RAPID Detection of Gene–Gene Interactions in Genome-Wide Association Studies

Bioinformatics, Vol. 26 no. 22 2010, pages 2856–2862

Lauris kaplinski
Bioinformatics Journal Club 08/11/2010

Motivation

- In complex disorders, independently evolving locus pairs might interact to confer disease susceptibility, **with only a modest effect at each locus**.
- With genome-wide association studies on large cohorts, testing all pairs for interaction confers a heavy computational burden, and a **loss of power** due to large Bonferroni like corrections.
- Correspondingly, limiting the tests to pairs that show marginal effect at either locus, also has reduced power.

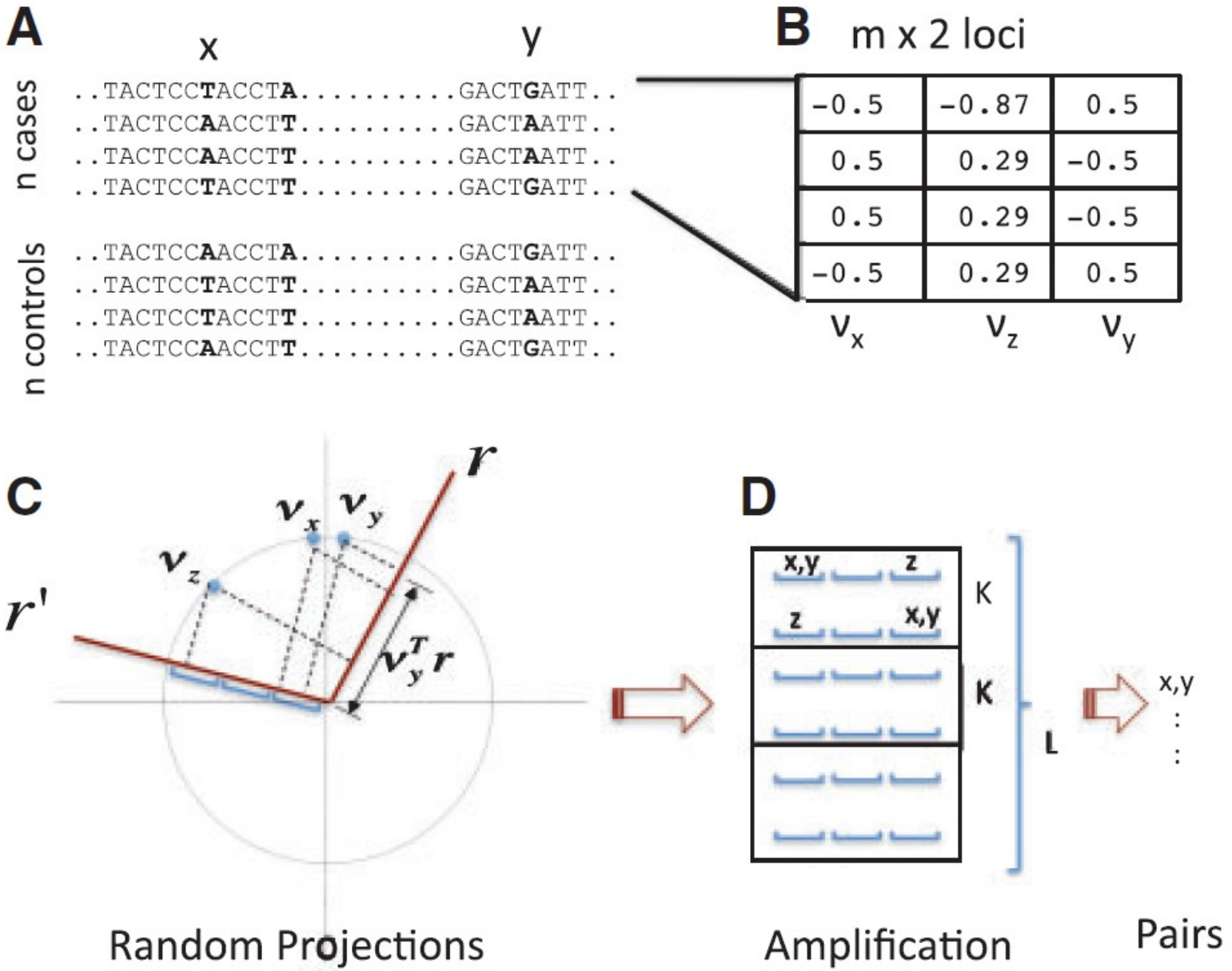
GWAS

- Detecting k-locus interactions in GWAS on large populations is computationally and statistically challenging, even when $k=2$.
- A test involving all pairs of m markers, with a case-control population of n individuals, involves $O(nm^2)$ computations.
- A straightforward (Bonferroni-like) correction for the multiple tests would result in significant loss of sensitivity.
- The **marginal effects of single loci** may be very small

The Idea

- A mathematical transformation that maps ‘statistical correlation between locus pairs’ to ‘**distance between two points in a Euclidean space**’.
- This enables the use of geometric properties to identify proximal points (correlated locus pairs), without testing each pair explicitly.
- The speed of the test allows for correction using permutation-based tests.
- **Rapid Pair IDentificator** is first-stage filtering tool for genome-wide analysis (as a bonus can do second stage association tests too)

Example of missing marginal effect



Rapid Pair IDentificator

- at most $\tau_1 \approx m^{1.07}$ tests
  - Naive implementation is m^2
- no more than $\tau_2 \approx m^{1.07}\ ln(1/\varepsilon)$ false positive pairs
  - NB! The total number of pairs is m^2
- n – the number of individuals
- m – the number of markers
- ε – false negative rate

Transformation

- Each locus x is described by a vector $x \in \{0,1\}^n$ of allelic values
- The case-control status of the individuals is described by a vector $d \in \{0,1\}^n$
- The null hypothesis - no association of a pair of loci x,y against d can be tested using a χ^2 test on a $2 \times 2 \times 2$ table
- If x,y,d jointly associate, then at least one of the following has association:
  - Marginal association between x and d, described by $\chi^2_{x,d}$
  - Marginal association between y and d, described by $\chi^2_{y,d}$
  - Association between x, and y, when the individuals are drawn only from cases. ($\chi^2_{x,y}$ is high for cases)
  - Association between x, and y, for controls ($\chi^2_{x,y}$ is high for controls)

2-way associations done fast

- Let Px denote the fraction of individuals in the population with allele 1 at locus x
- For a∈{0,1}, define

$$v_x(a) = \frac{a - P_x}{\sqrt{n}\sqrt{P_x(1-P_x)}}$$

- A vector $v_x = [v_x(x_1) v_x(x_2) \ldots v_x(x_n)]$ maps the allelic values at locus x for all n individuals onto a unit vector v_x
- Define the distance between 2 loci as

$$\mathrm{dist}(\boldsymbol{v_x}, \boldsymbol{v_y}) = \min(\| \boldsymbol{v_x} - \boldsymbol{v_y} \|, \| \boldsymbol{v_x} + \boldsymbol{v_y} \|)$$

The correlation

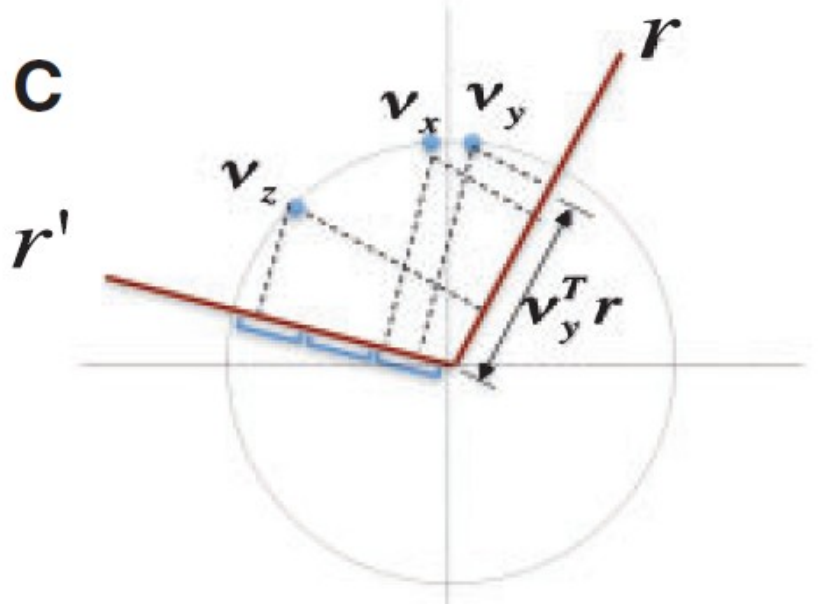
$$\mathrm{dist}(\boldsymbol{v_x}, \boldsymbol{v_y}) = \sqrt{2 - 2\sqrt{\chi^2_{x,y}/n}}$$

$$\mathrm{dist}(\boldsymbol{v_x}, \boldsymbol{v_y}) \leq \theta = \sqrt{2 - 2\sqrt{t/n}}$$

- Thus, we can transform the statistical problem of identifying interacting locus pairs $\{(x,y) : \chi_2 x,y \geq t\}$ (where t is a threshold) into a geometric problem of computing proximal vectors
- But finding distances is also O(n2) problem...

Locality Sensitive Hashing (LSH)

- To identify locus pairs (x,y) for which $dist(v_x, v_y) \leq \theta$, we choose a random unit vector r, and project each of the points onto r



- Then assign each locus to bin accoring to chosen value B

$$\text{HASH}(x, \boldsymbol{r}, B) = \left\lfloor \frac{|\boldsymbol{v_x} \cdot \boldsymbol{r}|}{B} \right\rfloor$$

Amplification of bias

- The bin size B is chosen to ensure that if $dist(vx,vy)<\theta$, then loci x,y fall in the same bin with high probability (denoted by $f1$).
- If x,y are non-interacting ($dist(vx,vy)$ is large), they fall into the same bin with a much lower probability (denoted by $f2<f1$)
- We have to amplify $f1/f2$ (i.e. (1 - false negative) / false positive ratio)
- $1-\varepsilon$ is the desired power (the fraction of true interacting pairs that are retained for a second-stage scoring)
- We run the hashing procedure LK times, and select only those pairs that fall in the same bin all K times, in at least one of the L iterations

$$K = \frac{\ln m}{\ln(1/f_2)} \qquad L = f_1^{-K} \ln(1/\varepsilon)$$

- Rapid will output a high fraction ($1-\varepsilon$) of all interacting pairs, but at most $m^c \ln(1/\varepsilon)$ non-interacting pairs

$$c = 1 + \frac{\ln(1/f_1)}{\ln(1/f_2)}$$

Processing genotypes

- Genotypes do not map immediately to required vectors
- If genotypes in locus i are *aa, aA* and *AA*, then map those to two alternative haplotypes X0 and X1

$$X_0[i] = \begin{cases} 0 & \text{if } x[i] = \text{`}aa\text{'} \text{ or } x[i] = \text{`}aA\text{'} \\ 1 & \text{otherwise} \end{cases}$$

$$X_1[i] = \begin{cases} 0 & \text{if } x[i] = \text{`}aa\text{'} \\ 1 & \text{otherwise} \end{cases}$$

- For any pair x,y of loci, if x,y are interacting ($\chi^2_{x,y}$ is high), then one of the four values $\chi^2_{X\,i,Yj}$ is high as well

Results

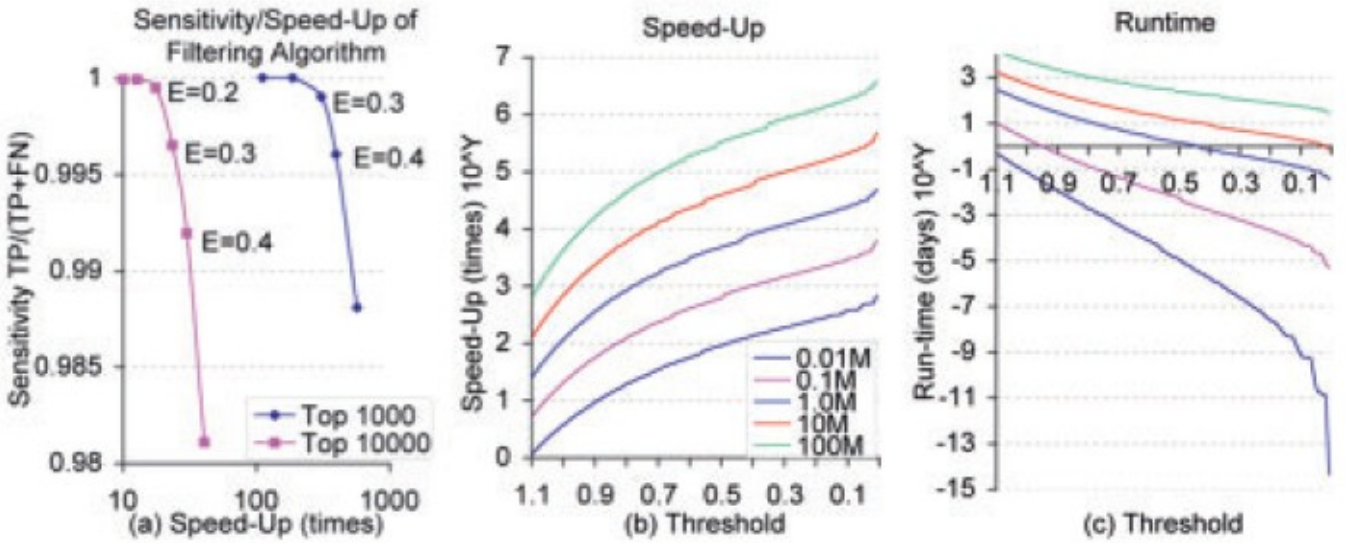


Fig. 2. Speed Speed sensitivity trade-offs in RAPID. The trade-offs are computed as a function of user-defined parameters θ, ε. (**a**) Speed-up versus sensitivity trade-offs are measured on a dataset of 50000 WTCCC control SNPs. Different thresholds θ are chosen to filter the top 1000 ($\theta \leq 0.1$), 10000 ($\theta \leq 0.4$) of SNP pairs, respectively. $\varepsilon \in \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$. (**b** and **c**) Runtime versus θ. Strongly interacting pairs ($\theta \leq 0.5$) can be identified on large datasets ($1M$ SNPs, 3000 genotypes) with 95% sensitivity in a few hours on a commodity PC. All experiments were run on a 1.8 GHz, 16 GB RAM, Linux machine. Axes have logarithmic scale.

Conclusion

- Many orders of magnitude speed-up in filtering stage without losing sensitivity
- Is capable of detecting true bilocal associations without any effect of single locus
- In theory should be possible to extend for more than 2 loci
- Because much faster and accurate analysis, it will be possible to use more precise 2-nd stage algorithms