A two-phase strategy for detecting recombination in nucleotide sequences

Cheong Xin Chan*, Robert G. Beiko†, Mark A. Ragan*

*ARC Centre in Bioinformatics and Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia.
†Faculty of Computer Science, Dalhousie University, 6050 University Avenue, Halifax, Nova Scotia, Canada B3H 1W5.

ABSTRACT

Genetic recombination can produce heterogeneous phylogenetic histories within a set of homologous genes. Delineating recombination events is important in the study of molecular evolution, as inference of such events provides a clearer picture of the phylogenetic relationships among different gene sequences or genomes. Nevertheless, detecting recombination events can be a daunting task, as the performance of different recombination-detecting approaches can vary, depending on evolutionary events that take place after recombination. We previously evaluated the effects of post-recombination events on the prediction accuracy of recombination-detecting approaches using simulated nucleotide sequence data. The main conclusion, supported by other studies, is that one should not depend on a single method when searching for recombination events. In this paper, we introduce a two-phase strategy, applying three statistical measures to detect the occurrence of recombination events, and a Bayesian phylogenetic approach to delineate breakpoints of such events in nucleotide sequences. We evaluate the performance of these approaches using simulated data, and demonstrate the applicability of this strategy to empirical data. The two-phase strategy proves to be time-efficient when applied to large datasets, and yields high-confidence results.

KEYWORDS: Recombination detection, sequence analysis, evolution and phylogenetics, bioinformatics

1 INTRODUCTION

Genetic recombination is the process in which an external fragment of genetic material is integrated into a recipient sequence. Recombination not only plays a critical role in the completion of DNA replication and DNA repair in prokaryotes [1], but also ensures proper pairing and correct segregation of chromosomes during meiosis in eukaryotes, a process essential for maintaining genome integrity throughout cell division [2, 3]. The process of recombination, or genetic transfer in general, contributes to genetic diversity and inconsistent phylogenetic signals across genomes of different species. Therefore, elucidating genetic transfers resulting from recombination events in biological sequences will enhance our understanding of the role selective forces play in shaping genomes. However, detecting recombination is not without problems. When the recombining sequences are very similar to each other, or when subsequent evolution has obscured the recombination signal, detecting recombination can be difficult. The scenario is more complicated when there are overlapping recombination events on the sequences, or when recombination events occurring constitutively one after another in regions within close proximity of each other.

A number of approaches are available for detecting recombination events in biological sequences. These approaches can be classified based on the different algorithms used: distance-based [4, 5], substitution distribution-based [6], compatibility-based [7, 8], and phylogenetic-based [9]. New approaches are also being developed such as one adopting a genetic algorithm [10] and one combining the use of two different statistical tests [11]. A number of these approaches were reviewed for their performance; the effects of sequence divergence, amount of recombination, and subsequent substitutions after the recombination event were examined [12, 13, 14, 15]. All these studies agree that recombination is easier to detect when the event involves sequences that are divergent, and that approaches based on compatibility and substitution distribution showed higher prediction accuracy than the conventional phylogenetic approach. While the conventional phylogenetic approaches were found to perform poorly compared to the other approaches examined, Bayesian phylogenetic approaches were found to show higher accuracy than all the other approaches [15]. The importance of not depending on a single method in isolation when detecting recombination was well demonstrated in these studies, as the different approaches have different advantages and drawbacks. A good method for detecting occurrences of recombination might not be good in identi-

Email: Cheong Xin Chan `c.chan@imb.uq.edu.au`, Robert G. Beiko `beiko@cs.dal.ca`, Mark A. Ragan `m.ragan@imb.uq.edu.au`

fying the breakpoints of such events, and vice versa. Using different approaches in succession can increase our confidence in detecting an event that has indeed taken place within the defined breakpoints at the sequences. When dealing with large datasets, it is desirable to use a quick method in first-pass screening for the presence of recombination in the datasets, then a more-accurate (albeit slower) method to delineate the recombination breakpoints among the positives [15]. Here we present a two-phase strategy in detecting recombination events in nucleotide sequences, and highlight its application in large-scale datasets.

2 TWO-PHASE APPROACH FOR DETECTING RECOMBINATION

The strategy involves two phases: (a) detecting occurrences of recombination events in the sequence dataset; and (b) identifying breakpoints of such events. The first phase involves a quick, first-pass screening for possible recombination events using a number of statistical measures for evaluating phylogenetic discrepancy across the set of sequences. Once significant phylogenetic discrepancy is detected, the second phase involves a slower but more-accurate Bayesian phylogenetic approach to delineate recombination breakpoints in sequence data.

2.1 Phase I: Detecting occurrences of recombination

To screen for occurrences of recombination events, we used PhiPack [8] to evalute three different statistical measures: neighbour similarity score (NSS) [7], maximal chi-squared (MaxChi) [16] and pairwise homoplasy index (PHI) [8]. These statistics measure the significance of phylogenetic discrepancy across sites in an alignment, each test yielding a p-value. Both NSS and PHI are based on compatibility of parsimoniously informative sites, whereas MaxChi is based on substitution distributions across sites. If all three p-values show high significance (e.g. each p-value ≤ 0.10), recombination is most likely present within the sequence set. The three tests are chosen because they are fast to run and the significance of the presence of a recombination event can be evaluated easily based on the p-values generated.

2.2 Phase II: Identification of recombination breakpoints

If the preliminary analysis suggests the presence of a recombination event, the corresponding breakpoints can then be identified with high accuracy using other approaches such as the Bayesian phylogenetic approach [15]. We used DualBrothers [17], which implements a Bayesian approach using reversible jump Markov chain Monte Carlo and dual multiple changepoint model in inferring changes in tree topologies and evolutionary rates across sites within a sequence set. While the prediction accuracy of recombination breakpoints comes at the expense of time and computational resources, the two-phase strategy avoids the use of time-consuming approaches in delineating recombination on sequence datasets for which there is no evidence of recombination in the first place.

3 EVALUATION OF PERFORMANCE

To evaluate the performance of the approaches used in our two-phase strategy, we simulated data with a single recombination event in the middle of a four-sequence set. The effects of subsequent substitutions after recombination and the sequence divergence prior to recombination were assessed. We used Seq-Gen [18] to simulate sequence evolution. Four-taxon sequence sets of length 1000 nt were generated using the HKY model of substitution [19] with nucleotide frequencies A = 0.20, C = 0.30, G = 0.30, T = 0.20, a transition/transversion ratio of 2, and a four-category discrete approximation to a gamma distribution of among-site rate variation with shape parameter alpha = 1.0. Different evolutionary histories prior to recombination and different number of subsequent substitutions after the recombination event were used in the simulations, and a total of 100 replicates was used for each combination. The recombination event was simulated by exchanging or replacing the fragments within the sequences resembling a reciprocal or non-reciprocal event respectively. After recombination, each lineage was evolved independently of the other with a different number of subsequent substitutions per site. See Figure 1 and Chan et al. [15] for more details on how the simulated sequence sets were generated.

The prediction accuracy for each of the three statistical measures implemented in PhiPack was evaluated by the p-value generated; a small p-value implies that the phylogeny discrepancy within the sequence set is significant, hence recombination is highly probable. Figure 2 shows the prediction accuracy (ρ) of each test across different combinations of prior evolutionary histories (showing sequence divergence) and subsequent substitutions (λ). The prediction accuracy (ρ) is calculated as the sum of true positives and true negatives over the sum of all cases in a set. A true positive was inferred when a p-value ≤ 0.10 was assigned on the recombinant sequence sets, whereas a false positive was inferred when a p-value ≤ 0.10 was assigned to the negative control sequence sets (that are void of recombination events). The results are shown for (a) reciprocal and (b) non-reciprocal sets, with prediction accuracies based on each test in isolation, any one test, any two of the three tests, and all three tests.

As shown in Figure 2, when more subsequent substitutions were simulated, the prediction accuracy decreased accordingly (e.g. non-reciprocal set, all tests $\rho < 0.60$ in L05/05 when $\lambda = 0.50$). In the reciprocal set (Figure 2a), NSS and PHI showed higher accuracy than MaxChi when used in isolation; PHI showed accuracy of 0.97 when the recombining sequences are divergent (L50) even when subsequent substitution is

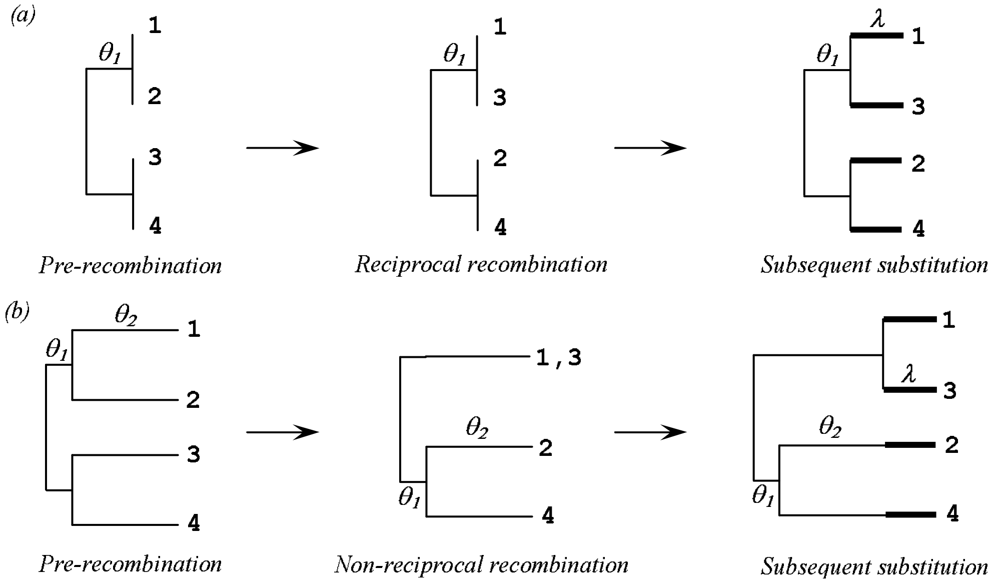
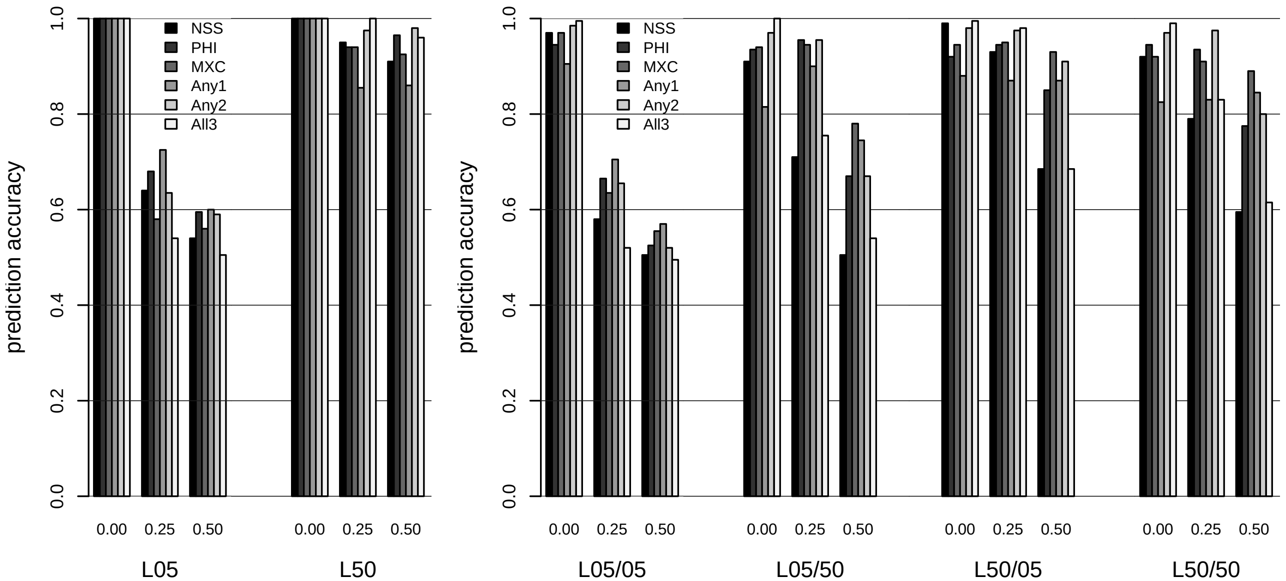


Figure 1: Tree topology of the recombined region in the simulations of (a) reciprocal and (b) non-reciprocal recombination. (i) Tree topology before recombination, with branch lengths represented by θ, e.g. the notation L05/50 represents θ_1 = 0.05 and θ_2 = 0.05 substitutions per site. Longer external branch lengths depict that the sequences are more divergent prior to recombination. (ii) A recombination event was simulated by exchanging (reciprocal) or replacing (non-reciprocal) the corresponding region between sequences 1 and 3, in the middle of the sequences. (iii) After recombination, subsequent substitutions (λ) were simulated on each sequence independently of each other. The signal of the recombination event is expected to diminish as λ increases, making detection of such an event more difficult. More-detailed description is provided in Chan et al. [15].



(a) Reciprocal set

(b) Non-reciprocal set

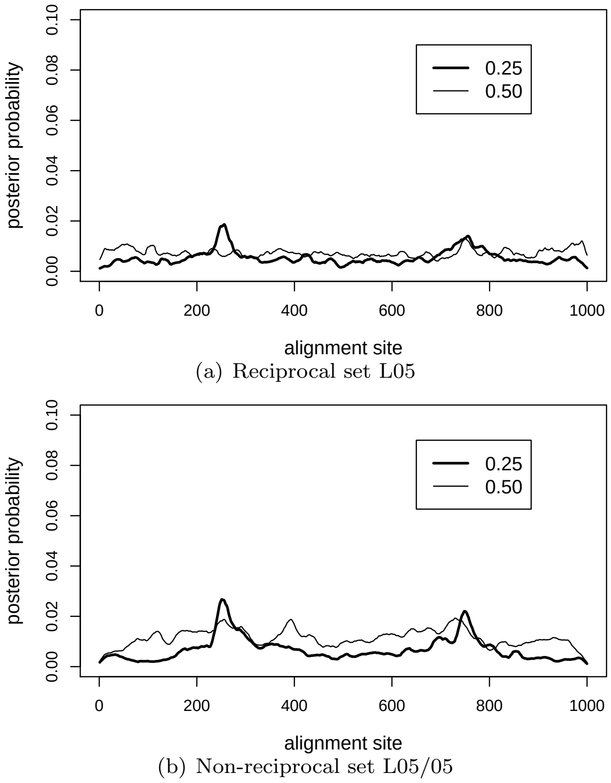
Figure 2: Prediction accuracies (ρ) of neighbour similarity score (NSS), pairwise homoplasy index (PHI) and maximal chi-squared (MXC) on simulated datasets of (a) reciprocal and (b) non-reciprocal recombination. The Y-axes represent the prediction accuracy; the X-axes represent different amounts of subsequent substitution (λ) across different tree topologies prior to the recombination event. For each set, each bar (from left to right) represents the prediction accuracy based on NSS; PHI; MXC; any one (Any1) of the three tests; any two (Any2) of the three tests; and all three (All3) tests.

high ($\lambda = 0.50$). In the non-reciprocal set (Figure 2b), MaxChi and PHI showed higher accuracy than NSS when used in isolation; MaxChi showed the highest accuracy among the three tests (e.g. prediction accuracy of MaxChi at 0.89 in cases of L50/50, $\lambda = 0.50$). In both reciprocal and non-reciprocal sets, the three tests showed higher prediction accuracy when the recombining sequences are more divergent (e.g. L50 in reciprocal set, L50/05 and L50/50 in non-reciprocal set) even when subsequent substitutions are high. When the recombining sequences are closely related to each other, the prediction accuracy of recombination event in the simulation sets decreased drastically when subsequent substitutions were increased (e.g. L05 in the reciprocal set and L05/05 in the non-reciprocal set where all methods showed $\rho \leq 0.70$ with $\lambda \geq 0.25$). This finding supports previous studies that have reported that recombination events involving closely related sequences are more difficult to detect than similar events involving sequences that are more divergent. A change in tree topology involving more-divergent sequences will yield a stronger recombination signal compared to a similar change involving closely related sequences; this effect has also been shown to create biases in phylogeny reconstruction [14].

Although the prediction accuracy of these statistical measures was sensitive to subsequent substitutions and prior evolutionary history (sequence divergence), PHI was found to be least sensitive to the number of subsequent substitutions and the prior evolutionary histories of the sequence sets (multiple regression analysis on non-reciprocal set, adjusted R^2 0.341, F statistic 156.1, 1195 degrees of freedom). All three statistical measures can be calculated quickly and are less dependent on certain parameter settings compared to other approaches evaluated in a previous study e.g. GENECONV (substitution distribution-based) or RecPars (phylogenetic-based) [15].

A recombination event can be inferred when all three tests show significant p-values (average accuracy 0.80 across both reciprocal and non-reciprocal sets), but this might be considered too strict. A better option is to rely on any two of these three tests; this yields an average prediction accuracy of 0.86, which is better than using a single method in isolation (PHI 0.85, MXC 0.85, NSS 0.80) or than using any one test (0.83). The prediction accuracies of all three methods are similar, but there are obvious biases in each of the three tests in detecting reciprocal and non-reciprocal recombination events, especially when the number of subsequent substitutions was high (Figure 2). One should not rely on a single test to detect a recombination event in set of sequences when the nature of such event (reciprocal or non-reciprocal) is unknown. Given the minimal time needed to compute p-values for each of the three tests, there is no great burden associated with computing all three tests.

For identifying recombination breakpoints, we used DualBrothers, a Bayesian phylogenetic approach using a multiple change-point model described by eight parameters related to location of breakpoints, tree topologies and evolutionary rates within a set of sequences. Using the same simulated dataset, the algorithm was found to be accurate in delineating recombination breakpoints, but at the expense of computational resources and time [15]. The two-phase approach with a first-pass screening reduces the number of datasets that require recombination breakpoint analysis with DualBrothers, as first-pass screening will avoid running computationally intensive DualBrothers on sequence sets for which there is no evidence of recombination.



(a) Reciprocal set L05

(b) Non-reciprocal set L05/05

Figure 3: Breakpoint detection using DualBrothers in false negative datasets resulted from the first-pass screening, in (a) reciprocal set L05, and (b) non-reciprocal set L05/05. The Y-axes represent the posterior probability of a site being proposed as a breakpoint. The X-axes represent sites in the alignment. The simulated breakpoints are at positions 250/251 and 750/751. The different lines on each graph represent the number of subsequent substitutions per site (λ): thick line, 0.25; thin line, 0.50.

To examine the effectiveness of first-pass screening in filtering out recombination-negative datasets, DualBrothers was run on the false negative datasets from the first-pass screening, i.e. cases that do not meet the requirement that any two of the three statistical tests yield p-values ≤ 0.10 in Figure 2. Figure 3 shows the relevant results on a selection of simulated datasets. DualBrothers was run with MCMC chain length = 1,000,000 generations and burn-in = 20,000 generations.

Within these false negative datasets, DualBrothers either failed to identify any recombination breakpoints (posterior probabilities of all sites < 0.010, re-

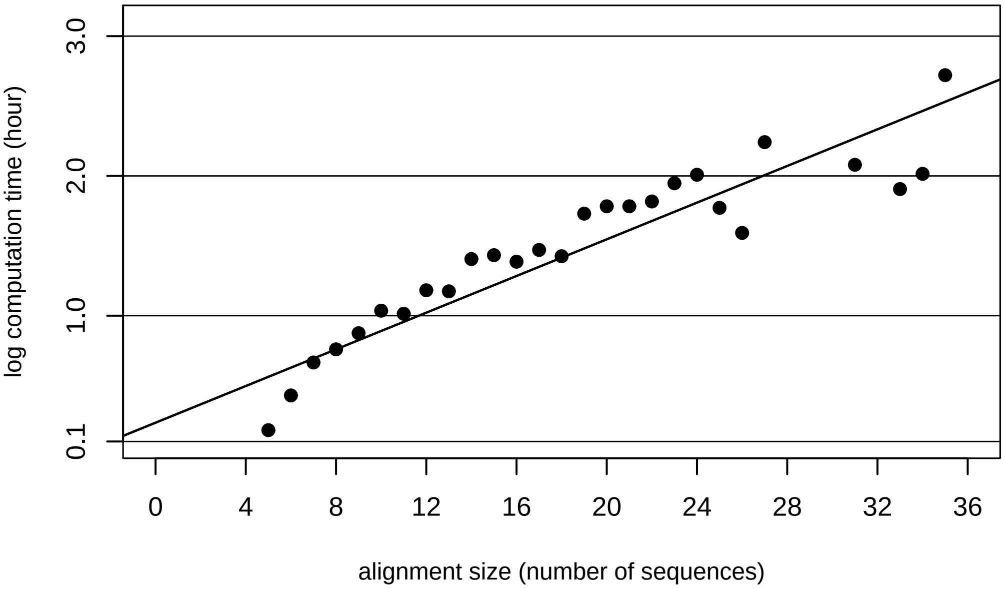


Figure 4: Computation time of DualBrothers across different sizes of alignments. The Y-axis represents computation time in the unit of hour on a 2.66GHz Pentium 4 processor (in a $\log_{10}$ scale), and the X-axis represents the number of sequences in an alignment. A linear relationship in this graph shows an exponential relationship between computation time and alignment size.

sults not shown), or, in the two instances shown in Figure 3, identified the breakpoints at a very low confidence level. The breakpoints in the simulated datasets are at positions 250/251 and 750/751. In the reciprocal case of L05 ($\lambda = 0.25$), breakpoints were identified at low posterior probability 0.019 (position 251/252) and 0.014 (position 755/756); whereas in the non-reciprocal case of L05/05 ($\lambda = 0.25$), both breakpoints 248/249 and 749/750 were identified at low posterior probabilities close to 0.025. The Bayes factor, as described in [17], was calculated for each of these false negative sets, to compare prior and posterior probabilities between the null and test hypotheses. The null hypothesis in this instance postulates that there is no recombination occurring in the sequences, whereas the test hypothesis postulates otherwise. A low Bayes factor (e.g. less than 5) would indicate a very strong support of the null hypothesis, i.e. DualBrothers failed to identify any recombination breakpoint. For the majority of the false negative datasets, the calculated Bayes factors were less than 50 (85.6%); a substantial number of the sequence sets (77.1%) with Bayes factor < 5. The weak, ambiguous breakpoint conclusions and low Bayes factors in the false negative cases show that the recombination events in these instances are indeed difficult to detect, and that first-pass screening using the three statistical tests is a good approach in filtering out datasets that are potentially recombination-negative. Therefore, having a first-pass screening in the recombination-detecting strategy will save time and computational resources, such that the accurate-but-slow method needs only be applied to those sequence sets that show evidence of recombination from the first-pass screening (e.g. at a certain cut-off threshold of p-value).

4 APPLICATION TO EMPIRICAL DATA

We applied the two-phase strategy in an attempt to infer recombination events among families of protein-coding sequences among prokaryotic genomes. The dataset consisted of 22437 putatively orthologous protein families obtained from 144 fully sequenced prokaryotic genomes [20]. Clustered via a hybrid approach of naïve and Markov clustering algorithms [21], protein alignments of these gene families were validated using a pattern-centric objective function [22], and converted into nucleotide sequence alignments for the analysis of recombination.

The first-pass screening for occurrence of recombination within the DNA alignments was carried out using the three statistical measures mentioned above, in which detection is treated as positive when two out of three measures (NSS, MaxChi and PHI) yield a p-value ≤ 0.10. Of 1462 DNA alignments of strictly orthologous gene families, 427 (29.2%) showed evidence of recombination. Based on this criterion, the largest gene family consisted of 48 sequences. The quick screening step greatly reduced the number of alignments needed for breakpoint detection in the next phase of the strategy.

Figure 4 shows the exponential relationship between the computation times of DualBrothers and the different number of sequences in an alignment. When the number of sequences approached 28, a runtime of over 100 hours was needed to run DualBrothers on a 2.66GHz Pentium 4 processor (MCMC chain length = 1,020,000 generations; burn-in = 20,000 generations; window length = 5; single start tree). The first-pass screening has reduced the number of datasets from 1462 to 427, and considering that it is necessary to run multiple replicates to increase confidence in our conclusions, the two-phase strategy greatly reduced

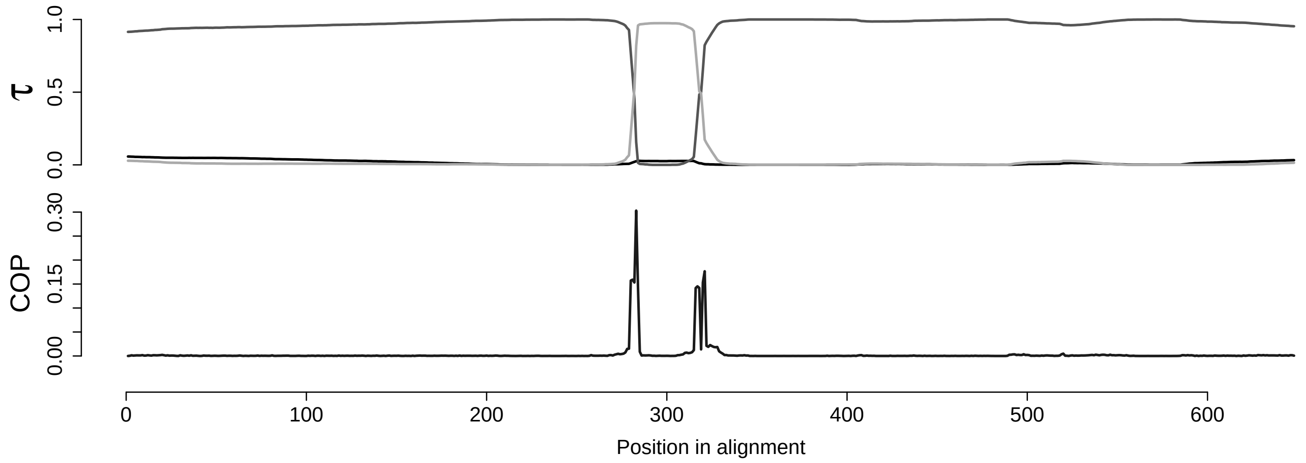


Figure 5: Identification of recombination breakpoints in gene family s1605. The upper graph shows posterior probability of different tree topology across the sites in the alignment; each line represents a different tree topology. The lower graph shows the COP (change-of-point) marginal posterior probability across sites in the alignments. A site with a sharp peak is a possible recombination breakpoint.

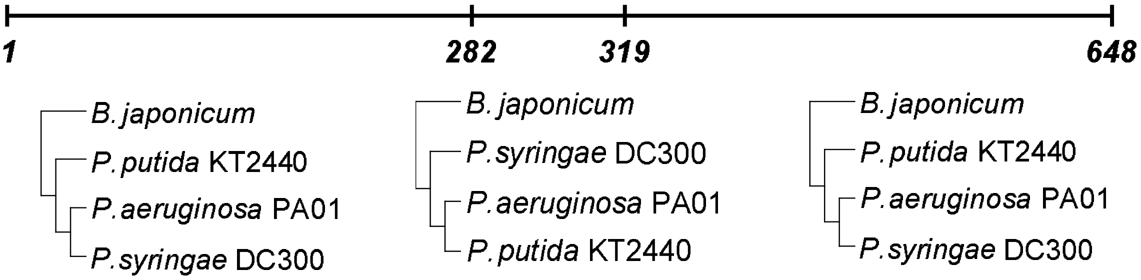


Figure 6: Tree topology within each partition separated by the recombination breakpoints. The line on top of the tree topologies depicts the positions of the DNA alignment (label shown is not to scale).

the required computational time and resources.

Figure 5 shows an example of a recombination event detected in a gene family, s1065, a family of hypothetical proteins consisting four sequences. As shown by the peaks in the change-of-point (COP) marginal posterior probability (mPP) plot and the corresponding change of tree topology, two possible breakpoint positions were detected, one between alignment positions 250-300, and another between positions 300-350.

Table 1: Inferred breakpoints in gene family s1605. Numbers shown are positions in the alignment.

COP(Breakpoint)	95% Bayesian Confidence Interval (BCI)
281/282	277-284
318/319	312-329

The exact breakpoints were determined by a statistical approach involving sub-sampling of COP mPP under a peak with 95% Bayesian confidence interval [17, 23], as shown in Table 1. Positions 281/282 and 318/319 in the DNA alignments were identified as the recombination breakpoints, separating the sequences into three partitions. Figure 6 shows the dominant tree topology within each partition of the alignment.

As shown in Figure 6, a recombination event was detected in this gene family, in which a recombined region was inferred in the middle of the alignment between positions 282-318. These topologies suggest several alternative explanations. A simple explanation is that the genetic fragments within that region could have been transferred from *Pseudomonas aeruginosa* PA01 to *P. putida* KT2440, or vice versa, making the two sequences more closely related to each other within the region 282-318.

5 CONCLUSIONS

Using different approaches in succession, we were able to detect recombination events more rapidly and with higher confidence than if a single method had been used in isolation. The first step of first-pass screening is quick and useful for filtering out sequence sets that show no evidence of recombination, making this approach suitable for detecting recombination in multi-genome scale data.

ACKNOWLEDGMENTS

This study is supported by an Australian Research Council (ARC) grant CE0348221. We thank Aaron

Darling and Vladimir Minin for invaluable input on the use of DualBrothers and for providing software codes. The first author is supported by a UQIPRS scholarship for his postgraduate study.

REFERENCES

[1] K. N. Kreuzer. "Interplay between DNA replication and recombination in prokaryotes". *Annual Review of Microbiology*, vol. 59, pp. 43–67, 2005.

[2] R. D. Camerini-Otero and P. Hsieh. "Homologous recombination proteins in prokaryotes and eukaryotes". *Annual Review of Genetics*, vol. 29, pp. 509–552, 1995.

[3] N. Kleckner. "Meiosis: how could it work?" *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 16, pp. 8167–8174, 1996.

[4] G. F. Weiller. "Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences". *Molecular Biology and Evolution*, vol. 15, no. 3, pp. 326–335, 1998.

[5] G. J. Etherington, J. Dicks and I. N. Roberts. "Recombination Analysis Tool (RAT): a program for the high-throughput detection of recombination". *Bioinformatics*, vol. 21, no. 3, pp. 278–281, 2005.

[6] S. Sawyer. "Statistical tests for detecting gene conversion". *Molecular Biology and Evolution*, vol. 6, no. 5, pp. 526–538, 1989.

[7] I. B. Jakobsen and S. Easteal. "A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences". *Computer Applications in the Biosciences*, vol. 12, no. 4, pp. 291–295, 1996.

[8] T. C. Bruen, H. Philippe and D. Bryant. "A simple and robust statistical test for detecting the presence of recombination". *Genetics*, vol. 172, no. 4, pp. 2665–2681, 2006.

[9] J. Hein. "Reconstructing evolution of sequences subject to recombination using parsimony". *Mathematical Biosciences*, vol. 98, no. 2, pp. 185–200, 1990.

[10] S. L. Kosakovsky Pond, D. Posada, M. B. Gravenor, C. H. Woelk and S. D. Frost. "Automated phylogenetic detection of recombination using a genetic algorithm". *Molecular Biology and Evolution*, vol. 23, no. 10, pp. 1891–1901, 2006.

[11] J. Graham, B. McNeney and F. Seillier-Moiseiwitsch. "Stepwise detection of recombination breakpoints in sequence alignments". *Bioinformatics*, vol. 21, no. 5, pp. 589–595, 2005.

[12] D. Posada and K. A. Crandall. "Evaluation of methods for detecting recombination from DNA sequences: computer simulations". *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 24, pp. 13757–13762, 2001.

[13] C. Wiuf, T. Christensen and J. Hein. "A simulation study of the reliability of recombination detection methods". *Molecular Biology and Evolution*, vol. 18, no. 10, pp. 1929–1939, 2001.

[14] D. Posada. "Evaluation of methods for detecting recombination from DNA sequences: empirical data". *Molecular Biology and Evolution*, vol. 19, no. 5, pp. 708–717, 2002.

[15] C. X. Chan, R. G. Beiko and M. A. Ragan. "Detecting recombination in evolving nucleotide sequences". *BMC Bioinformatics*, vol. 7, no. Art. 412, 2006.

[16] J. Maynard Smith. "Analyzing the mosaic structure of genes". *Journal of Molecular Evolution*, vol. 34, no. 2, pp. 126–129, 1992.

[17] V. N. Minin, K. S. Dorman, F. Fang and M. A. Suchard. "Dual multiple change-point model leads to more accurate recombination detection". *Bioinformatics*, vol. 21, no. 13, pp. 3034–3042, 2005.

[18] A. Rambaut and N. C. Grassly. "Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees". *Computer Applications in the Biosciences*, vol. 13, no. 3, pp. 235–238, 1997.

[19] M. Hasegawa, H. Kishino and T. A. Yano. "Dating of the human ape splitting by a molecular clock of mitochondrial DNA". *Journal of Molecular Evolution*, vol. 22, no. 2, pp. 160–174, 1985.

[20] R. G. Beiko, T. J. Harlow and M. A. Ragan. "Highways of gene sharing in prokaryotes". *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 40, pp. 14332–14337, 2005.

[21] T. J. Harlow, J. P. Gogarten and M. A. Ragan. "A hybrid clustering approach to recognition of protein families in 114 microbial genomes". *BMC Bioinformatics*, vol. 5, no. Art. 45, 2004.

[22] R. G. Beiko, C. X. Chan and M. A. Ragan. "A word-oriented approach to alignment validation". *Bioinformatics*, vol. 21, no. 10, pp. 2230–2239, 2005.

[23] M. A. Suchard, R. E. Weiss, K. S. Dorman and J. S. Sinsheimer. "Inferring spatial phylogenetic variation along nucleotide sequences: a multiple change-point model". *Journal of the American Statistical Association*, vol. 98, no. 462, pp. 427–437, 2003.