Relatedness detection with an embedded likelihood ratio test

Jesse M. Rodriguez[1], Sivan Bercovici[2], Serafim Batzoglou[2]

1. Biomedical Informatics Program 2. Department of Computer Science

Background

- Knowing when two or more people are related is important for medical and population genomics. For example, genotype-phenotype studies assume that all individuals in the study are unrelated. Distant, previously-unknown relatives in a study can convolute the results and decrease study power.
- Methods to detect relatedness (aka IBD, identity-by-descent) are important for detecting previously unknown relationships between individuals in a population or cohort.
- A major challenge is the fact that relatedness is a property between two individuals, which means that one must consider all possible pairs of people that can be related to each other.
- Therefore, because there are O(N^2) pairs, a method inferring relatedness must
  - have a very low false positive rate, and
  - be very computationally efficient

Aims

To develop ***Parente***, a method to detect relatedness that has the following properties:

1. Highly accurate with a **low false positive rate**
2. Sensitive to detect very small related segments.
3. Computationally efficient, able to analyze large cohorts of **tens to hundreds of thousands** of individuals.
4. Conducive to sub-linear retrieval algorithms (future work).

Methods

- We define a set of SNP **windows** (**w**) that each contain a few SNPs (5-20).
- We define a set of **blocks** (**B**) that contain many windows. These blocks span several megabases (eg 4 cM).
- We have data that we used to generate a training data set, **T**, that contains related (T_R) and unrelated ($T_{\bar{R}}$) pairs of individuals.
- For each window, we compute a **likelihood-ratio *score***, (L) as follows:

$$L(w)=\frac{p(g_w^1,g_w^2\mid R)}{p(g_w^1,g_w^2\mid \bar{R})}$$

- where the following values indicate the likelihood of the data given that the **observed genotypes** (**g**) in the window were produced by two **related** (R) individuals that share one haplotype (ie, there are 3 unique haplotypes between the two of them), or by two **unrelated** (R) individuals that have 4 unique haplotypes between them.

$$p(g_w^1,g_w^2\mid R)=\sum_{h_1,h_2,h_3} p(g^1\mid h_1,h_2)p(g^2\mid h_1,h_3)p(h_1)p(h_2)p(h_3)$$

$$p(g_w^1,g_w^2\mid \bar{R})=\sum_{h_1,h_2,h_3,j_4} p(g_w^1\mid h_1,h_2)p(g_w^2\mid h_3,h_4)p(h_1)p(h_2)p(h_3)p(h_4)$$

- During training, we look at the related and unrelated pairs, and fit a **Gaussian distribution** to the distribution of *L(w)* scores in each set. This lets us calculate the probability of observing a particular score at a window, assuming it was generated from previously-observed related or unrelated pairs:

$$p\left(L(w)\mid T_R\right) \quad \text{and} \quad p\left(L(w)\mid T_{\bar{R}}\right)$$

- What we really care about is longer stretches of relatedness (not just the 5 SNPs in a window), so we compute the **embedded likelihood-ratio** score for an entire block of windows:

$$\boxed{eLR(B)=\prod_{w\in B}\frac{p\left(L(w)\mid T_R\right)}{p\left(L(w)\mid T_{\bar{R}}\right)}=\prod_{w\in B}\frac{p\left(\dfrac{p(g_w^1,g_w^2\mid R)}{p(g_w^1,g_w^2\mid \bar{R})}\mid T_R\right)}{p\left(\dfrac{p(g_w^1,g_w^2\mid R)}{p(g_w^1,g_w^2\mid \bar{R})}\mid T_{\bar{R}}\right)}}$$

Results

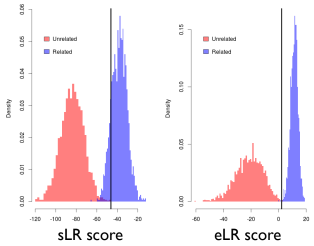
The eLR statistic

- To provide insight into why we use the embedded likelihood ratio statistic, we also computed the "standard" likelihood ratio stastistic (sLR) for each block of windows as follows:

$$sLR(B)=\prod_{w\in B} L(w)$$

- We compared eLR to sLR in terms of the score histograms of known related and unrelated individuals. An ideal score would have perfect separation between the scores of unrelated individuals and the scores of related individuals:

Score histograms for sLR and eLR

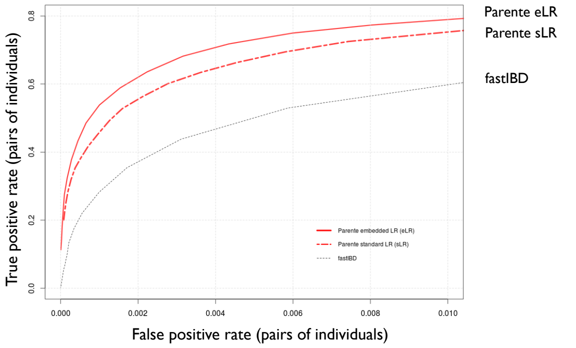


The vertical lines indicate the highest observed score for unrelated individuals. Note that if this line were used as a classification threshold, we expect to find a higher true positive rate for this block.

Performance versus leading methods

- We compared the sensitivity and specificity of our method, Parente using the eLR statistic as well as the sLR statistic), against fastIBD, the most accurate high-speed detection tools to date.

Performance curves for Parente and fastIBD


Parente eLR
Parente sLR
fastIBD
True positive rate (pairs of individuals)
False positive rate (pairs of individuals)
Parente embedded LR (eLR)
Parente standard LR (sLR)
fastIBD