BIOINFORMATICS LETTER TO THE EDITOR Vol. 30 no. 5 2014, pages 743–745
doi:10.1093/bioinformatics/btt588

Genetics and population analysis Advance Access publication October 15, 2013

Comment on 'Bayesian parentage analysis with systematic accountability of genotyping error, missing data and false matching'

Eric C. Anderson[1,2,*] and Thomas C. Ng[3]

[1]Fisheries Ecology Division, Southwest Fisheries Science Center, National Marine Fisheries Service, NOAA, 110 Shaffer Road, Santa Cruz, CA 95060, [2]Department of Applied Math and Statistics and [3]Department of Biomolecular Engineering, University of California, Santa Cruz, CA, USA

Associate Editor: Jeffrey Barrett

Summary: We show the software SOLOMON is improved by using the likelihood ratio instead of an *ad hoc* statistic.

Code: github.com/eriqande/solidmon/releases/tag/v0.1-bioinformatics

Contact: eric.anderson@noaa.gov

Received on August 7, 2013; revised on September 23, 2013; accepted on October 4, 2013

1 INTRODUCTION

In a recent *Bioinformatics* article, Christie, Tennessen and Blouin (hereafter CTB) present the software SOLOMON for parentage inference from genotype data (Christie *et al.*, 2013). They propose a method of estimating the number of true parent-offspring pairs amongst all the candidate pairs in a dataset and show it provides better results than the program CERVUS (Marshall *et al.*, 1998) when the fraction of sampled candidates (an input to CERVUS) is misspecified.

Here we show that the performance of SOLOMON can be improved by making full use of the genotype data, replacing CTB's statistic with the likelihood ratio as prescribed by a well-established literature (Marshall *et al.*, 1998; Meagher and Thompson, 1987). This is particularly beneficial when using single nucleotide polymorphisms (SNPs) chosen to have minor allele frequency near 0.5, as desired for SNP-based pedigree reconstruction (Anderson and Garza, 2006).

2 METHODS

We start with an overview of SOLOMON's methodology, adopting new notation where it permits a more complete succinct description.

2.1 SOLOMON's methodology

Assume L unlinked loci used to identify the true parents of n_2 diploid offspring from amongst n_1 diploid candidate parents. CTB focus first on the case where all candidate parents are of the same sex and the goal is to identify true parent–offspring pairs from amongst the $n_1 n_2$ candidate pairs. Let z_j count the number of loci not compatible with Mendelian inheritance (barring a genotyping error) between the candidate parent and offspring of pair j. The method of CTB then proceeds in four steps.

Step 1: Partition all $n_1 n_2$ pairs into disjoint sets $\mathcal{P}_0, \ldots, \mathcal{P}_L$, where $\mathcal{P}_i$ includes all pairs j such that $z_j = i$. Denote by $\#\{\mathcal{P}_i\}$ the cardinality of $\mathcal{P}_i$.

Step 2: Estimate π_i, the unknown fraction of pairs in each set $\mathcal{P}_i$ that are unrelated pairs (i.e. the two members of the pair are unrelated), with

$$\tilde{\pi}_i = \min\left\{1, \frac{n_1 n_2 P(z_j = i \mid \text{pair } j \text{ is unrelated})}{\#\{\mathcal{P}_i\}}\right\}. \tag{1}$$

The probability in the numerator is computed by simulation from the estimated allele frequencies in the population. CTB denote $\tilde{\pi}_i$ by $\Pr(\phi)$, and refer to it as the prior probability that a pair in $\mathcal{P}_i$ is unrelated.

CTB define a statistic λ to describe allele sharing between a pair of individuals. Let $y_{s,\ell}$ and $y_{k,\ell}$ be genotypes at locus ℓ in the putative sire s and the putative kid k, respectively. Then,

$$\lambda(y_{s,\ell}, y_{k,\ell}) = \begin{cases} 0 & \text{if } s \text{ and } k \text{ share no alleles at locus } \ell, \\ h & \text{if } s \text{ and } k \text{ share at least one allele and } h \text{ is the} \\ & \text{index of the least frequent of the shared alleles,} \end{cases}$$

i.e. if a single allele is shared, h is its index, if two alleles are shared, then h is the index of the shared allele that is the least frequent (in the population) of the two shared alleles.

Step 3: Calculate for each pair j in $\mathcal{P}_i$ with $\pi_i < 1$ the quantities:

$$T_j^{\mathrm{U}} = \prod_{\ell \in \mathcal{L}} P(\lambda(y_{s,\ell}, y_{k,\ell}) | R_{sk} = \mathrm{U}) \tag{2}$$

$$T_j^{\mathrm{PO}} = \prod_{\ell \in \mathcal{L}} f(\lambda(y_{s,\ell}, y_{k,\ell})) \tag{3}$$

where s and k are the members of the pair, R_{sk} denotes the assumed relationship between s and k, U and PO denote 'unrelated' and 'parent–offspring', respectively, $\mathcal{L}$ is the set of loci at which s and k share at least one allele, and $f(\lambda(y_{s,\ell}, y_{k,\ell}))$ is the relative frequency in the population of the allele with index $\lambda(y_{s,\ell}, y_{k,\ell})$. To compute T_j^{U}, CTB approximate $P(\lambda(y_{s,\ell}, y_{k,\ell})|R_{sk} = \mathrm{U})$ by Monte Carlo simulation.

Step 4: Combine $\tilde{\pi}_i$ and λ to compute a 'posterior probability' of parentage for each pair in each $\mathcal{P}_i$ with $\pi_i < 1$. Let X_i^{U} represent a random variable defined as the product in (2) when s and k are unrelated individuals simulated conditional on sharing at least one allele at every locus, and $\mathcal{L}$ is a randomly chosen set of $L - i$ loci at which s and k share alleles. Let X_i^{PO} be a random variable that is the product in (3) when s and k are a simulated true parent and offspring. CTB define the 'posterior probability of parentage' for a pair j in $\mathcal{P}_i$, with observed values of T_j^{U} and T_j^{PO}, as:

$$Q(\mathrm{PO}|\lambda) = \frac{(1-\pi_i)P(X_i^{\mathrm{PO}} \geq T_j^{\mathrm{PO}})}{\pi_i P(X_i^{\mathrm{U}} \geq T_j^{\mathrm{U}}) + (1-\pi_i)P(X_i^{\mathrm{PO}} \geq T_j^{\mathrm{PO}})}. \tag{4}$$

[*]To whom correspondence should be addressed.

$Q(\mathrm{PO}|\lambda)$, more a P-value than a posterior probability, is computed by simulation and used to accept or reject pairs as being parent–offspring pairs.

2.2 Analysis

The statistics T_j^{PO} and T_j^{U} reflect the fact that 'pairs sharing rare alleles are much more likely to be true parent–offspring pairs' (Christie *et al.*, 2013, p. 726); however, they do not reflect the fact that 'a homozygous male that is compatible with a given offspring really does have a higher likelihood of producing it than a heterozygous male sharing one allele with the offspring' (Jones and Ardren, 2003, p. 2512). As such, SOLOMON is not making full use of the data; T_j^{PO} and T_j^{U} are not sufficient statistics.

An example demonstrates this. Consider $L = 60$ biallelic loci, each with one allele at frequency of 0.501 and the other at 0.499. Clearly, little information is gained by knowing whether the first or the second allele is shared in a candidate parent–offspring pair, so, in effect, the statistic $\lambda(y_{s,\ell}, y_{k,\ell})$ is merely recording whether or not at least one allele is shared. An unrelated pair shares at least one allele at every locus with probability close to $(1-\frac{1}{8})^{60} = 3.3\times10^{-4}$. Imagine that $n_1 = n_2 = 500$ and $\#\{\mathcal{P}_0\} = 165$. By (1) we have $\tilde{\pi}_0 = (3.3\times10^{-4}\times500^2)/165 = 0.5$, so we expect that half of the pairs in $\mathcal{P}_0$ are true parent–offspring pairs. Those true pairs cannot, however, be identified by CTB's method because the observed values T_j^{U} and T_j^{PO} are effectively identical for every pair j in $\mathcal{P}_0$, whether it is a true parent–offspring pair or not.

We implemented CTB's approach, but based it on the likelihood ratio:

$$\mathrm{LR}(\boldsymbol{y}_s, \boldsymbol{y}_k) = \log\left(\frac{\prod_{\ell\in\mathcal{L}} P(y_{s,\ell}, y_{r,\ell}|R_{s,k} = \mathrm{PO})}{\prod_{\ell\in\mathcal{L}} P(y_{s,\ell}, y_{r,\ell}|R_{s,k} = \mathrm{U})}\right)$$

Kids are assigned to the candidate sire with highest $\mathrm{LR}(y_s, y_k)$. Confidence in those assignments can still be assessed with a P-value, $Q^*(\mathrm{PO}|y_s, y_k)$, like that in (4), by substituting $\mathrm{LR}(y_s, y_k)$ for both T_j^{PO} and T_j^{U} in (4) and redefining X_i^{U} and X_i^{PO} to be the random variable $\mathrm{LR}(y_s, y_k)$ for an s and k drawn randomly from the population conditional on s and k being either unrelated or parental, respectively, and sharing at least one allele at a randomly determined $L-i$ loci. We call this the 'likelihood ratio (LR) approach'.

To compare the performance of the LR approach to SOLOMON, we simulated 20 datasets under two different scenarios with no genotyping error: (i) $n_1 = n_2 = 500$, number of true pairs $= 83$ and $L = 60$ biallelic loci with equifrequent alleles; and (ii) $n_1 = n_2 = 200$ with 50 true pairs, and $L = 10$ loci each with 10 alleles and with the frequency of allele $v = 1, \ldots, 10$, proportional to $1/v$. π_0 is expected to be 0.50 and 0.48 for the scenarios, respectively. Each dataset was analyzed using the LR approach and SOLOMON, and accuracy was compared graphically by plotting the receiver-operating characteristic (ROC) curve (Green and Swets, 1966) for each dataset.

3 RESULTS AND CONCLUSIONS

Figure 1a shows that, as predicted, SOLOMON's criterion for parentage contains no extra information beyond Mendelian incompatibility in data scenario 1, with equifrequent alleles. The ROC curves indicate SOLOMON does not rank true pairs any higher than false pairs in $\mathcal{P}_0$. The LR approach performs considerably better. Figure 1b shows that on data with 10 alleles at frequencies that are far from equal (scenario 2), the LR approach still outperforms SOLOMON, as expected.

Our implementation of the LR approach, using 10^5 simulation replicates to approximate $Q^*(\mathrm{PO}|y_s, y_k)$, requires roughly 0.1 and 1 min for each dataset from scenarios 1 and 2, respectively, on a single core from a Mac Pro running at 2.8 GHz. SOLOMON, using the default 1000 replicates, requires 16.6 and 225 min.

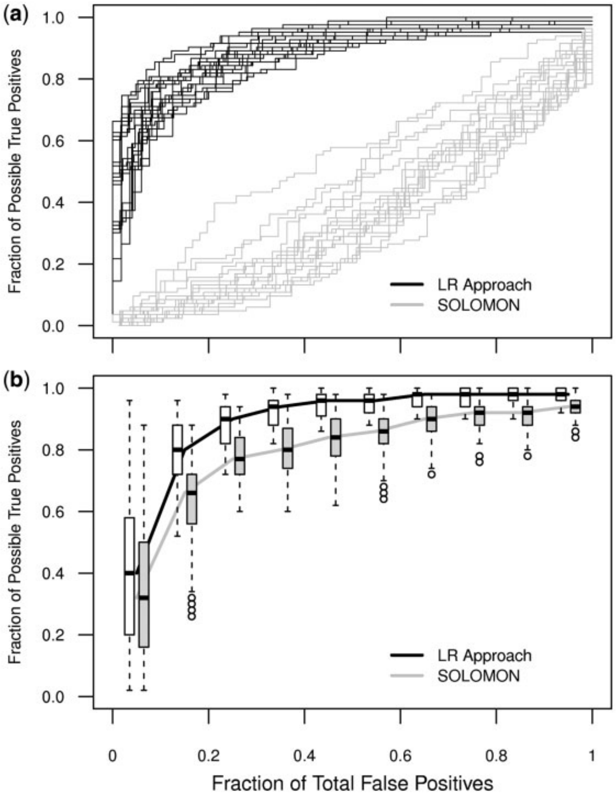

Fig. 1. ROC curves for the LR approach and SOLOMON. (**a**) Data scenario 1 (SNPs). (**b**) Data scenario 2 (microsatellites)—because individual ROC curves are overlapping, we plot the median as black and gray lines and depict variability with boxplots (white = LR approach; gray = SOLOMON)

CTB propose an interesting way of accounting for the unknown sampling fraction of parents, and we appreciate their efforts to make approachable software. We recommend SOLOMON be updated to use the likelihood instead of λ, so as to make it more accurate, faster and better suited for general use. In the future it will be interesting to compare how SOLOMON performs against methods that estimate the sampling fraction directly from the genotype probabilities (Koch *et al.*, 2008; Nielsen *et al.*, 2001), rather than formulating a 'prior' based on Mendelian incompatibilities.

ACKNOWLEDGEMENTS

The authors acknowledge helpful comments from Mark Christie and two anonymous referees.

Funding: T.N. was partially supported with funds from the US Section of the Chinook Technical Committee of the International Pacific Salmon Commission.

Conflict of Interest: none declared.

REFERENCES

Anderson,E.C. *et al.* (2006) The power of single nucleotide polymorphisms for large-scale parentage inference. *Genetics*, **172**, 2567–2582.

Christie,M.R. *et al.* (2013) Bayesian parentage analysis with systematic accountability of genotyping error, missing data and false matching. *Bioinformatics*, **29**, 725–732.

Green,D.M. *et al.* (1966) *Signal Detection Theory and Psychophysics*. John Wiley & Sons Inc., New York.

Jones,A.G. *et al.* (2003) Methods of parentage analysis in natural populations. *Mol. Ecol*., **12**, 2511–2523.

Koch,M. *et al.* (2008) Pedigree reconstruction in wild cichlid fish populations. *Mol. Ecol*., **17**, 4500–4511.

Marshall,T.C. *et al.* (1998) Statistical confidence for likelihood-based paternity inference in natural populations. *Mol. Ecol*., **7**, 639–655.

Meagher,T.R. *et al.* (1987) Analysis of parentage for naturally established seedlings within a population of *Chamaelirium luteum* (Liliaceae). *Ecology*, **68**, 803–812.

Nielsen,R. *et al.* (2001) Statistical approaches to paternity analysis in natural populations and applications to the North Atlantic humpback whale. *Genetics*, **157**, 1673–1682.