovinos com a quantidade de proteína e gordura no leite e vida útil do rebanho em vacas leiteiras e com a susceptibilidade a infestação por nematóides intestinais em animais da raça Angus, respectivamente.

Diferentes metodologias podem ser utilizadas na identificação de CNVs em larga escala, sendo as principais a hibridização genômica comparativa usando microarranjos de DNA (aCGH), a genotipagem de SNPs usando *chips* de DNA de alta densidade e o sequenciamento genômico. A genotipagem de animais participantes dos programas de melhoramento genético, por meio da utilização de *chips* de DNA tem sido bastante utilizada na Embrapa, como ferramenta para a realização de estudos visando à introdução de seleção genômica nos rebanhos e plantéis, e também para estudos de associação genética. A possibilidade de utilização desses *chips* para a identificação de CNVs com precisão e resolução consideravelmente altas (Wang e Bucan, 2008) abre então uma nova oportunidade para a realização de estudos que buscam a caracterização da arquitetura genômica das CNVs em animais de produção.

Assim, esse trabalho teve por objetivo implementar um *pipeline* de análises para a identificação e visualização de CNVs a partir dos relatórios de genotipagem de SNPs usando *chips* de DNA, gerados pelos programas de melhoramento genético da Embrapa. A partir das CNVs identificadas, estudos de associação com características fenotípicas dos animais poderão ser realizados, visando identificar regiões no genoma importantes na determinação de características de interesse econômico relevantes e também na descoberta de genes e mecanismos de regulação relacionados a essas características.

MATERIAL E MÉTODOS

Dada a importância das CNVs, que têm sido associadas a doenças e outras características quantitativas em humanos e animais, diferentes algoritmos para a sua detecção, a partir de dados de intensidade dos *chips* de DNA foram desenvolvidos, incluindo o PennCNV, QuantiSNP, HMMSeg e cnvPartition (Tsuang et al., 2010). O presente trabalho utilizou o PennCNV (Wang et al. 2007), uma ferramenta de *software*

livre, desenvolvida em Perl – Pratical Extract and Report Language, que incorpora vários componentes em um modelo oculto de Markov (HMM). O PennCNV analisa os padrões de intensidade do sinal em todo o genoma e identifica, por meio de marcadores (SNPs) consecutivos, as mudanças no número de cópias de trechos do genoma.

O PennCNV utiliza duas medidas de intensidade de sinal calculadas para cada SNP. O LRR (log_2 da razão de R, onde R é o valor normalizado da intensidade total para os dois alelos do SNP) e o BAF (frequência do alelo B; medida normalizada da relação de intensidade alélica para cada SNP), gerados a partir das plataformas de genotipagem. A combinação dos valores de LRR e BAF pode ser usada para inferir o número de cópias de regiões do genoma, como ilustrado na Figura 1. Quando ocorre uma deleção, há diminuição dos valores de LRR e uma ausência de heterozigotos nos valores de BAF (os agrupamentos de genótipos de SNPs localizam-se ao redor de 0 ou 1); na presença de uma duplicação, há um aumento nos valores de LRR e uma separação do genótipo heterozigoto em dois grupos.

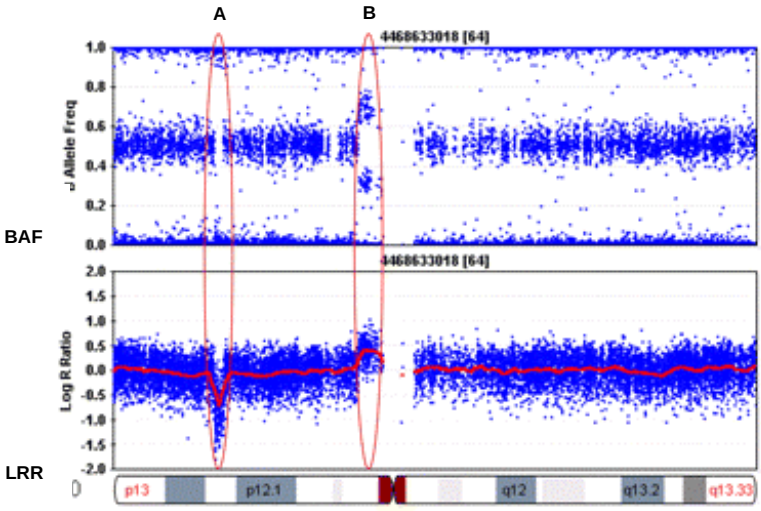

FIGURA 1. Valores de BAF (Frequência do Alelo B) e LRR (log_2 da razão de R) em uma região do cromossomo 20 de um indivíduo apresentando uma deleção (A) e uma duplicação (B). Em A, os genótipos se agrupam ao redor de 0 e 1 e os valores de LRR são reduzidos. Em B, aparecem 2 grupos do genótipo heterozigoto e os valores de LRR são elevados. Adaptado de Wang et al. (2007).

Para a execução da ferramenta, o PennCNV (http://www.openbioinformatics.org/penncnv/penncnv_download.html) foi instalado em um servidor do Laboratório de Bioinformática Aplicada da Embrapa (máquina IBM System x3850 X5, com 512GB de memória RAM, 8 processadores Xeon 6-core E7540 2.00GHz HT com 18 MB de memória cache, 8 HDs SAS de 300GB utilizados em RAID1 e RAID5, disponibilizando volumes de 570GB e 820GB, respectivamente, 2 placas controladoras HBA 8 Gb Fibre Channel, 2 placas Ethernet Gigabit) e executado em Sistema Operacional Linux.

Foram utilizados dados da genotipagem de SNPs de 400 bovinos machos da raça Canchim, obtidos no projeto em Rede do Sistema Embrapa de Gestão, Rede Genômica Animal. A genotipagem dos animais foi feita utilizando-se o *chip* BovineHD BeadChip (Illumina®), que permite a investigação de mais de 770 mil SNPs simultaneamente, dispersos ao longo de todo o genoma. A partir dos relatórios brutos de genotipagem, programas em Perl foram desenvolvidos para a obtenção dos 2 arquivos de entrada (arquivos texto com os dados de intensidade de sinal), requeridos pelo PennCNV e aqui denominados LRR/BAF e PFB. O LRR/BAF é um arquivo com 6 campos de informações para cada SNP: nome do SNP, cromossomo, posição, genótipo do SNP, LRR e BAF. Já o PFB possui 4 campos: nome do SNP, cromossomo, posição e valor de PFB (que traz a informação de frequência do alelo B na população sob estudo). A frequência do alelo B foi calculada a partir dos relatórios de genotipagem de todos os animais amostrados na população, e o arquivo PFB foi gerado por meio do programa ***compile_pfb.pl****,* disponível no PennCNV. Modelados os arquivos (21GB de dados), o programa ***run.pl****,* desenvolvido em Perl, fez a chamada do executável ***detect_cnv.pl*** contido no PennCNV, responsável pela detecção das CNVs. Um exemplo do comando necessário para a execução do programa ***detect_cnv. pl*** é exibido abaixo:

~$ perl detect_cnv.pl -test -hmm .. /lib/example.hmm -pfb .. /lib/example.pfb -log -ex3.log -out dados.rawcnv ../partes2/* -conf>

No comando acima, o programa *detect_cnv.pl* é executado com os argumentos: ***-hmm***, o qual indica o modelo oculto de Markov para a detecção de CNV, contido no pacote do PennCNV, e renomeado aqui como *example.hmm*; ***-pfb****,* com o caminho do arquivo chamado de *example.pfb*; ***-out****,* que faz a saída de dados, o caminho onde estão as amostras a serem analisadas é o *../partes2/;* ***-conf****,* lança nas últimas colunas do relatório de saída a pontuação de confiança dos resultados obtidos.

Para facilidade de conversão de arquivos, o PennCNV disponibiliza o executável ***convert_cnv.pl,*** que converte o arquivo em um modelo delimitado por tabulação, que pode ser aberto em planilhas, facilitando a organização e visualização dos resultados. Um comando de execução do programa ***convert_cnv.pl*** é exemplificado abaixo:

~$ path_convert_cnv -intype ../partes2 -outtype tab -output ex1.tabcnv>

O executável ***convert_cnv.pl****,* também pode ser utilizado para converter formatos de arquivos de CNVs de outros programas, para o formato de saída do PennCNV, o que pode ser útil para se fazer uma análise comparativa dos resultados obtidos com outros algoritmos.

A análise visual das CNVs identificadas por meio da execução do programa ***detect_cnv. pl*** frequentemente é realizada, e pode auxiliar na identificação de resultados falso-positivos. O programa ***visualize_cnv.pl***, exemplificado no comando abaixo*,* ilustra os dados, plotando os valores de intensidade de sinal (LRR e BAF) para cada CNV chamada, identificando o cromossomo e a região, de modo que o usuário possa examinar visualmente e decidir se as chamadas são confiáveis ou não. Essa função exige a instalação do programa R.

~$ perl visualize_cnv.pl -format plot -signal dados_div_ai_cabec.txt dados_ai.rawcnv>

O comando ***visualize_cnv.pl*** é executado com os seguintes argumentos: ***-format****,* que especifica o formato plot; -***signal,*** o arquivo de intensidade de sinal e o arquivo de saída com o resultado da chamada da CNV.

RESULTADOS E DISCUSSÃO

A Figura 2 mostra um exemplo do arquivo de entrada LRR/BAF, requerido pelo PennCNV. A primeira linha do arquivo especifica o significado de cada coluna delimitada por tabulação. São seis campos em cada linha, correspondentes ao nome do SNP, cromossomo, posição do SNP, genótipo do SNP, $\log_2$ da razão de R (LRR) e frequência do alelo B (BAF).

Para geração do segundo arquivo de entrada, o PFB (Figura 3), foi necessário o cálculo da frequência do alelo B na população e para tanto considerou-se os dados dos 400 animais genotipados.

Name	Chr	Position	Gtype.Allele Calls	B Allele Freq	Log R Ratio
ARS-BFGL-BAC-10172	4	6371334	BB	0.9964	-0.1636
ARS-BFGL-BAC-1020	4	7928189	AB	0.4034	0.5906
ARS-BFGL-BAC-10245	4	31819743	AB	0.4980	-0.0866
ARS-BFGL-BAC-10345	4	6133529	AB	0.4267	0.1027
ARS-BFGL-BAC-10365	4	27005721	BB	1.0000	-0.0118

FIGURA 2. Modelo do arquivo de entrada LRR/BAF, requerido pelo PennCNV. Name: nome do SNP; Chr: cromossomo; Position: posição do SNP no cromossomo; Gtype.Allele Calls: genótipo do SNP; B Allele Freq: frequência do alelo B; Log R Ratio: $\log_2$ da razão de R.

Assim como para o LRR/BAF, os primeiros 3 campos designam o nome e as coordenadas do SNP e o último campo contém a informação de frequência do alelo B.

Name	Chr	Position	PFB
ARS-BFGL-BAC-10172	14	6371334	0.957393483709273
ARS-BFGL-BAC-1020	14	7928189	0.653553299492386
ARS-BFGL-BAC-10245	14	31819743	0.913316582914573
ARS-BFGL-BAC-10345	14	6133529	0.134517766497462
ARS-BFGL-BAC-10365	14	27005721	0.99874686716792

FIGURA 3. Modelo do arquivo de entrada PFB, requerido pelo PennCNV. Name: nome do SNP; Chr: cromossomo; Position: posição do SNP no cromossomo; PFB: frequência do alelo B na população.

A execução do programa ***detect_cnv.pl*** gerou o arquivo de saída ilustrado na Figura 4, que informa as CNVs identificadas. Os campos correspondem às coordenadas da região contendo a CNV no cromossomo, o número de marcadores SNP na região de CNV, o comprimento da CNV, o estado da CNV e a estimativa do número do cópias, o caminho e nome do arquivo de intensidade de sinal, o SNP inicial

e o final e a pontuação de confiança da chamada (escore que reflete a confiança de que a região detectada realmente existe).

chr1:16021282-16023240	numsnp=3	length=1,959	state1,cn=0	../partes2/dados_div_aa	start=Bovine HD070004885	end=Bovine HD070004887	conf=13.638
chr2:46477034-46485436	numsnp=4	length=8,403	state1,cn=0	../partes2/dados_div_aa	start=Bovine HD0200013459	end=Bovine HD0200013462	conf=12.316
chr5:55135352-55176436	umsnp=10	length=41,085	state1,cn=0	../partes2/dados_div_aa	start=Bovine HD0500015712	end=Bovine HD0500015722	conf=38.704
chr5:117524110-117639815	numsnp=12	length=115,706	state1,cn=0	../partes2/dados_div_aa	start=Bovine HD0500034144	end=Bovine HD0500034161	conf=22.519
chr7:18541538-18566035	numsnp=5	length=24,498	state1,cn=0	../partes2/dados_div_aa	start=Bovine HD5189	end=Bovine HD5195	conf=12.850

FIGURA 4. Modelo do arquivo de saída, com as CNVs identificadas pelo programa PennCNV. Cada linha representa uma CNV identificada, onde o 1º campo representa a região de CNV identificada, o 2º campo mostra o numero de SNPs presentes na CNV, o 3º campo informa o comprimento da CNV, o 4º campo traz informações sobre o estado e o número de cópias da CNV (o estado da CNV está relacionado ao número de cópias e designa uma distribuição dos genótipos de CNV).

No exemplo apresentado na Figura 4, tem-se que a primeira linha do arquivo especifica uma CNV identificada no cromossomo 1, contendo 3 SNPs e com tamanho de 1.958bp. O estado 1, associado ao número de cópias igual a 0 informa que na região especificada houve uma deleção das 2 cópias do genoma, sendo nulo o genótipo da CNV em questão.

Conforme já citado, a inspeção visual das CNVs identificadas permite uma avaliação para confirmação da identidade das mesmas. A Figura 5 mostra a imagem (formato JPG) gerada pelo programa ***visualize_cnv.pl***, a partir dos valores de LRR e BAF de uma região de CNV identificada no cromossomo 14 de uma das amostras utilizadas. Na região que compreende a posição 62.650.000 - 62.850.000pb, nota-se que os valores de LRR encontram-se abaixo de 0, enquanto que os valores de BAF agrupam-se ao redor de 0 e 1, o que representa a deleção de uma cópia e CNV com genótipo A, B. O programa visualiza_cnv.pl também converte o relatório de saída do PennCNV para o formato BED, permitindo a visualização das CNVs em um Genome Browser (visualizador de genomas). Observando-se a região em questão no Genome Browser da UCSC (http://genome.ucsc.edu/cgi-bin/hgGateway, dados não mostrados), pode-se observar que a CNV em questão, caracterizada como uma deleção, está presente em uma região rica em genes. Um estudo mais profundo acerca dos genes presentes nessa região, assim como estudos de associação com características

fenotípicas dos animais amostrados podem levar a descobertas significativas relacionadas a essa variação estrutural do genoma.

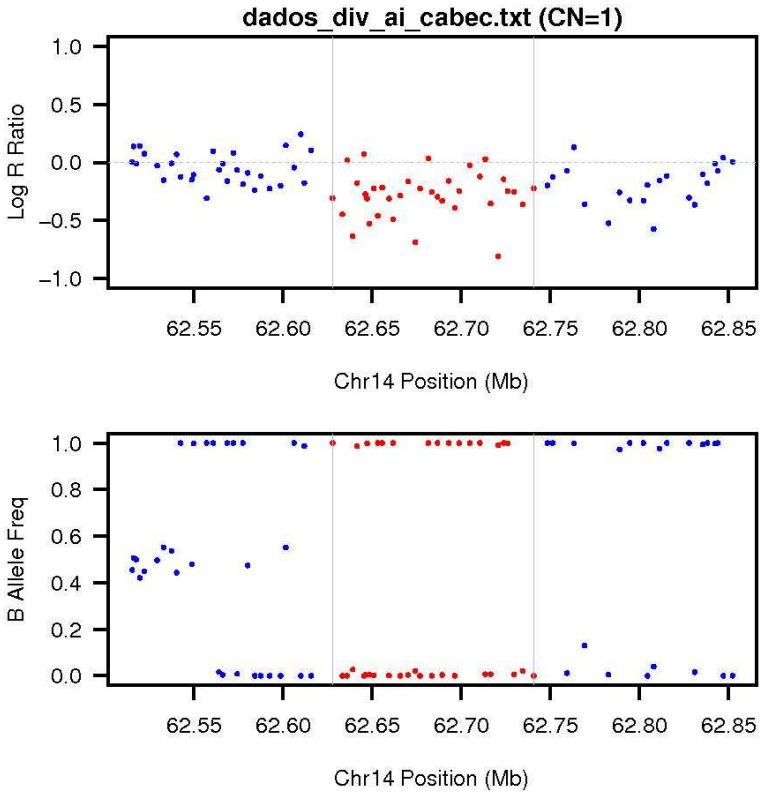

Figura 5. Imagem gerada a partir dos valores de LRR e BAF de uma região de CNV identificada no cromossomo 14, na amostra dados_div_ai_cabec.txt. CN=1 (número de cópias = a 1 e estado 2) informa a deleção de de 1 cópia do genoma, com genótipo da CNV A, B.

CONCLUSÃO

Foi elaborado um *pipeline* de identificação de CNVs a partir de dados de genotipagem usando o *chip* BovineHD BeadChip. Esse *pipeline* inclui programas para a geração dos arquivos de entrada requeridos pela ferramenta PennCNV, a execução dos programas da ferramenta PennCNV para identificação das CNVs e programas para a visualização das regiões identificadas em planilhas, gráficos com os valores de LRR e BAF e também em um Genome Browser. Prevê-se, como próximos passos, a inclusão do *pipeline* proposto na plataforma Galaxy (ferramenta de gerenciamento de *workflows* baseada em Web), e sua disponibilização pelo Laboratótio Multiusuário de

Bioiformática da Embrapa Galaxy para ampla utilização por toda a comunidade científica.

AGRADECIMENTOS

Ao CNPQ – PIBIC, pela bolsa concedida. A Empresa Brasileira de Pesquisa Agropecuária (Embrapa), pelos dados de genotipagem dos animais. A Embrapa Informática Agropecuária, pela oportunidade de estágio e ao pesquisador Dr. Roberto H. Herai, pelas contribuições.

REFERÊNCIAS

HENRICHSEN, C.N.; CHAIGNAT, E.; REYMOND, A. Copy number variants, diseases and gene expression. **Human Molecular Genetics**, v.18, p.R1-8, 2009.

LIU, G.E., BROWN, T., HEBERT, D.A. et al. Initial analysis of copy number variations in cattle selected for resistance or susceptibility to intestinal nematodes. **Mammalian Genome**, v.22, p.111-21, 2011.

ROVELET-LECRUX, A., HANNEQUIN, D., RAUX, G., et al. APP locus duplication causes autosomal dominant early-onset alzheimer disease with cerebral amyloid angiopathy. **Nature Genetics**, v.38, p.24–26, 2006.

SEBAT, J., LAKSHMI, B., TROGE, J., et al. Large-scale copy number polymorphism in the human genome. **Science**, v.305, p.525–528, 2004.

SEROUSSI, E., GLICK, G., SHIRAK, A., et al. Analysis of copy loss and gain variations in Holstein cattle autosomes using BeadChip SNPs. **BMC Genomics**, v.11, p.673, 2010.

SIMON-SANCHEZ, J., SCHOLZ, S., MATARIN, M.D.E.L.M., FUNG, H.C., HERNANDEZ, D., GIBBS, J.R., BRITTON, A., HARDY, J., SINGLETON, A. Genomewide SNP assay reveals mutations underlying Parkinson disease. **Human Mutations**, v.29, p.315–322, 2008.

TSUANG, D.W., MILLARD, S.P., ELY, B., et al. The effect of algorithms on copy number variant detection. **PLoS ONE**, v.5, n.12, e14456, 2010.

WANG, K., BUCAN, M. Copy number variation detection via high-density SNP genotyping. **CSH Protocols**, v.3, p.6, 2008.

WANG, K., LI, M., HADLEY, D., LIU, R., GLESSNER, J., GRANT, S., HAKONARSON, H., BUCAN, M. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. **Genome Research**, v.17, p.1665-1674, 2007.