Janelas de influência em polimorfismos de um único nucleotídeo

André Jucovsky Bianchi

Departamendo de Ciência da Computação
Instituto de Matemática e Estatística
Universidade de São Paulo

Trabalho de Formatura Supervisionado, 2009
Orientadora: Florencia Leonardi

Sumário

Outline

Genes?

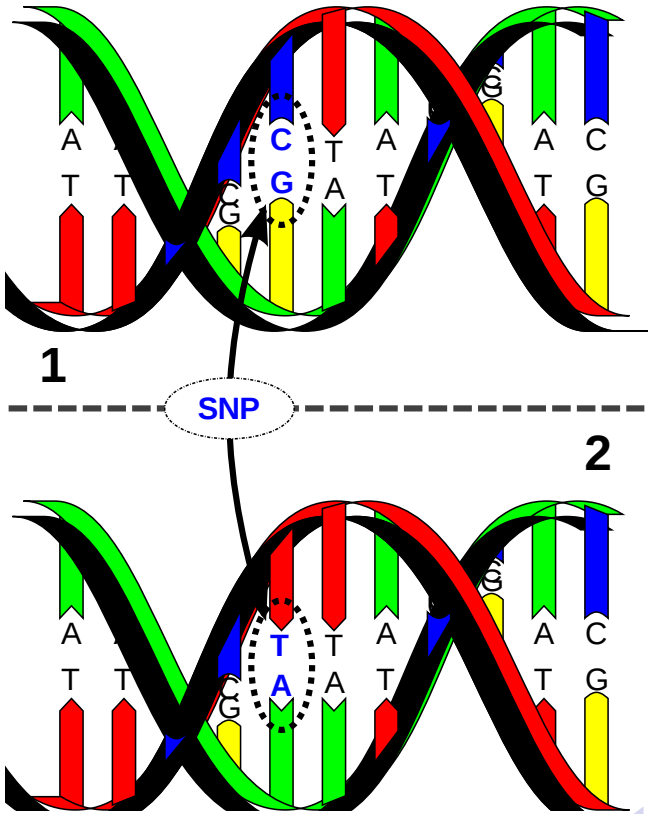
A concepção de “gene" tem mudado com o avanço das pesquisas:

- Genes sobrepostos.
- Genes contidos em genes.
- Genes segmentados: recorta e cola, um mesmo gene codifica diferentes tipos de proteínas.
- RNA como mecanismo corretor de falhas[2].

Sequenciamento do DNA

Formas de sequenciamento (marcadores):

- Microsatélites. Ex: $(CA)^n$, $n = 10 \ldots 100$
- **S**ingle-**N**ucletotide **P**olimorphisms. Longas sequências onde bases alternativas ocorrem com alta frequência (1%!) em apenas um locus[4].



Outline

O conjunto de dados

- **GAW16**: Genetic Analysis Workshop - 2008.
  - NARAC: Consórcio norte-americano para a Artrite Reumatóide
  - 2.062 indivíduos: 868 caso, 1194 controle.
  - 500.180 **SNP**s.

- Contagem de frequência alélica:
  - Homozigoto com maior frequência dentro do **SNP** recebe 0.
  - Heterozigoto recebe 1.
  - Homozigoto com menor frequencia dentro do **SNP** recebe 2.

Outline

Objetivos da análise dos dados

Objetivo principal: determinação de regiões de influência entre **SNP**s.

1. Determinação de *vizinhanças de influência* para cada **SNP** através da maximização da pseudo-verossimilhança penalizada[1].
2. Agrupamento das vizinhanças em *janelas de influência*.
3. Utilização de diferentes medidas de distância para determinar a relevância de cada janela na diferenciação entre entre indivíduos caso e controle.

Objetivos da análise dos dados

Objetivo principal: determinação de regiões de influência entre **SNP**s.

1. Determinação de *vizinhanças de influência* para cada **SNP** através da maximização da pseudo-verossimilhança penalizada[1].
2. Agrupamento das vizinhanças em *janelas de influência*.
3. Utilização de diferentes medidas de distância para determinar a relevância de cada janela na diferenciação entre entre indivíduos caso e controle.

Objetivos da análise dos dados

Objetivo principal: determinação de regiões de influência entre **SNP**s.

1. Determinação de *vizinhanças de influência* para cada **SNP** através da maximização da pseudo-verossimilhança penalizada[1].
2. Agrupamento das vizinhanças em *janelas de influência*.
3. Utilização de diferentes medidas de distância para determinar a relevância de cada janela na diferenciação entre entre indivíduos caso e controle.

Objetivos da análise dos dados

Objetivo principal: determinação de regiões de influência entre **SNP**s.

1. Determinação de *vizinhanças de influência* para cada **SNP** através da maximização da pseudo-verossimilhança penalizada[1].
2. Agrupamento das vizinhanças em *janelas de influência*.
3. Utilização de diferentes medidas de distância para determinar a relevância de cada janela na diferenciação entre entre indivíduos caso e controle.

Outline

Modelo

- $S = 500.180$ **SNP**s, $N = 2.062$ indivíduos.
- Um **SNP** é uma variável aleatória X_j em $A = \{0, 1, 2\}$, $1 \leq j \leq S$.
- A realização de uma sequência de **SNP**s é o código genético de um indivíduo: $(x_1, x_2, \ldots x_S)$.
- Uma *vizinhança* do **SNP** j é um subconjunto de $\{1, \ldots, S\} \setminus \{j\}$.
- $\mathfrak{D}_{S \times N}$ é a matriz de amostras.

Modelo

- $S = 500.180$ **SNP**s, $N = 2.062$ indivíduos.
- Um **SNP** é uma variável aleatória X_j em $A = \{0, 1, 2\}$, $1 \leq j \leq S$.
- A realização de uma sequência de **SNP**s é o código genético de um indivíduo: $(x_1, x_2, \ldots x_S)$.
- Uma *vizinhança* do **SNP** j é um subconjunto de $\{1, \ldots, S\} \setminus \{j\}$.
- $\mathfrak{D}_{S \times N}$ é a matriz de amostras.

Modelo

- $S = 500.180$ **SNP**s, $N = 2.062$ indivíduos.
- Um **SNP** é uma variável aleatória X_j em $A = \{0, 1, 2\}$, $1 \leq j \leq S$.
- A realização de uma sequência de **SNP**s é o código genético de um indivíduo: $(x_1, x_2, \ldots x_S)$.
- Uma *vizinhança* do **SNP** j é um subconjunto de $\{1, \ldots, S\} \setminus \{j\}$.
- $\mathfrak{D}_{S \times N}$ é a matriz de amostras.

Modelo

- $S = 500.180$ **SNP**s, $N = 2.062$ indivíduos.
- Um **SNP** é uma variável aleatória X_j em $A = \{0, 1, 2\}$, $1 \leq j \leq S$.
- A realização de uma sequência de **SNP**s é o código genético de um indivíduo: $(x_1, x_2, \ldots x_S)$.
- Uma *vizinhança* do **SNP** j é um subconjunto de $\{1, \ldots, S\} \setminus \{j\}$.
- $\mathfrak{D}_{S \times N}$ é a matriz de amostras.

Modelo

- $S = 500.180$ **SNP**s, $N = 2.062$ indivíduos.
- Um **SNP** é uma variável aleatória X_j em $A = \{0, 1, 2\}$, $1 \leq j \leq S$.
- A realização de uma sequência de **SNP**s é o código genético de um indivíduo: $(x_1, x_2, \ldots x_S)$.
- Uma *vizinhança* do **SNP** j é um subconjunto de $\{1, \ldots, S\} \setminus \{j\}$.
- $\mathfrak{D}_{S \times N}$ é a matriz de amostras.

Outline

Vizinhanças de influência

A *vizinhança de influência* de um **SNP** j é uma *vizinhança* V_j tal que:

$$P(X_j = x_j \mid X_k = x_k,\ k \neq j) = P(X_j = x_j \mid X_k = x_k,\ k \in V_j)$$

Obs: Consideramos apenas vizinhanças adjacentes centradas em j, com comprimento l para a esquerda e r para a direita..

*A pseudo-verossimilhança penalizada para o **SNP** j em função de uma vizinhança de tamanho l para a esquerda e r para a direita* é dada por:

$$\overline{L}_j^{\mathcal{D}}(l,r) = \sum_{\omega \in A^l} \sum_{\tau \in A^r} \sum_{a \in A} \left(N_j^{\mathcal{D}}(\omega, a, \tau) \log_{|A|} \left(\frac{N_j^{\mathcal{D}}(\omega, a, \tau)}{N_j^{\mathcal{D}}(\omega, \cdot, \tau)} \right) \right) - \frac{(|A|-1)}{2} |A|^{|\omega\tau|} . \log_{|A|}(n)$$

Vizinhanças de influência

A *vizinhança de influência* de um **SNP** j é uma *vizinhança* V_j tal que:

$$P(X_j = x_j \mid X_k = x_k,\ k \neq j) = P(X_j = x_j \mid X_k = x_k,\ k \in V_j)$$

Obs: Consideramos apenas vizinhanças adjacentes centradas em j, com comprimento l para a esquerda e r para a direita..

A *pseudo-verossimilhança penalizada para o* ***SNP*** j *em função de uma vizinhança de tamanho* l *para a esquerda e* r *para a direita* é dada por:

$$\overline{L}_j^{\mathcal{D}}(l,r) = \sum_{\omega \in A^l} \sum_{\tau \in A^r} \sum_{a \in A} \left(N_j^{\mathcal{D}}(\omega, a, \tau) \log_{|A|} \left(\frac{N_j^{\mathcal{D}}(\omega, a, \tau)}{N_j^{\mathcal{D}}(\omega, \cdot, \tau)} \right) \right) - \frac{(|A|-1)}{2} |A|^{|\omega\tau|} . \log_{|A|}(n)$$

Vizinhanças de influência

A *vizinhança de influência* de um **SNP** j é uma *vizinhança* V_j tal que:

$$P(X_j = x_j \mid X_k = x_k,\ k \neq j) = P(X_j = x_j \mid X_k = x_k,\ k \in V_j)$$

Obs: Consideramos apenas vizinhanças adjacentes centradas em j, com comprimento l para a esquerda e r para a direita..

A *pseudo-verossimilhança penalizada para o* ***SNP*** j *em função de uma vizinhança de tamanho* l *para a esquerda e* r *para a direita* é dada por:

$$\overline{L}_j^{\mathcal{D}}(l,r) = \sum_{\omega \in A^l} \sum_{\tau \in A^r} \sum_{a \in A} \left(N_j^{\mathcal{D}}(\omega, a, \tau) \log_{|A|} \left(\frac{N_j^{\mathcal{D}}(\omega, a, \tau)}{N_j^{\mathcal{D}}(\omega, \cdot, \tau)} \right) \right) - \frac{(|A|-1)}{2} |A|^{|\omega\tau|} . \log_{|A|}(n)$$

Pseudo-verossimilhança penalizada: resultados

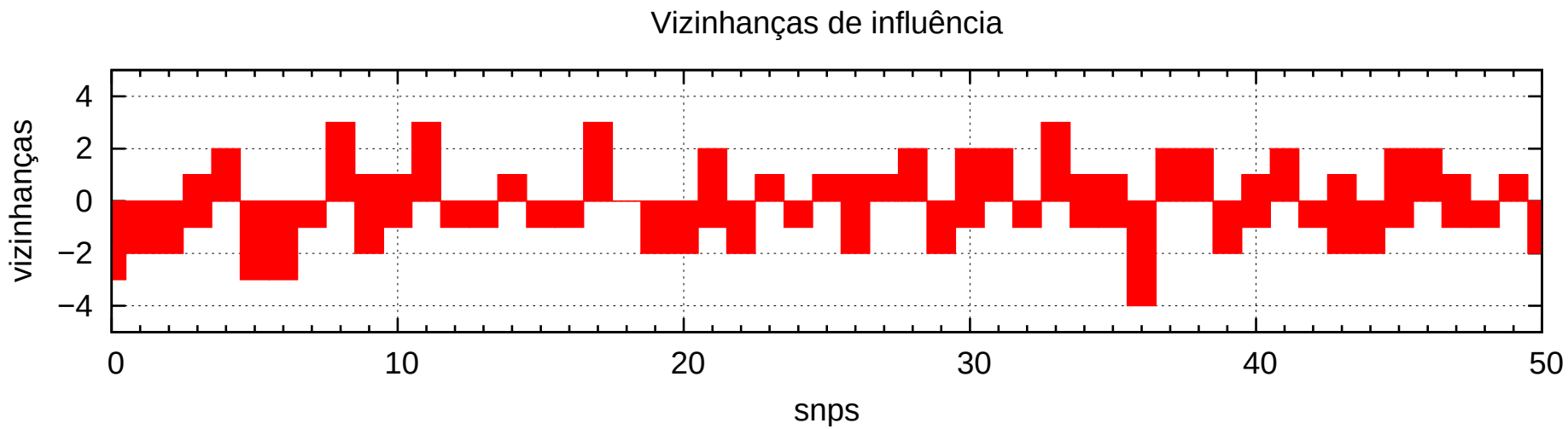


Figura: Vizinhanças de influência determinadas pela pseudo-verossimilhança penalizada.

Outline

Janelas de influência

Definição: Uma *janela de influência* é uma sequência $J \subseteq \{1, \ldots, S\}$ adjacente de **SNP**s tal que para todo $j \in J$, $V_j \subseteq J$.

Distância entre os grupos de caso e controle

- $\hat{P}_i(\omega) = \dfrac{N_{J_i}^{\mathfrak{D}_K}(\omega)}{|\mathfrak{D}_K|}$ e $\hat{Q}_i(\omega) = \dfrac{N_{J_i}^{\mathfrak{D}_L}(\omega)}{|\mathfrak{D}_L|}$
- $D_1(\hat{P}_i, \hat{Q}_i) = \sum_{\omega \in A^{|J_i|}} |\hat{P}_i(\omega) - \hat{Q}_i(\omega)|$
- $D_2(\hat{P}_i, \hat{Q}_i) = \max_{\omega \in A^{|J_i|}} |\hat{P}_i(\omega) - \hat{Q}_i(\omega)|$
- $D_3(\hat{P}_i, \hat{Q}_i) = \sum_{\omega \in A^{|J_i|}} \hat{P}_i(\omega) . \log \dfrac{\hat{P}_i(\omega)}{\hat{Q}_i(\omega)}$

Distância entre os grupos de caso e controle

- $\hat{P}_i(\omega) = \dfrac{N_{J_i}^{\mathfrak{D}_K}(\omega)}{|\mathfrak{D}_K|}$ e $\hat{Q}_i(\omega) = \dfrac{N_{J_i}^{\mathfrak{D}_L}(\omega)}{|\mathfrak{D}_L|}$
- $D_1(\hat{P}_i, \hat{Q}_i) = \sum_{\omega \in A^{|J_i|}} |\hat{P}_i(\omega) - \hat{Q}_i(\omega)|$
- $D_2(\hat{P}_i, \hat{Q}_i) = \max_{\omega \in A^{|J_i|}} |\hat{P}_i(\omega) - \hat{Q}_i(\omega)|$
- $D_3(\hat{P}_i, \hat{Q}_i) = \sum_{\omega \in A^{|J_i|}} \hat{P}_i(\omega).\log \dfrac{\hat{P}_i(\omega)}{\hat{Q}_i(\omega)}$

Distância entre os grupos de caso e controle

- $\hat{P}_i(\omega) = \dfrac{N_{J_i}^{\mathfrak{D}_K}(\omega)}{|\mathfrak{D}_K|}$ e $\hat{Q}_i(\omega) = \dfrac{N_{J_i}^{\mathfrak{D}_L}(\omega)}{|\mathfrak{D}_L|}$
- $D_1(\hat{P}_i, \hat{Q}_i) = \sum_{\omega \in A^{|J_i|}} |\hat{P}_i(\omega) - \hat{Q}_i(\omega)|$
- $D_2(\hat{P}_i, \hat{Q}_i) = \max_{\omega \in A^{|J_i|}} |\hat{P}_i(\omega) - \hat{Q}_i(\omega)|$
- $D_3(\hat{P}_i, \hat{Q}_i) = \sum_{\omega \in A^{|J_i|}} \hat{P}_i(\omega).\log \dfrac{\hat{P}_i(\omega)}{\hat{Q}_i(\omega)}$

Distância entre os grupos de caso e controle

- $\hat{P}_i(\omega) = \dfrac{N_{J_i}^{\mathfrak{D}_K}(\omega)}{|\mathfrak{D}_K|}$ e $\hat{Q}_i(\omega) = \dfrac{N_{J_i}^{\mathfrak{D}_L}(\omega)}{|\mathfrak{D}_L|}$
- $D_1(\hat{P}_i, \hat{Q}_i) = \sum_{\omega \in A^{|J_i|}} |\hat{P}_i(\omega) - \hat{Q}_i(\omega)|$
- $D_2(\hat{P}_i, \hat{Q}_i) = \max_{\omega \in A^{|J_i|}} |\hat{P}_i(\omega) - \hat{Q}_i(\omega)|$
- $D_3(\hat{P}_i, \hat{Q}_i) = \sum_{\omega \in A^{|J_i|}} \hat{P}_i(\omega).\log \dfrac{\hat{P}_i(\omega)}{\hat{Q}_i(\omega)}$

Outline

Distância entre distribuições: $D_1(\hat{P}_i, \hat{Q}_i)$

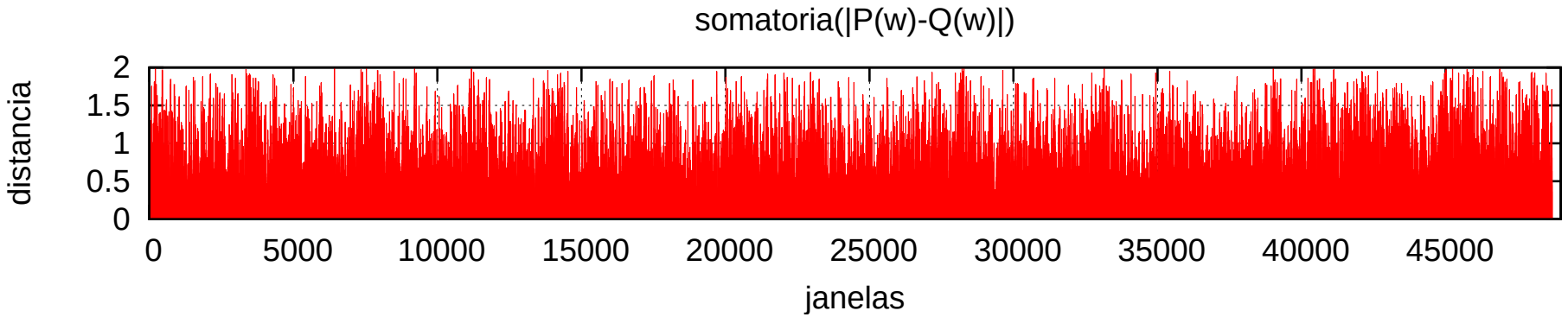


Figura: Distância entre indivíduos caso e controle, dada pela somatória das diferenças de frequência.

Distância entre distribuições: $D_2(\hat{P}_i, \hat{Q}_i)$

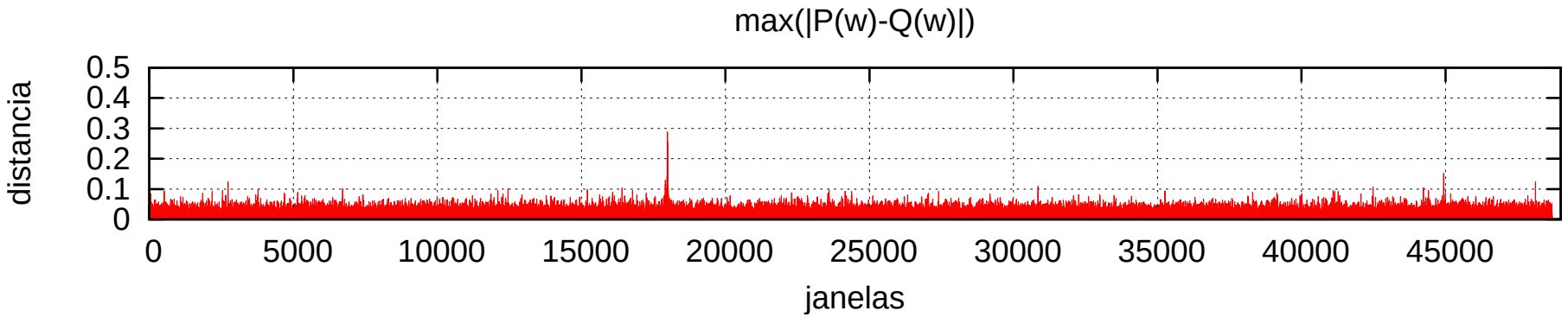


Figura: Distância entre indivíduos caso e controle, dada pela diferença máxima entre frequências.

Distância entre distribuições: $D_2(\hat{P}_i, \hat{Q}_i)$ (zoom)

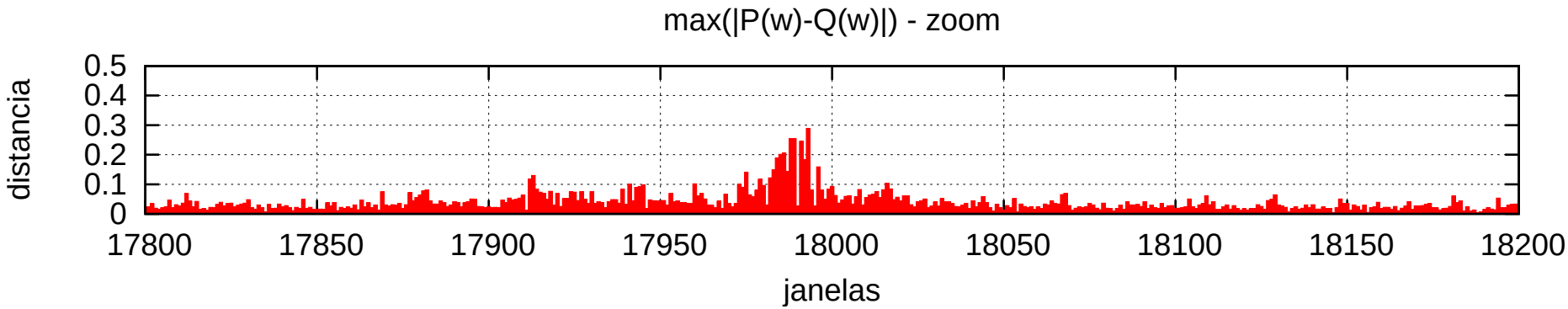


Figura: Distância entre indivíduos caso e controle, dada pela diferença máxima entre frequências (zoom).

Distância entre distribuições: $D_3(\hat{P}_i, \hat{Q}_i)$

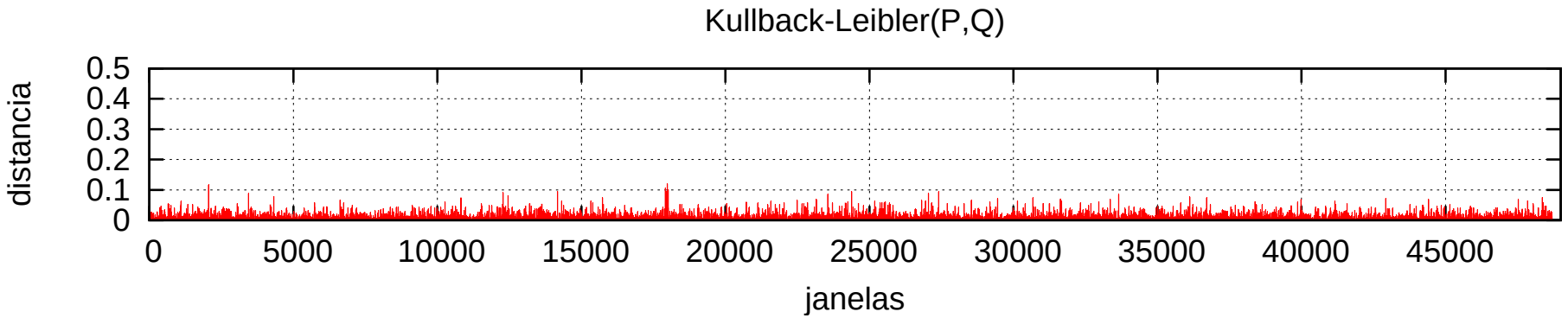


Figura: Distância entre indivíduos caso e controle, dada pela Divergência de Kullback-Leibler entre $\hat{P}_i$ e $\hat{Q}_i$.

Distância entre distribuições: $D_3(\hat{Q}_i, \hat{P}_i)$

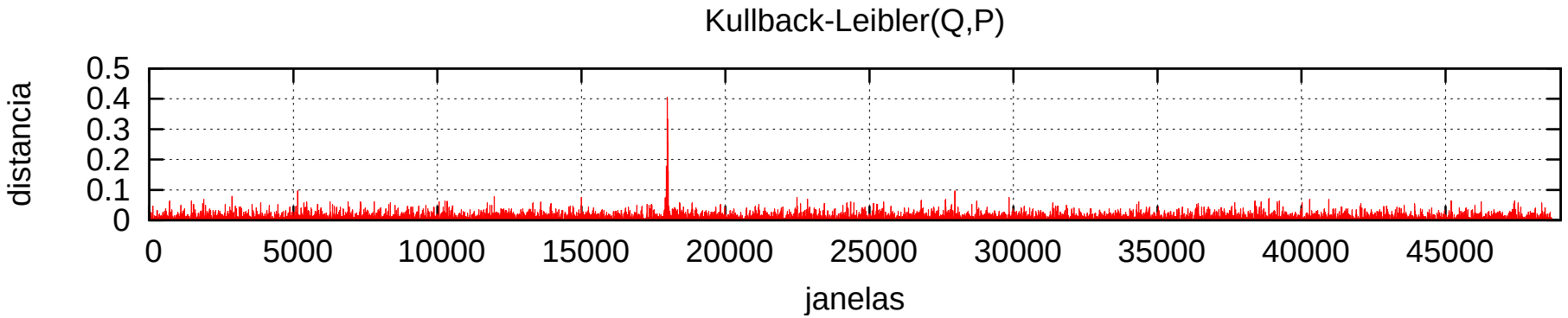


Figura: Distância entre indivíduos caso e controle, dada pela Divergência de Kullback-Leibler entre $\hat{Q}_i$ e $\hat{P}_i$.

Distância entre distribuições: $D_3(\hat{Q}_i, \hat{P}_i)$ (zoom)

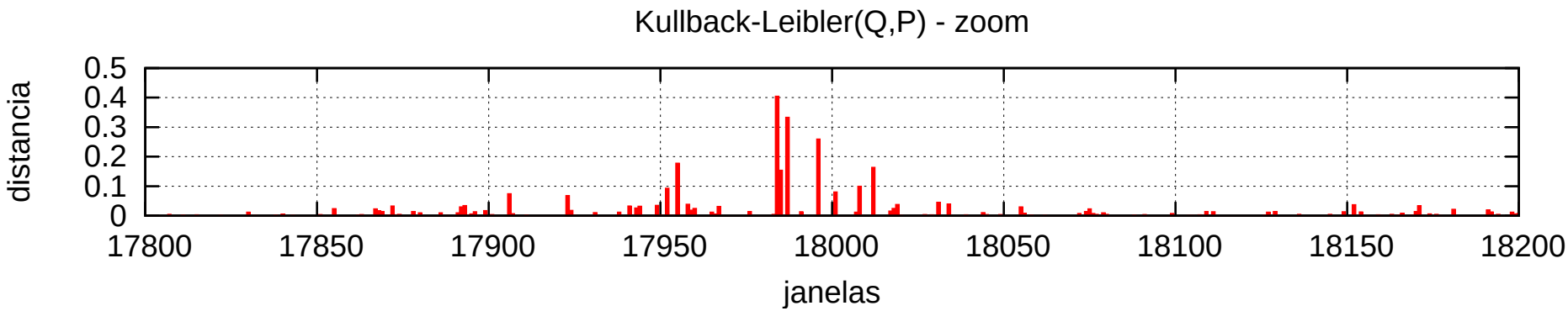


Figura: Distância entre indivíduos caso e controle, dada pela Divergência de Kullback-Leibler entre $\hat{Q}_i$ e $\hat{P}_i$ (zoom).

Outline

Resumo

- É possível determinar regiões de influência entre **SNP**s adjacentes.
- Estas regiões evidenciam diferenças entre indivíduos com fenótipos diferentes[3].

- A fazer:
  - Descobrir outros estimadores para janelas de influênca.
  - Determinar quais medidas de distância são úteis e por quê.
  - Integrar com outros métodos de análise de DNA utilizando filtragens e plataformas já existentes.

Outline

Bibliografia

[1] Julian Besag. Statistical analysis of non-lattice data. *The Statistician*, 24(3):179–195, 1975.

[2] Helen Pearson. Genetics: what is a gene? *Nature*, 441(7092):398–401, May 2006.

[3] Robert M. et al. Plenge. Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nature Genetics*, 8(39):1477–1482, 2007.

[4] D. G. Wang, J. B. Fan, C. J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, L. Hsie, T. Topaloglou, E. Hubbell, E. Robinson, M. Mittmann, M. S. Morris, N. Shen, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T. J. Hudson, R. Lipshutz, M. Chee, and E. S. Lander. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, 280(5366):1077–1082, May 1998.