Troost – Busca de interações entre trios de SNPs em estudos de associação de genoma inteiro

Troost – Search for interactions among trios of SNPs in genome-wide association studies

José Osório de Oliveira Azevedo Neto

TESE APRESENTADA AO PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA DA UNIVERSIDADE DE SÃO PAULO PARA OBTENÇÃO DO TÍTULO DE DOUTOR EM CIÊNCIAS

Orientador: Prof. Dr. Sergio Russo Matioli

Co-orientadora: Profª. Dra. Júlia Maria Pavan Soler

São Paulo, 7 de novembro de 2013

Ficha Catalográfica

Azevedo Neto, José Osório de Oliveira
Troost – Busca de interações entre trios de SNPs em estudos de associação de genoma inteiro
154 páginas

Tese (Doutorado) - Programa Interunidades em Bioinformática da USP.

1. Epistasia 2. Estudos de associação I. Universidade de São Paulo. Instituto de Biociências. Departamento de Genética e Biologia Evolutiva.

Comissão Julgadora:

Prof[a]. Dr[a]. Elida Benquique Ojopi	Prof[a]. Dr[a]. Helena Brentani
Prof. Dr. Carlos Alberto Bragança Pereira	Prof. Dr. Paulo Alberto Otto

Prof. Dr. Sergio Russo Matioli
Orientador

Resumo

Os estudos de associação de genoma inteiro têm encontrado alguns marcadores associados a doenças notoriamente hereditárias com herança complexa, mas, muitas vezes, estes marcadores somente explicam uma pequena parte da herdabilidade. Este relativo insucesso é atribuído, entre outras causas, à epistasia, ou seja, interação entre diferentes locos genéticos. A busca por epistasia é complexa e exige intensos recursos computacionais. Diversos métodos têm sido propostos para abordar este problema, incluindo métodos estatísticos tradicionais, busca estocástica e métodos heurísticos. Poucos destes métodos são capazes de processar as grandes massas de dados produzidas nos estudos caso-controle de genoma inteiro, e ainda menos métodos buscam conjuntos de três ou mais marcadores. A busca exaustiva de conjuntos de marcadores epistáticos é inviável hoje em dia para estes conjuntos, mas o algoritmo BOOST (WAN et al., 2010) mostrou que ela é relativamente fácil para pares de locos, em especial com o uso de placas gráficas como processadores (GPGPU). Partindo deste recente sucesso, propomos um algoritmo em fases para a busca de trios de locos que interagem, utilizando a busca de pares como passo inicial, uma abordagem ainda não utilizada. Outra ideia fundamental do algoritmo proposto é a extensão da concepção de trio de marcadores para um trio de blocos haplotípicos, onde cada bloco é formado por marcadores próximos entre si. Usando os dados do WTCCC, o Troost (de TRio+bOOST) sugeriu trios potencialmente epistáticos em todas a sete doenças. Quando submetidos à confirmação em amostra independente, os trios não puderam ser confirmados, exceto os trios para diabetes tipo 1 (T1D). Duzentos e oito trios foram confirmados para T1D, com baixos valores-P e genótipos combinados de risco com altas razões de chances. Os SNPs que compõem estes trios estão todos na região MHC, sabidamente associada à doença, exceto por um deles que está no cromossomo cinco e não havia sido previamente relacionado à T1D.

Abstract

Genome-wide association studies have found some markers associated with diseases with complex inheritance. However, these markers explain only a fraction of the previously estimated heritability of the trait. This relative failure has been credited, among other causes, to epistasis, *i.e.* the interaction among genotypes at different loci. The search for epistasis is complex and requires intense computational resources. Many methods have been proposed to approach this problem, including traditional statistics, stochastic search, and heuristic methods. Few of them are capable of extracting, from the large amount of data produced in genome-wide case-control studies, useful information about sets of markers associated with the trait in question. Exhaustive search of sets of interacting markers is unfeasible nowadays for sets of three or more markers, but the BOOST algorithm (WAN et al., 2010) showed that the search is relatively easy for pairs of SNPs, in particular with the use of graphic cards for general processing (GPGPU). Starting from this recent success, we propose an algorithm in phases for the search for trios of interacting loci, using the search for pairs as the initial step, an approach not tried yet, to our knowledge. Another important idea of our algorithm is the extension of the concept of trio of markers to a trio of haplotypic blocks, where each block is formed by neighbor markers. Using data from WTCCC, the Troost (from TRio+bOOST) algorithm suggested potentially epistatic trios in all seven diseases. When submitted to a confirmation in an independent sample, the results could not be confirmed, except for type-1 diabetes (T1D). Two hundred eight trios were confirmed for T1D, with low p-values and risk combined genotypes with high odds ratio. The SNPs that form those trios are all in the MHC region, which is known to be strongly associated to T1D, except by one SNP in chromosome five that has not been previously associated with T1D.

Índice

Resumo 3
Abstract 4
Índice 5
1 Introdução aos estudos de associação de genoma inteiro 9
1.1 Doenças de herança complexa 9
Exemplo de doença de herança complexa: o autismo 9
1.2 Marcadores genéticos 10
1.3 Estudos caso-controle com marcador único 11
Teste em tabela de contingência 11
Regressão linear 12
Extensão para incluir efeito de dominância 13
Regressão logística 14
1.4 DbSNP e plataformas de genotipagem 15
DbSNP 15
Plataformas de genotipagem de alta densidade 16
1.5 Estudos de genomas inteiros. 16
1.6 Problemas com os GWAS 18
2 Introdução à epistasia – Interação de fatores genéticos entre si 20
2.1 Conceito biológico de epistasia 20
2.2 Conceito estatístico de epistasia 21
2.3 Modelos de epistasia 22
2.4 A maldição da dimensionalidade 23
3 Métodos para detecção de epistasia 25
3.1 Técnicas de busca condicional em duas fases 25
3.2 MDR – Multifactor Dimensionality Reduction 26
Como funciona o MDR 26
Crítica ao MDR 27
3.3 Ganho de Informação (IG) 27
3.4 SNP Harvester 27
3.5 AntEpiSeeker 27
3.6 AGR – Redução de Grafos de Associação 28
3.7 Comparação dos métodos 28
3.8 Busca exaustiva de pares de locos 30
BOOST 30
4 Objetivos 34
5 Material e métodos 35
5.1 Compilação dos tempos de busca de epistasia 35
5.2 Dados para a pesquisa – o estudo do WTCCC 35
Descrição dos dados do WTCCC 36
Qualidade dos dados do WTCCC 37
5.3 Análise exploratória dos pares de locos 38
5.4 A proposta: o algoritmo Troost 40
Ideia 1: Busca de trios epistáticos a partir dos pares 40
Ideia 2: Construir blocos haplotípicos entre os SNPs presentes nos pares fortes 43
Duplo-haplótipos 44
Esquema geral do Troost 46
Módulos do Troost 47
Detalhamento dos Módulos do Troost 49
5.5 Confirmação dos resultados com amostras independentes 59
Razão de chances 60

5.6 Recursos utilizados na pesquisa 60
Hardware 60
Software 61
6 Resultados 62
6.1 Tempo de execução de testes de epistasia 62
6.2 Resultados da análise exploratória 64
Entre os SNPs que aparecem nos pares, há predominância daqueles que têm maior associação marginal? 64
Os SNPs com maior efeito aditivo obscurecem os pares, ao aparecer num grande número deles? 64
Quantos SNPs existem na lista de pares? 64
Há muitos pares de SNPs próximos, com alto desequilíbrio de ligação (LD) entre si? 65
SNPs com alto LD aparecem pareados com os mesmos outros SNPs? 65
É possível escolher um único SNP em grupos que têm alto LD entre si? 66
Existem padrões encadeados do tipo AB, BC, CD? 66
6.3 Resultados da pesquisa 66
Resumo numérico das pesquisas 66
Transtorno Bipolar – BD 68
Doença arterial coronariana – CAD 75
Doença de Crohn – CD 83
Hipertensão arterial – HT 96
Artrite reumatoide – RA 102
Artrite reumatoide – RA 102
Diabetes tipo 2 – T2D 106
Diabetes tipo 1 – T1D 110
7 Discussão 120
7.1 Uso das distâncias em centimorgans 120
7.2 Blocos haplotípicos definidos por pares 121
7.3 Desempenho computacional 122
7.4 Sucesso: o Troost encontrou trios, mas não nos controles 123
7.5 BD, CAD, RA, T2D – A maldição do vencedor 123
7.6 CD – Resultado confirmado, mas rejeitado 124
7.7 HT – Mais um erro nos dados 126
7.8 T1D – Novo SNP fora da MHC interage com ela 128
7.9 T1D – Genótipos triplos com alto efeito 129
7.10 Acerto na escolha do BOOST 130
7.11 Perspectivas 130
8 Conclusão 131
9 Referências Bibliográficas 132
Apêndice – Programas-fonte 137
SNPdist.c 137
RefAllele.c 141
MakeSNPsFile.R 143
SampleDisease.R 144
tped2boost.c 144
Troost.R 147
TriosFromPairs.c 149
mM2snp.R 150
TabTrio.R 151
TrioStat.R 154

Agradecimentos

À minha esposa Cris e meus filhos Tomás e Laura, pelo apoio e por suportar os muitos fins de semana em que eu estava trabalhando nesta tese;

Ao meu orientador Sérgio e co-orientadora Júlia, pela sempre pronta e valiosa ajuda;

À minha irmã Sílvia, pelo apoio e comentários precisos;

À nossa secretária Patrícia, sempre disponível e competente;

À IBM, onde meu emprego em *home-office* com horário flexível permitiu-me cursar o doutorado e fazer esta tese, sem necessidade de bolsa;

Ao WTCCC pelo fornecimento dos preciosos dados genotípicos.

This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113 and 085475.

1 Introdução aos estudos de associação de genoma inteiro

1.1 Doenças de herança complexa

Uma das principais motivações do projeto Genoma Humano foi a identificação de genes causadores de doenças, para posterior descoberta das vias metabólicas envolvidas e possível desenvolvimento de drogas curativas ou preventivas. Entre estas importantes doenças podemos citar diabetes, asma, hipertensão arterial, esclerose múltipla, autismo, transtorno bipolar, câncer, esquizofrenia e obesidade.

Algumas destas doenças têm incidência elevada em populações do ocidente, outras nem tanto, mas nenhuma delas é classificada como doença rara. Estas doenças apresentam fortes evidências de possuírem causas genéticas, por serem recorrentes nas mesmas famílias. No entanto, elas definitivamente não possuem herança mendeliana simples. As buscas por genes causadores têm produzido alguns resultados, mas muito aquém do que era esperado quando grande quantidade de recursos humanos, financeiros e materiais foi investida no projeto genoma humano e nas centenas de estudos de associação já realizados, muitas vezes com a colaboração de dezenas de instituições de pesquisa distribuídas por diversos países e continentes.

Exemplo de doença de herança complexa: o autismo

O autismo é uma doença que desafia os geneticistas há muitas décadas. Nos Estados Unidos, uma criança em cada 150 é portadora de doença do espectro autista. (WANG et al., 2009a). É portanto uma doença grave, causadora não apenas de desconforto e gastos no sistema de saúde, mas também de sofrimento para vítimas e seus familiares. Em relação aos aspectos genéticos: se o casal já tiver um filho autista, a chance de um novo bebê ser também autista é de 10% (WANG et al., 2009a). Esta chance cresce de 1 em 150 na população para 1 em 10 nessa família. No caso de gêmeos idênticos, se um deles apresenta sintomas de autismo, a chance de o outro também os apresentar é de 90% (WANG et al., 2009a). Estes resultados sugerem que o autismo tenha um importante componente genético.

As buscas de marcadores relacionados a doenças têm se concentrado nos efeitos principais de cada marcador, supondo-os independentes dos demais genes.

Genes associados à incidência de autismo foram encontrados, mas a maior parte da herdabilidade permanece inexplicada. (ALTSHULER; DALY, 2007)

O autismo é citado aqui apenas como exemplo de doença genética. No decorrer da nossa pesquisa, em nenhum momento lidamos com dados sobre autismo.

1.2 Marcadores genéticos

O genoma humano, representado pela sequência de referência, é uma sequência de consenso entre os poucos indivíduos cujos genomas foram sequenciados com esta finalidade. Ele reflete a uniformidade do genoma para a humanidade, mas não retrata as diferenças entre as pessoas. Cerca de 0,1% do genoma é diferente entre duas pessoas não aparentadas, o que equivale a cerca de um a cada mil nucleotídeos sequenciados. (HAPMAP CONSORTIUM, 2003). Esta variação genética entre as pessoas pode explicar porquê alguns indivíduos são mais suscetíveis a uma doença do que outros, e também pode explicar a recorrência de doenças em determinadas famílias. O projeto Genoma Humano pode ser visto como um primeiro passo para a análise profunda da relação entre a variação genética e a propensão a doenças hereditárias, pois produziu um mapa detalhado sobre o qual outras descobertas podem ser adicionadas. O próximo passo foi encontrar as diferenças comuns entre os genomas, e essas diferenças são determinadas por *marcadores moleculares*. O tipo mais comum de marcador molecular é o SNP (*Single Nucleotide Polymorphism),* ou polimorfismo de nucleotídeo único. É uma posição específica no genoma, a qual em geral tem uma base nitrogenada (A, C, G ou T) como a mais comum na população, o alelo comum, e outra como a menos comum, o alelo alternativo ou menos frequente. Os indivíduos normalmente são homozigotos para o alelo comum, mas podem ser heterozigotos ou homozigotos para o alelo alternativo. Embora possa haver mais que duas bases nitrogenadas polimórficas em um SNP, a sua imensa maioria é bialélica.

Há outros tipos de marcadores, tais como minissatélites, microssatélites e variações estruturais (inserções, exclusões, inversões e translocações). Mas os SNPs são, atualmente, os mais adotados nos estudos genômicos, por serem mais numerosos, proporcionarem uma cobertura mais densa do genoma e por serem mais facilmente abordáveis por métodos de genotipagem em larga escala.