

RelocaTE2: a high resolution transposable element insertion site mapping tool for population resequencing

Jinfeng Chen^{1,2,3}, Travis R Wrightsman³, Susan R Wessler^{2,3}, Jason E. Stajich^{Corresp. 1,2}

¹ Department of Plant Pathology & Microbiology, University of California, Riverside, Riverside, CA, United State

² Institute for Integrative Genome Biology, University of California, Riverside, Riverside, CA, United States

³ Department of Botany and Plant Sciences, University of California, Riverside, Riverside, CA, United States

Corresponding Author: Jason E. Stajich

Email address: jason.stajich@ucr.edu

Background Transposable element (TE) polymorphisms are important components of population genetic variation. The functional impacts of TEs in gene regulation and generating genetic diversity have been observed in multiple species, but the frequency and magnitude of TE variation is under appreciated. Inexpensive and deep sequencing technology has made it affordable to apply population genetic methods to whole genomes with methods that identify single nucleotide and insertion/deletion polymorphisms. However, identifying TE polymorphisms, particularly transposition events or non-reference insertion sites can be challenging due to the repetitive nature of these sequences, which hamper both the sensitivity and specificity of analysis tools.

Methods We have developed the tool RelocaTE2 (<http://github.com/stajichlab/RelocaTE2>) for identification of TE insertion sites at high sensitivity and specificity. RelocaTE2 searches for known TE sequences in whole genome sequencing reads from second generation sequencing platforms such as Illumina. These sequence reads are used as seeds to pinpoint chromosome locations where TEs have transposed. RelocaTE2 detects target site duplication (TSD) of TE insertions allowing it to report TE polymorphism loci with single base pair precision.

Results and Discussion The performance of RelocaTE2 is evaluated using both simulated and real sequence data. RelocaTE2 demonstrate high level of sensitivity and specificity, particularly when the sequence coverage is not shallow. In comparison to other tools tested, RelocaTE2 achieves the best balance between sensitivity and specificity. In particular, RelocaTE2 performs best in prediction of TSDs for TE insertions. Even in highly repetitive regions, such as those tested on rice chromosome 4, RelocaTE2 is able to report up to 95% of simulated TE insertions with less than 0.1% false positive rate using 10-fold genome coverage resequencing data. RelocaTE2 provides a robust solution to identify TE insertion sites and can be incorporated into analysis workflows in support of describing the complete genotype from light coverage genome sequencing.

1 **RelocaTE2: a high resolution transposable element insertion site mapping tool for**
2 **population resequencing**

3 Jinfeng Chen^{1,2,3}, Travis R. Wrightsman², Susan R. Wessler^{2,3} and Jason E. Stajich^{1,3,*}

4 ¹ Department of Plant Pathology and Microbiology, University of California-Riverside,
5 Riverside CA 92521, USA.

6 ² Department of Botany and Plant Sciences, University of California-Riverside, Riverside CA
7 92521, USA.

8 ³ Institute for Integrative Genome Biology, University of California-Riverside, Riverside, CA
9 92521, USA.

10 *To whom correspondence should be addressed. E-mail: jason.stajich@ucr.edu

11 Abstract**12 Background**

13 Transposable element (TE) polymorphisms are important components of population genetic
14 variation. The functional impacts of TEs in gene regulation and generating genetic diversity have
15 been observed in multiple species, but the frequency and magnitude of TE variation is under
16 appreciated. Inexpensive and deep sequencing technology has made it affordable to apply
17 population genetic methods to whole genomes with methods that identify single nucleotide and
18 insertion/deletion polymorphisms. However, identifying TE polymorphisms, particularly
19 transposition events or non-reference insertion sites can be challenging due to the repetitive
20 nature of these sequences, which hamper both the sensitivity and specificity of analysis tools.

21 Methods

22 We have developed the tool RelocaTE2 (<http://github.com/stajichlab/RelocaTE2>) for
23 identification of TE insertion sites at high sensitivity and specificity. RelocaTE2 searches for
24 known TE sequences in whole genome sequencing reads from second generation sequencing
25 platforms such as Illumina. These sequence reads are used as seeds to pinpoint chromosome
26 locations where TEs have transposed. RelocaTE2 detects target site duplication (TSD) of TE
27 insertions allowing it to report TE polymorphism loci with single base pair precision.

28 Results and Discussion

29 The performance of RelocaTE2 is evaluated using both simulated and real sequence data.
30 RelocaTE2 demonstrate high level of sensitivity and specificity, particularly when the sequence
31 coverage is not shallow. In comparison to other tools tested, RelocaTE2 achieves the best
32 balance between sensitivity and specificity. In particular, RelocaTE2 performs best in prediction
33 of TSDs for TE insertions. Even in highly repetitive regions, such as those tested on rice
34 chromosome 4, RelocaTE2 is able to report up to 95% of simulated TE insertions with less than
35 0.1% false positive rate using 10-fold genome coverage resequencing data. RelocaTE2 provides
36 a robust solution to identify TE insertion sites and can be incorporated into analysis workflows in
37 support of describing the complete genotype from light coverage genome sequencing.

38 **Introduction**

39 Transposable elements (TE), mobile DNA of the genome, are drivers of genomic innovation
40 (Bennetzen & Wang 2014; Cordaux & Batzer 2009). They can act as mutagens to disrupt gene
41 functions or induce novel gene functions by providing enhancers or promoters that alter host
42 gene expression (Feschotte 2008; Lisch 2013). In plants, TEs have been shown to contribute to
43 several key trait innovations in crop domestication (Lisch 2013). Systematic analysis of TE
44 insertions and gene expression also suggests widespread roles of TEs in altering gene regulation
45 (Kunarso et al. 2010; Lynch et al. 2011; Sundaram et al. 2014). It was found that 600-2000
46 genetic variants between individuals in the human population and 200-300 variants between
47 Arabidopsis accessions could be attributed to TE polymorphisms (Quadrana et al. 2016; Stewart
48 et al. 2011). Although the magnitude of these polymorphisms is small compared to SNPs or
49 other insertion/deletions, some TE polymorphisms are proximal to protein coding genes and can
50 have large impacts on gene function or gene regulation (Cowley & Oakey 2013; Quadrana et al.
51 2016; Stewart et al. 2011).

52 Two categories of bioinformatics tools have been developed to identify TE polymorphisms from
53 population resequencing data. One type employs a strategy similar to that used to discover
54 structural variations. These tools identify discordant pairs of sequence reads based on the
55 chromosomal position of read alignments to indicate genomic inversions, insertions, deletions or
56 other complex rearrangements (Campbell et al. 2008; Korbel et al. 2007). Software for TE
57 mapping scrutinize genomic loci with discordant read pairs to see if known TE sequences are can
58 be implicated near the rearrangement site. These tools, such as Retroseq (Keane et al. 2013) and
59 TEMP (Zhuang et al. 2014), are generally highly sensitive and can locate TE insertions and
60 absences to a 10-50 bp resolution. A second category of tools operates by first identifying by
61 similarity any sequence reads containing partial or complete known TE sequences. Any sequence
62 containing a TE is a "junction-read" which contains partial TE and partial unique host genomic
63 sequence. These tools eliminate the TE sequence from these junction-reads and search the
64 remaining 5' or 3' flanking sequence against the host organism genome sequence to identify the
65 element's location. These junctions-based tools, including RelocaTE (Robb et al. 2013) and ITIS
66 (Jiang et al. 2015), are able to detect the exact location and TSD characteristic of TE insertion
67 sites. This second category of tools is ideal for identifying new insertions from population
68 resequencing data because it can accurately detect an insertion location along with the TSD.

69 However, most of these tools are designed to search a single transposable element representative
70 sequence at a time, which sacrifices speed for increased sensitivity and specificity. The extended
71 runtime limits the feasibility of applying these tools when searching thousands of TEs in
72 hundreds or thousands of individuals.

73 RelocaTE2 is an improved version of RelocaTE where we have implemented a junction-based
74 approach that can search multiple template TEs in the same pass through short read sequencing
75 data, streamlining the computational approach. Using simulated datasets, we show that
76 RelocaTE2 is highly sensitive even in low coverage resequencing data or on chromosomes with
77 a high repetitive sequence content with a specificity of greater than 99%. In comparing
78 performance of related tools, RelocaTE2 was the most sensitive and specific tool in our tests
79 profiling human and rice population genomics data. The tool is presented as a useful resource for
80 analyzing population dynamics of TEs in genomic resequencing data.

81 **Materials & Methods**

82 **RelocaTE2 Workflow**

83 RelocaTE2 is based on the previous algorithm implemented in RelocaTE (Robb et al. 2013),
84 which uses junction reads to find insertion sites of TEs. In RelocaTE2, we re-implement the
85 search strategy to enable identification of multiple TEs in a single search, greatly increasing the
86 speed and enabling searches for hundreds or thousands of candidate TE families in a genome
87 (Fig. 1). We also implement new features in the algorithm to automatically identify TSDs and
88 remove false junction reads (Fig. 1).

89 Briefly, the workflow initiates by matching a library of known repeat elements against short
90 sequence reads generated by next generation sequencing, typically Illumina paired-end reads,
91 using BLAT with the sensitive setting "`-minScore=10 -tileSize=7`" (Kent 2002; Robb
92 et al. 2013). Every read with similarity to repeat elements is denoted as an informative read and
93 will contain a partial or complete copy of a TE. Informative reads that contain partial matches at
94 the boundaries of the repeat elements are trimmed to remove the TE region so that only the
95 regions flanking the element remain in either one or both of the paired-end reads (denoted as
96 junction reads). Untrimmed versions of each junction read and its pair (denoted as full reads) are
97 used as controls to filter false positive junction reads.

98 Sequence reads comprised entirely of repeat elements are ignored, but their read pair is kept
99 (denoted as supporting reads). These junction, full, and supporting reads, are all aligned to the
100 reference genome using BWA (v0.6.2) with the default setting "-l 32 -k 2" (Li & Durbin
101 2009). Mapped reads are sorted by chromosome order and windows of 2,000 bp are evaluated to
102 define insertion clusters. In each insertion cluster, additional subclusters are further refined based
103 on the mapping position of junction reads to address the possible scenario of multiple insertion
104 sites within a window. TSD position and sequence are identified if the subcluster is supported by
105 junction reads from both upstream and downstream of the TE insertion site.

106 Next, a series of cleaning steps are used to filter low quality candidate insertion sites: i.) remove
107 insertion sites that are only supported by low quality junction reads (map quality < 29); ii.)
108 remove insertion site only supported by less than 3 junction reads total on the left or right flank
109 when there are additional insertion sites which pass these filters in the same window. iii.) remove
110 insertion sites only supported by junction reads and located within 10 bp range of an annotated
111 TE in the reference genome. RelocaTE2 output reports the number of junction reads and
112 supporting reads from both upstream and downstream of candidate TE insertion sites. Only
113 confident insertions, defined as having at least one supporting junction read flanking the
114 upstream or downstream of insertion sites and at least one junction read or one supporting read
115 supporting the other end of TE insertion, are provided in the default output:

116 "ALL.all_nonref_insert.gff". Additional information about all candidate sites are provided in
117 alternative output file: "ALL.all_nonref_insert.raw.gff".

118 **Simulated data for evaluation of TE insertion tools**

119 Simulated datasets were created by randomly inserting TEs into sequence records of
120 chromosomes 3 (OsChr3) and 4 (OsChr4) of rice (*Oryza sativa japonica*). OsChr3 is primarily
121 made up of euchromatic regions, whereas OsChr4 has the sequence complexity consistent with
122 heterochromatic regions and is a typical feature of many plant genomes (Zhao et al. 2002).
123 Fourteen TE families found in rice genomes comprised of 7 DNA Transposons: *mPing*, *nDart*,
124 *Gaijin*, *spmlike*, *Truncator*, *mGing*, *nDarz* and 7 RNA Retrotransposons: *Bajie*, *Dasheng*,
125 *Retro1*, *RIRE2*, *RIRE3*, *Copia2*, *karma*, were used. The insertion simulations were performed by
126 choosing 200 random insertion sites on each chromosome in three independent replicates. Each
127 simulated insertion site was generated by selecting one random chromosome position, one
128 random TE, and inserting the element along with the expected TSDs. After generating 200

129 random insertions, a new genome sequence was generated along with a GFF3 file containing the
130 recorded insertion locations to support the performance evaluation of the dataset. Paired-end
131 reads of all simulated chromosomes were simulated by pIRS (`pirs simulate -l 100 -x coverage -`
132 `m 500 -v 100`) (Hu et al. 2012). For each dataset, simulate sequence reads at sequence depths of
133 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 40-fold coverage were generated.

134 **Real sequence data for evaluation of TE insertion tools**

135 Three sets of data, an individual human genome, HuRef, an individual rice genome, IR64, and
136 population resequencing data of 50 rice and wild rice genomes (Levy et al. 2007; Schatz et al.
137 2014; Xu et al. 2012), were used to evaluate the performance of RelocaTE2 and TEMP. High
138 quality genome assemblies of HuRef and IR64 were used to evaluate the accuracy of TE
139 genotyping tools by comparing the assembled sequences to the reference genome. The HuRef
140 (also known as Venter genome) has been extensively studied for TE insertions (Xing et al.
141 2009). Previous work identified 574 *Alu* elements that have been experimentally verified and can
142 be treated as a gold standard data set for evaluation (Hormozdiari et al. 2010; Xing et al. 2009).
143 Paired-end sequence reads of 10-fold depth were simulated from HuRef as test dataset by pIRS
144 (`pirs simulate -l 100 -x coverage -m 500 -v 100`) (Hu et al. 2012).
145 RelocaTE2 and TEMP were tested and their results compared to the Human Genome Reference
146 Consortium genome (GRCh36 or hg18). A second dataset, the finished reference genome
147 assembly of rice strain IR64, was explored utilizing available Illumina sequencing reads (Schatz
148 et al. 2014). RelocaTE2 and TEMP were tested on 20-fold genome coverage of 100 bp paired-
149 end Illumina short reads (SRA accession: SRR546439) aligned to the rice reference genome
150 (MSU7). A third dataset of resequencing data from 50 strains of rice and wild rice population
151 with an average sequencing depth of 17-fold (Xu et al. 2012). RelocaTE2 and TEMP were tested
152 on the sequencing libraries from each of these 50 strains to assess their consistency across
153 datasets with varying sequence depth and genetic diversity. RelocaTE and ITIS were not
154 included in the biological data testing because of the prohibitively long run times on these large
155 datasets and their poor performance on simulated datasets.

156 **Detection of TE insertions using RelocaTE2, RelocaTE, TEMP and ITIS**

157 RelocaTE2, RelocaTE, TEMP and ITIS were run with default parameter settings on simulated
158 data. The results were filtered to evaluate the best performing parameters for each tool.

159 RelocaTE2 was tested with parameters "`--len_cut_match 10 --len_cut_trim 10`

160 `--mismatch 2 --aligner blat`", which uses BLAT as the search engine (`--aligner`
161 `blat`), allows for 2 mismatches (`--mismatch 2`) in matched sequence between reads and
162 repeat elements (`--len_cut_match 10`), and only keeps sequence fragments that have at
163 least 10 bp after trimming repeat elements from reads (`--len_cut_trim 10`). RelocaTE
164 was tested using parameters "`--len_cutoff 10 --mismatch 0`", which uses BLAT as
165 search engine by default and allowed 0 bp mismatch (`--mismatch 0`) for matched sequence
166 between reads and repeat elements (`--len_cutoff 10`). It should be noted that the mismatch
167 setting in RelocaTE is the ratio of base pairs in the alignment between reads and repeat elements
168 that can be mismatched, not an integer number of allowed mismatches, as used in RelocaTE2.
169 Singleton calls from RelocaTE's results, which are sites supported by only one read, were
170 removed. TEMP was tested with parameters "`-m 3`", which allow for three mismatches
171 between reads and repeat elements. Singleton calls from TEMP's results were removed when
172 testing on simulated data to achieve a balance between sensitivity and specificity. ITIS was
173 tested with default parameters, which filtered TE calls with at least one read supporting from
174 both ends of TE insertions. For analysis of the HuRef genome, the IR64 genome and the 50 rice
175 and wild rice strains, RelocaTE2 and TEMP were run with default parameter settings as
176 described above. The TEMP results were filtered to keep only TE calls with supporting and/or
177 junction reads from both ends of TE insertions to achieve a comparable balance between
178 sensitivity and specificity.

179 **Results and Discussions**

180 **Performance of RelocaTE2, RelocaTE, TEMP and ITIS on simulated data**

181 RelocaTE2 was first compared to RelocaTE, TEMP and ITIS using the simulated datasets. Each
182 dataset of simulated rice chromosomes, OsChr3 and OsChr4, was virtually sheared to simulated
183 paired-end short reads at a coverage ranging from 1 to 40-fold. At high sequencing coverage
184 (≥ 10 -fold), RelocaTE2, TEMP and ITIS were able to identify $>99\%$ of simulated insertions on
185 OsChr3, whereas the performance of RelocaTE was much lower (85%) (Fig. 2A). At lower
186 sequencing coverage, e.g. 3-fold, only RelocaTE2 and TEMP were able to achieve $\geq 95\%$
187 sensitivity on OsChr3 (Fig. 2A). Furthermore, TEMP was able to identify 83% and 93% of
188 simulated insertions on OsChr3 at very low sequence coverage of 1-fold and 2-fold, respectively
189 (Fig.

190 2A). RelocaTE2 had a sensitivity of 53% and 83% on OsChr3 for the 1-fold and 2-fold coverage
191 due to the removal of TE insertions supported by only one read (singleton) or supported by reads
192 from only one end of TE insertions (insufficient insertions), which can result in many false
193 positives (Fig. 2A).

194 RelocaTE2, RelocaTE and TEMP showed >99% specificity on OsChr3 at multiple levels of
195 sequence coverage (Fig. 2B). In contrast, the specificity of ITIS was much lower (<90%), even
196 when run on the high sequence coverage dataset on OsChr3 (Fig. 2B). In comparing recall rates
197 of TSDs, RelocaTE2 and ITIS had similar performance and achieved the highest recall rate of
198 98% and 91% respectively, on OsChr3 at ~10-fold coverage (Fig. 2C). The recall rate of TSDs
199 for both TEMP and RelocaTE depended on sequence depth and achieved only 37% and 60%,
200 respectively, at 10-fold coverage (Fig. 2C). All the tools performed worse on OsChr4 as
201 compared to OsChr3 (Fig. 2D-F). RelocaTE2 demonstrated a lower average sensitivity (92%) on
202 OsChr4 when compared OsChr3 (96%) (Fig. 2A, D). Similarly, TEMP had a slightly lower
203 sensitivity (95%) on OsChr4 than on OsChr3 (97%) (Fig. 2A, D). However, RelocaTE2 and
204 RelocaTE demonstrated high level of the specificity (>99%) while TEMP performed at a slightly
205 lower specificity (98%) on OsChr4 compared to >99% on OsChr3 (Fig. 2B, E). In comparing
206 TSD accuracy on OsChr4, on average 81% of RelocaTE2 calls correctly identified the TSD,
207 whereas only 31% of TEMP calls were correct (Fig. 2C, F).

208 **Evaluation of RelocaTE2 and TEMP on biological datasets**

209 We evaluated TE identifying tools in the HuRef genome and benchmark the sensitivity and
210 specificity of these tools using 574 experimental verified *Alu* insertions in HuRef genome and
211 genomic comparison between HuRef genome and GRCh36. RelocaTE2 and TEMP reported
212 similar results and identified 83% (479/574) and 76% (438/574) of standard insertion sites (Fig.
213 3A). Comparing the HuRef genome with GRCh36 suggested that 89% and 95% of insertions
214 identified by RelocaTE2 and TEMP, respectively, were real insertions (Fig. 3A). In addition,
215 RelocaTE2 predicted TE insertion sites with higher precision (9 ± 6 bp) compared to TEMP (366
216 ± 170 bp).

217 RelocaTE2 and TEMP were used to analyze data from the rice strain IR64 and the results were
218 evaluated by comparing the genome assembly of IR64 with MSU7. RelocaTE2 identified 648
219 insertion sites while the genome comparison revealed that 93% of insertions were true positives
220 (Fig. 3A). TEMP identified 362 insertions, of which 50% (183/362) overlapped with RelocaTE2

221 (Fig. 3A). The specificity of TEMP was estimated to be 86%, slightly lower than RelocaTE2
222 (93%) (Fig. 3A). However, TEMP was found to be less sensitive than RelocaTE2 in the rice
223 genome, only calling 362 sites as compared to 648 by RelocaTE2 (38% vs. 90%, Fig. 3A).
224 Moreover, RelocaTE2 predicted TE insertion junctions of 3 ± 1 bp, which was much smaller
225 than TEMP (393 ± 199 bp).

226 RelocaTE2 and TEMP were used to identify TE polymorphisms in 50 resequenced rice and wild
227 rice strains, which contain substantial sequence diversity and population structure (Xu et al.
228 2012). The results from these two tools were well correlated ($R^2 = 0.96$, P value = $2.2e-16$) and
229 predicted more TE insertions in the diverged population of wild rice, *O. nivara* and *O. rufipogon*,
230 and even in the *indica* population than *japonica* rice which close to the reference genome (Fig.
231 3B). On average 72% of the sites predicted in these 50 rice and wild rice strain by RelocaTE2
232 and TEMP overlapped. Many insertion sites from TEMP were predicted with only supporting
233 read flanking one end of an insertion, which produced large variations in predicted junctions of
234 TE insertion sites (118 ± 151 bp). In contrast, RelocaTE2 reported most of TE insertions
235 supported by junction reads or supporting reads on both ends, which resulted in accurate
236 insertion junction predictions (3 ± 2 bp).

237 **Runtime performance**

238 We implemented the searching process for TE insertion to run on multiple processors in Python.
239 The process is relatively memory efficient. When searching TEs in the rice genome for example,
240 one process generally uses less than 1 Gb memory. The running time of RelocaTE2 depends on
241 number of processors used. Searching 3000 templates of transposable elements with 20-fold
242 genome coverage sequencing data of the rice genome takes 3-4 hours for RelocaTE2 using 32
243 CPUs including the alignment steps. TEMP identifies transposable element insertions from a
244 BAM file. It takes ~1 hours for TEMP for the same project using single process. RelocaTE
245 (version 1) and ITIS take at least days for the same rice datasets and can be prohibitively difficult
246 to run on large datasets with multiple templates due to the serial searching approach of their
247 implementation.

248

249 **Conclusions**

250 We present RelocaTE2 as a new tool for mapping TE insertions to base-pair resolution from
251 resequencing data. RelocaTE2 identifies multiple TE families in a single search with high

252 sensitivity and specificity. The evaluation of these tools on simulated and biological datasets
253 support the use of RelocaTE2 for analysis of genomes of plants and animals and indicate it can
254 generate very high quality genotyping of TE insertions from resequencing datasets of modest
255 sequencing depths. TEMP is fast and achieves well balance between sensitivity and specificity in
256 identifying TE polymorphisms, particularly when the sequence depth is shallow. Although the
257 precision of TSD prediction of TEMP is not very sensitive compared to RelocaTE2 and other
258 tools. We recommend using TEMP in scenarios when accuracy of TSD is not very critical. The
259 high resolution mapping of TE insertions sites will enable detailed analysis of the interaction of
260 TEs and genes and as structural variations that vary in populations.

261

262 **Data Availability**

263 The source code in Python, manual, and sample data of RelocaTE2 are available for download at
264 <https://github.com/stajichlab/RelocaTE2>.

265

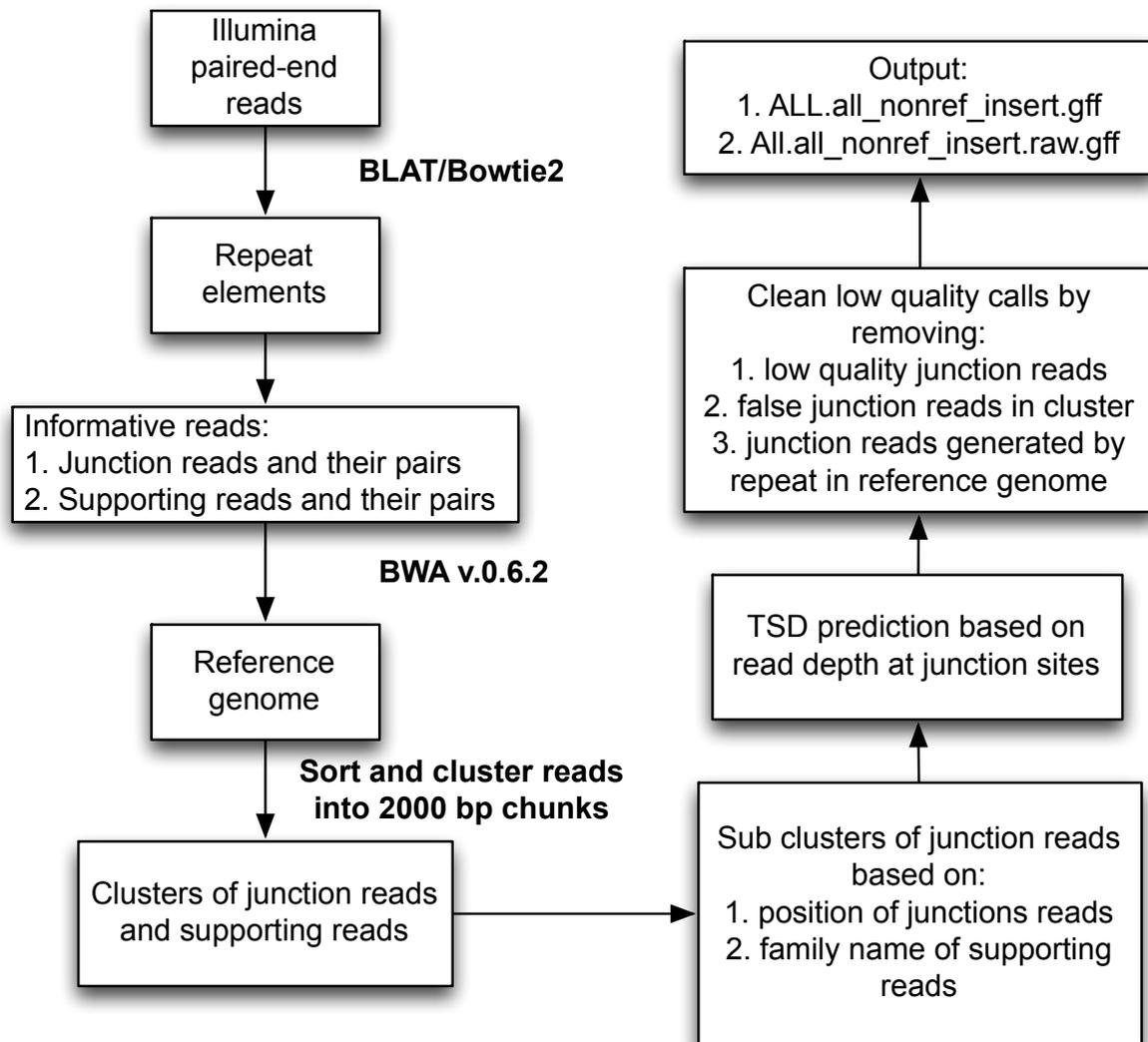
266

267 **References**

- 268 Bennetzen JL, and Wang H. 2014. The contributions of transposable elements to the
269 structure, function, and evolution of plant genomes. *Annu Rev Plant Biol* 65:505-530.
270 10.1146/annurev-arplant-050213-035811
- 271 Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, Stebbings LA, Leroy C,
272 Edkins S, Hardy C, Teague JW, Menzies A, Goodhead I, Turner DJ, Clee CM, Quail MA,
273 Cox A, Brown C, Durbin R, Hurler ME, Edwards PA, Bignell GR, Stratton MR, and
274 Futreal PA. 2008. Identification of somatically acquired rearrangements in cancer
275 using genome-wide massively parallel paired-end sequencing. *Nat Genet* 40:722-
276 729. 10.1038/ng.128
- 277 Cordaux R, and Batzer MA. 2009. The impact of retrotransposons on human genome
278 evolution. *Nat Rev Genet* 10:691-703. 10.1038/nrg2640
- 279 Cowley M, and Oakey RJ. 2013. Transposable elements re-wire and fine-tune the
280 transcriptome. *PLoS Genet* 9:e1003234. 10.1371/journal.pgen.1003234
- 281 Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nat*
282 *Rev Genet* 9:397-405. 10.1038/nrg2337
- 283 Hormozdiari F, Hajirasouliha I, Dao P, Hach F, Yorukoglu D, Alkan C, Eichler EE, and
284 Sahinalp SC. 2010. Next-generation VariationHunter: combinatorial algorithms for
285 transposon insertion discovery. *Bioinformatics* 26:i350-357.
286 10.1093/bioinformatics/btq216
- 287 Hu X, Yuan J, Shi Y, Lu J, Liu B, Li Z, Chen Y, Mu D, Zhang H, Li N, Yue Z, Bai F, Li H, and Fan
288 W. 2012. pIRS: Profile-based Illumina pair-end reads simulator. *Bioinformatics*
289 28:1533-1535. 10.1093/bioinformatics/bts187

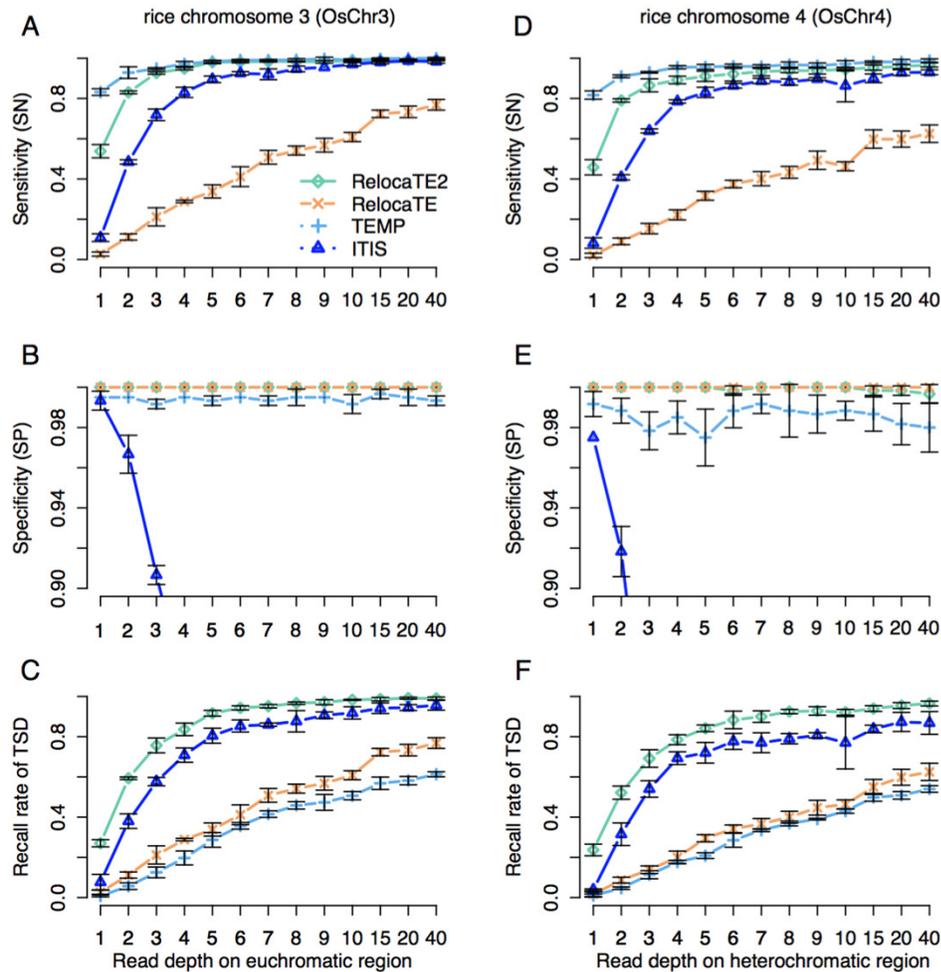
- 290 Jiang C, Chen C, Huang Z, Liu R, and Verdier J. 2015. ITIS, a bioinformatics tool for accurate
291 identification of transposon insertion sites using next-generation sequencing data.
292 *BMC Bioinformatics* 16:72. 10.1186/s12859-015-0507-2
- 293 Keane TM, Wong K, and Adams DJ. 2013. RetroSeq: transposable element discovery from
294 next-generation sequencing data. *Bioinformatics* 29:389-390.
295 10.1093/bioinformatics/bts697
- 296 Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* 12:656-664.
297 10.1101/gr.229202. Article published online before March 2002
- 298 Korbelt JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D,
299 Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP,
300 Hurler ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, and Snyder M. 2007.
301 Paired-end mapping reveals extensive structural variation in the human genome.
302 *Science* 318:420-426. 10.1126/science.1149504
- 303 Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, and Bourque G. 2010.
304 Transposable elements have rewired the core regulatory network of human
305 embryonic stem cells. *Nat Genet* 42:631-634. 10.1038/ng.600
- 306 Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF,
307 Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A,
308 Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA,
309 Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, and
310 Venter JC. 2007. The diploid genome sequence of an individual human. *PLoS Biol*
311 5:e254. 10.1371/journal.pbio.0050254
- 312 Li H, and Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler
313 transform. *Bioinformatics* 25:1754-1760. 10.1093/bioinformatics/btp324
- 314 Lisch D. 2013. How important are transposons for plant evolution? *Nat Rev Genet* 14:49-61.
315 10.1038/nrg3374
- 316 Lynch VJ, Leclerc RD, May G, and Wagner GP. 2011. Transposon-mediated rewiring of gene
317 regulatory networks contributed to the evolution of pregnancy in mammals. *Nat*
318 *Genet* 43:1154-1159. 10.1038/ng.917
- 319 Quadrana L, Bortolini Silveira A, Mayhew GF, LeBlanc C, Martienssen RA, Jeddloh JA, and
320 Colot V. 2016. The Arabidopsis thaliana mobilome and its impact at the species level.
321 *Elife* 5. 10.7554/eLife.15716
- 322 Robb SM, Lu L, Valencia E, Burnette JM, 3rd, Okumoto Y, Wessler SR, and Stajich JE. 2013.
323 The use of RelocaTE and unassembled short reads to produce high-resolution
324 snapshots of transposable element generated diversity in rice. *G3 (Bethesda)* 3:949-
325 957. 10.1534/g3.112.005348
- 326 Schatz MC, Maron LG, Stein JC, Hernandez Wences A, Gurtowski J, Biggers E, Lee H, Kramer
327 M, Antoniou E, Ghiban E, Wright MH, Chia JM, Ware D, McCouch SR, and McCombie
328 WR. 2014. Whole genome de novo assemblies of three divergent strains of rice,
329 *Oryza sativa*, document novel gene space of aus and indica. *Genome Biol* 15:506.
330 10.1186/PREACCEPT-2784872521277375
- 331 Stewart C, Kural D, Stromberg MP, Walker JA, Konkel MK, Stutz AM, Urban AE, Grubert F,
332 Lam HY, Lee WP, Busby M, Indap AR, Garrison E, Huff C, Xing J, Snyder MP, Jorde LB,
333 Batzer MA, Korbelt JO, Marth GT, and Genomes P. 2011. A comprehensive map of
334 mobile element insertion polymorphisms in humans. *PLoS Genet* 7:e1002236.
335 10.1371/journal.pgen.1002236

- 336 Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, and Wang T. 2014. Widespread
337 contribution of transposable elements to the innovation of gene regulatory
338 networks. *Genome Res* 24:1963-1976. 10.1101/gr.168872.113
- 339 Xing J, Zhang Y, Han K, Salem AH, Sen SK, Huff CD, Zhou Q, Kirkness EF, Levy S, Batzer MA,
340 and Jorde LB. 2009. Mobile elements create structural variation: analysis of a
341 complete human genome. *Genome Res* 19:1516-1526. 10.1101/gr.091827.109
- 342 Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, Li J, He W,
343 Zhang G, Zheng X, Zhang F, Li Y, Yu C, Kristiansen K, Zhang X, Wang J, Wright M,
344 McCouch S, Nielsen R, Wang J, and Wang W. 2012. Resequencing 50 accessions of
345 cultivated and wild rice yields markers for identifying agronomically important
346 genes. *Nat Biotechnol* 30:105-111. 10.1038/nbt.2050
- 347 Zhao Q, Zhang Y, Cheng Z, Chen M, Wang S, Feng Q, Huang Y, Li Y, Tang Y, Zhou B, Chen Z, Yu
348 S, Zhu J, Hu X, Mu J, Ying K, Hao P, Zhang L, Lu Y, Zhang LS, Liu Y, Yu Z, Fan D, Weng
349 Q, Chen L, Lu T, Liu X, Jia P, Sun T, Wu Y, Zhang Y, Lu Y, Li C, Wang R, Lei H, Li T, Hu
350 H, Wu M, Zhang R, Guan J, Zhu J, Fu G, Gu M, Hong G, Xue Y, Wing R, Jiang J, and Han
351 B. 2002. A fine physical map of the rice chromosome 4. *Genome Res* 12:817-823.
352 10.1101/gr.48902
- 353 Zhuang J, Wang J, Theurkauf W, and Weng Z. 2014. TEMP: a computational method for
354 analyzing transposable element polymorphism in populations. *Nucleic Acids Res*
355 42:6826-6838. 10.1093/nar/gku323



357

358 **Figure 1. Workflow for identification of transposable element insertions in population**359 **resequencing data using Illumina paired-end reads.**



360

361 **Figure 2. Performance of RelocaTE2, RelocaTE, TEMP and ITIS on simulated rice data.**

362 Comparison of tool performance on rice chromosome 3 for Sensitivity (A) Specificity (B) Recall

363 rate of TSD (C) and comparison of performance on rice chromosome 4 for Sensitivity (D),

364 Specificity (E), Recall rate of TSD (F). Simulations of 200 random transposable element (TE)

365 insertions were generated for rice chromosome 3 (OsChr3) and rice chromosome 4 (OsChr4)

366 with three replicates. A series of datasets with varying sequence depths (from 1 to 40) were

367 generated for each simulation dataset. Sensitivity (SN), Specificity (SP) and Recall rate of target

368 site duplication (TSD) of each tool were estimated for each of these datasets and plotted against

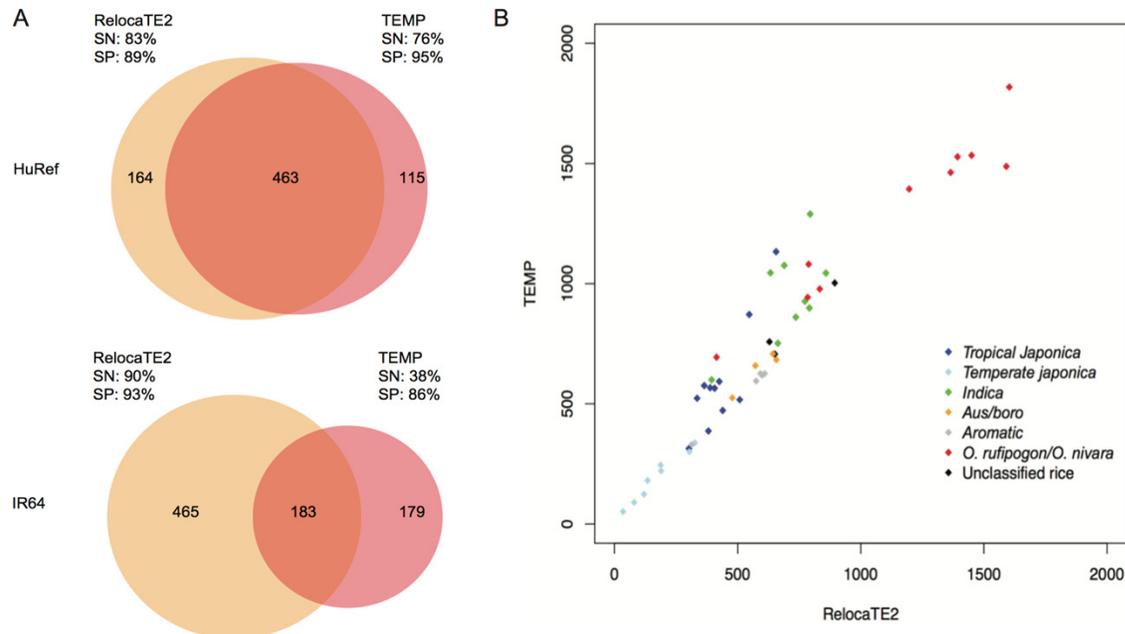
369 sequence depth. The error bars show the standard deviation of three replicates with different sets

370 of 200 random TE insertions. SN was defined as the percentage of calls within 100 base pairs of

371 200 random TE insertions. SP was defined as the percentage of calls not within 100 base pairs of

372 200 random TE insertions. Recall rate of TSD was defined as the percentage of true positive

373 calls that correctly matched the simulated TSD of TE insertions.



374
 375 **Figure 3. Performance of RelocaTE2 and TEMP on biological dataset in HuRef genome,**
 376 **IR64 genome, and 50 rice and wild rice strains. A.** Venn diagram of the overlap in non-
 377 reference TE insertions identified in the HuRef genome and the rice IR64 genome using
 378 RelocaTE2 and TEMP. Sensitivity (SN) and Specificity (SP) were assessed by comparing the
 379 assembled HuRef genome to the GRCh36 reference genome and the assembled IR64 genome to
 380 the MSU7 reference genome. SN was defined as the percentage of validated calls out of all
 381 validated calls by either RelocaTE2 or TEMP. SP was defined as the percentage of validated
 382 calls out of all calls by each tool. **B.** Comparison of the number of non-reference TE insertions of
 383 14 TE families in 50 rice and wild rice strains identified by RelocaTE2 and TEMP. Strains are
 384 color-coded based on subpopulation classification.