Default = Best?
High Performance Computation Assisted Parameter Refinement in Genomic Scan for Long Terminal Repeat (LTR) Retrotransposons

Shujun Ou, Ning Jiang (jiangn@msu.edu)
Department of Horticulture, Michigan State University

Abstract
Long Terminal Repeat (LTR) retrotransposon is a subclass of mobile sequences that dominate the genome ingredients. LTRharvest is a de novo LTR sequence detection program that helps to develop comprehensive genome annotations. However, even utilizing the best methods, with inappropriate parameters notable false positives will be generated. In this research, we selected four parameters of LTRharvest and tested with multiple levels on three rice genomes. By evaluating sensitivities and specificities of these parameters, we found that the recovered LTR candidate pool could be more than 10 times difference. The optimum parameters we found were SEED=30 (default), SIMILAR=90 (dft=85), VIC=5 (dft=60), MAXTSD=20 (dft), indicating that the optimum parameters are not always the default ones. We also developed two versions of scripts to run the RepeatMasker either parallelly or with limit computation resources (PC). The outcome knowledge and tools would enhance cooperation and application in the CI community.

Introduction
The structure of Long Terminal Repeat (LTR) retrotransposons have been well characterized (Figure 1) and very suited for computational identification. This research intended to refine the parameters of LTRharvest [1], a de novo LTR sequence detection program, so that the false detection rate could be controlled while maximizing the sensitivity and specificity.

![Figure 1. Structure of the Long Terminal Repeat (LTR) retrotransposon. ILTR, left LTR. rLTR, right LTR. TSD, target site duplication. Motif, conserved dinucleotide at the 5’ and 3’ end of the LTRs. SIMILAR, similarity between the left and right LTR. SEED, minimum seed length for alignment extension. MAXTSD, maximum length for each TSD. VIC, the range (bp) for TSDs and/or motifs searching at the 5’ and 3’ end of LTR candidates.](image)

By pre-study screening, we selected four parameters with several levels of LTRharvest that potentially affect the detection result significantly (Table 1).

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIMILAR</td>
<td>95, 90, 85 (dft), 80</td>
</tr>
<tr>
<td>SEED</td>
<td>50, 40, 30 (dft), 20, 10</td>
</tr>
<tr>
<td>MAXTSD</td>
<td>20 (dft), 15, 10, 9, 8, 7, 6</td>
</tr>
<tr>
<td>VIC</td>
<td>60 (dft), 20, 15, 10, 9, 8, 7, 6, 5, 4, 3</td>
</tr>
</tbody>
</table>

Table 1. Selected parameters and levels of LTRharvest. Legend for the terms refer to Figure 1.

Results
Sensitivity of LTRharvest decreased when increasing SEED and SIMILAR parameters, and increased when increasing the VIC. Specificity increased when increasing SEED and SIMILAR. But when increasing the VIC, the specificity increased at first and then dropped afterward. Although the ranges of sensitivity and specificity are not large, 4.52% and 8.00% respectively, the outcome could be more than 10 times difference, which recovered LTR candidate pool is 33.4m in max and 3.3m in min in the Kasalath genome.

The optimum parameters of LTRharvest for rice were suggested as SEED=30 (default), SIMILAR=90 (dft=85), VIC=5 (dft=20), MAXTSD=20 (dft). Base on the structural stability and conservation of LTR element, this parameter set is expected to be applied to de novo searching of LTR candidates in new genomes.

What we learnt is that for LTRharvest, probably can be extended to any other software, the optimum parameters are not always the default ones. End users should carefully modify their parameters base on biological sense and experience.

Discussion

Materials and methods
We used three assembled rice genomes as biological replicated samples: Oryza sativa ssp. japonica var. Nipponbare [2], O. sativa ssp. indica var. R9311 [3], O. sativa ssp. aas var. Kasalath [4].

To estimate the sensitivity and specificity (Figure 2) of these parameter combinations, we used RepeatMasker[3], an alignment based TE sequence detection program, to verify the outcome of LTRharvest.

![Figure 2. Illustration of sensitivity and specificity. a, the genomic regions that masked by the established LTR library were taken out and masked by the LTR candidates, the final masked rate is sensitivity. b, Mask rate of the LTR candidate by the established LTR library is specificity.](image)

All analyses were performed in HPCG with automated scripts. Performing genome level parameter testing would consume massive computation powers. The following problems may be not so unfamiliar to HPCG users:

- Infinite queue time? Almost finished jobs got killed?

No problem! We developed 2 versions of scripts that can perform RepeatMasker analysis with either parallel computing algorithm or split task scheme. Also, both of the scripts have been equipped with a check point scheme, allowing users interrupt and continue their analysis at any time. With these tools users can even perform genome level analysis of LTRharvest and RepeatMasker with limit computation resources (PC).

References