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

Jinfeng Chen[1,2,3], Travis R. Wrightsman[2], Susan R. Wessler[2,3] and Jason E. Stajich[1,3,*]

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

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

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

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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2447v1 | CC BY 4.0 Open Access | rec: 13 Sep 2016, publ: 13 Sep 2016

Abstract

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 transposition events or polymorphisms 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 polymorphisms 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 demonstrates a higher level of sensitivity and specificity when compared to other tools. 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 polymorphisms and can be incorporated into analysis workflows in support of describing the complete genotype from light coverage genome sequencing.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2447v1 | CC BY 4.0 Open Access | rec: 13 Sep 2016, publ: 13 Sep 2016

Introduction

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

Two categories of bioinformatics tools have been developed to identify TE polymorphisms from population resequencing data. One type employs a strategy similar to that used to discover structural variations. These tools identify discordant pairs of sequence reads based on the chromosomal position of read alignments to indicate genomic inversions, insertions, deletions or other complex rearrangements (Campbell et al. 2008; Korbel et al. 2007). Software for TE mapping scrutinize genomic loci with discordant read pairs to see if known TE sequences are can be implicated near the rearrangement site. These tools, such as Retroseq (Keane et al. 2013) and TEMP (Zhuang et al. 2014), are generally highly sensitive and can locate insertion sites to a 10-50 bp resolution. A second category of tools first identify by similarity, any sequence reads containing known TE sequences. The tools excise the TE sequence from the reads and search the remaining 5' or 3' flanking sequence against the host organism genome sequence to find the element's genomic location. These tools, including RelocaTE (Robb et al. 2013), T-lex2 (Fiston-Lavier et al. 2015), and ITIS (Jiang et al. 2015), are able to detect the exact location of insertion sites and TSDs characteristic of TEs. This second category of tools is ideal for identifying new insertions from population resequencing data because it can accurately detect an insertion location and identify the sequence of TSD. However, most of these tools are designed to search with only a single TE at a time, which sacrifices speed for increased sensitivity and specificity.

The extended runtime limits the feasibility of applying these tools when searching thousands of TEs in hundreds or thousands of individuals.

In RelocaTE2, an improved version of RelocaTE, we implement a junction-based approach that can search multiple template TEs in the same pass through the sequencing data, streamlining the computational approach. Using simulated datasets, we show that RelocaTE2 is highly sensitive even in low coverage resequencing data or on chromosomes with high repetitive sequence content has a specificity of greater than 99%. RelocaTE2 performed as the most sensitive and specific tool in our tests profiling human and rice population genomics data and can be widely used for analyzing population dynamics of TEs.

Materials & Methods

RelocaTE2 Workflow

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

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

Sequence reads comprised entirely of repeat elements are ignored, but their read pair is kept (denoted as supporting reads). These junction, full, and supporting reads, are all aligned to the reference genome using BWA (v0.6.2) with the default setting `"-l 32 -k 2"` (Li & Durbin 2009). Mapped reads are sorted by chromosome order and windows of 2,000 bp are evaluated to

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2447v1 | CC BY 4.0 Open Access | rec: 13 Sep 2016, publ: 13 Sep 2016

define insertion clusters. In each insertion cluster, additional subclusters are further refined based on the mapping position of junction reads to address the possible scenario of multiple insertion sites within a window. TSD position and sequence are identified if the subcluster is supported by junction reads from both upstream and downstream of the TE insertion site.

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

Simulated data for evaluation of TE insertion tools

Simulated datasets were created by randomly inserting TEs into sequence records of chromosomes 3 (OsChr3) and 4 (OsChr4) of rice (*Oryza sativa japonica*). OsChr3 is primarily made up of euchromatic regions, whereas OsChr4 has the sequence complexity consistent with heterochromatic regions and is a typical feature of many plant genomes (Zhao et al. 2002). Fourteen TEs families found in rice genomes comprised of 7 DNA Transposons: *mPing*, *nDart*, *Gaijin*, *spmlike*, *Truncator*, *mGing*, *nDarz* and 7 RNA Retrotransposons: *Bajie*, *Dasheng*, *Retro1*, *RIRE2*, *RIRE3, Copia2*, *karma*, were used. The insertion simulations were performed by choosing 200 random insertion sites on each chromosome in three independent replicates. Each simulated insertion site was generated by selecting one random chromosome position and then one random TE and TSDs of the expected size was generated for each TE. After generating 200 random sites and TE assignments, a new genome sequence was generated with the TEs inserted at corresponding locations. A GFF3 file with the insertion locations recorded was produced to support the evaluation of the performance of each tool on the dataset. Paired-end reads of all simulated chromosomes were simulated by pIRS (pirs simulate –l 100 –x coverage –m 500 –v

100) (Hu et al. 2012). For each dataset, simulate sequence reads at sequence depths of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 40-fold coverage were generated.

Real sequence data for evaluation of TE insertion tools

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

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

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

Results and Discussions

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

RelocaTE2 was first compared to RelocaTE, TEMP and ITIS using the simulated datasets. Each dataset of simulated rice chromosomes, OsChr3 and OsChr4, was virtually sheared to simulated paired-end short reads at a coverage ranging from 1X to 40X. At high sequencing coverage (≥10X), RelocaTE2, TEMP and ITIS were able to identify >99% of simulated insertions on OsChr3, whereas the performance of RelocaTE was much lower (85%) (Fig.2A). At lower sequencing coverage, e.g. 3X, only RelocaTE2 and TEMP were able to achieve ≥95% sensitivity on OsChr3 (Fig.2A). Furthermore, TEMP was able to identify 83% and 93% of simulated insertions on OsChr3 at very low sequence coverage of 1X and 2X, respectively (Fig.2A). RelocaTE2 had a sensitivity of 53% and 83% on OsChr3 for the 1X and 2X coverage due to the removal of TE insertions supported by only one read (singleton) or supported by reads from only one end of TE insertions (insufficient insertions), which can result in many false positives (Fig.2A).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2447v1 | CC BY 4.0 Open Access | rec: 13 Sep 2016, publ: 13 Sep 2016

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

Evaluation of RelocaTE2 and TEMP on biological datasets

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

RelocaTE2 and TEMP were used to analyze data from the rice strain IR64 and the results were evaluated by comparing the genome assembly of IR64 with MSU7. RelocaTE2 identified 648 insertion sites while the genome comparison revealed that 93% of insertions were true positives (Fig.3A). TEMP identified 362 insertions, of which 50% (183/362) overlapped with RelocaTE2 (Fig.3A). The specificity of TEMP was estimated to be 86%, slightly lower than RelocaTE2 (93%) (Fig.3A). However, TEMP was found to be less sensitive than RelocaTE2 in the rice genome, only calling 362 sites as compared to 648 by RelocaTE2 (38% vs. 90%, Fig.3A).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2447v1 | CC BY 4.0 Open Access | rec: 13 Sep 2016, publ: 13 Sep 2016

Moreover, RelocaTE2 predicted TE insertion junctions of 3 ± 1 bp, which was much smaller than TEMP (393 ± 199 bp).

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

Runtime performance

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

Conclusions

We present RelocaTE2 as a new tool for mapping TE polymorphisms to base-pair resolution from resequencing data. RelocaTE2 identifies multiple TE families in a single search with high sensitivity and specificity. The evaluation of these tools on simulated and biological datasets support the use of RelocaTE2 for analysis of genomes of plants and animals and indicate it can generate very high quality genotyping of TE insertions from resequencing datasets of modest

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2447v1 | CC BY 4.0 Open Access | rec: 13 Sep 2016, publ: 13 Sep 2016

sequencing depths. The high resolution mapping of TE insertions sites will enable detailed analysis of the interaction of TEs and genes and as structural variations that vary in populations.

Competing Interests

The authors declare that there are no competing interests.

Author Contributions

Jinfeng Chen conceived and designed the study, wrote the code, analyzed the data, wrote the paper, and prepared figures and/or tables.

Travis R. Wrightsman tested the code, wrote the manual, and reviewed drafts of the papers.

Susan R. Wessler and Jason E. Stajich conceived and designed the study, wrote the paper, and reviewed drafts of the papers.

Data Availability

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

Acknowledgements

Funding

This research is supported by US National Science Foundation grant IOS-1027542 and a grant from the W. M. Keck Foundation to SR Wessler and JE Stajich and funds to SR Wessler from the Howard Hughes Medical Institute 52008110 and US Department of Agriculture Grant 2013-38422-20955.

References

Bennetzen JL, and Wang H. 2014. The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu Rev Plant Biol* 65:505-530. 10.1146/annurev-arplant-050213-035811

Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, Stebbings LA, Leroy C, Edkins S, Hardy C, Teague JW, Menzies A, Goodhead I, Turner DJ, Clee CM, Quail MA, Cox A, Brown C, Durbin R, Hurles ME, Edwards PA, Bignell GR, Stratton MR, and Futreal PA. 2008. Identification of somatically acquired rearrangements in cancer

using genome-wide massively parallel paired-end sequencing. *Nat Genet* 40:722-729. 10.1038/ng.128
Cordaux R, and Batzer MA. 2009. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* 10:691-703. 10.1038/nrg2640
Cowley M, and Oakey RJ. 2013. Transposable elements re-wire and fine-tune the transcriptome. *PLoS Genet* 9:e1003234. 10.1371/journal.pgen.1003234
Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* 9:397-405. 10.1038/nrg2337
Fiston-Lavier AS, Barron MG, Petrov DA, and Gonzalez J. 2015. T-lex2: genotyping, frequency estimation and re-annotation of transposable elements using single or pooled next-generation sequencing data. *Nucleic Acids Res* 43:e22. 10.1093/nar/gku1250
Hormozdiari F, Hajirasouliha I, Dao P, Hach F, Yorukoglu D, Alkan C, Eichler EE, and Sahinalp SC. 2010. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics* 26:i350-357. 10.1093/bioinformatics/btq216
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 W. 2012. pIRS: Profile-based Illumina pair-end reads simulator. *Bioinformatics* 28:1533-1535. 10.1093/bioinformatics/bts187
Jiang C, Chen C, Huang Z, Liu R, and Verdier J. 2015. ITIS, a bioinformatics tool for accurate identification of transposon insertion sites using next-generation sequencing data. *BMC Bioinformatics* 16:72. 10.1186/s12859-015-0507-2
Keane TM, Wong K, and Adams DJ. 2013. RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics* 29:389-390. 10.1093/bioinformatics/bts697
Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* 12:656-664. 10.1101/gr.229202. Article published online before March 2002
Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, and Snyder M. 2007.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2447v1 | CC BY 4.0 Open Access | rec: 13 Sep 2016, publ: 13 Sep 2016

Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318:420-426. 10.1126/science.1149504

Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, and Bourque G. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* 42:631-634. 10.1038/ng.600

Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ, 3rd, Lohr JG, Harris CC, Ding L, Wilson RK, Wheeler DA, Gibbs RA, Kucherlapati R, Lee C, Kharchenko PV, Park PJ, and Cancer Genome Atlas Research N. 2012. Landscape of somatic retrotransposition in human cancers. *Science* 337:967-971. 10.1126/science.1222077

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, and Venter JC. 2007. The diploid genome sequence of an individual human. *PLoS Biol* 5:e254. 10.1371/journal.pbio.0050254

Li H, and Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754-1760. 10.1093/bioinformatics/btp324

Lisch D. 2013. How important are transposons for plant evolution? *Nat Rev Genet* 14:49-61. 10.1038/nrg3374

Lynch VJ, Leclerc RD, May G, and Wagner GP. 2011. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat Genet* 43:1154-1159. 10.1038/ng.917

Quadrana L, Bortolini Silveira A, Mayhew GF, LeBlanc C, Martienssen RA, Jeddeloh JA, and Colot V. 2016. The Arabidopsis thaliana mobilome and its impact at the species level. *Elife* 5. 10.7554/eLife.15716

Robb SM, Lu L, Valencia E, Burnette JM, 3rd, Okumoto Y, Wessler SR, and Stajich JE. 2013. The use of RelocaTE and unassembled short reads to produce high-resolution snapshots of transposable element generated diversity in rice. *G3 (Bethesda)* 3:949-957. 10.1534/g3.112.005348

Schatz MC, Maron LG, Stein JC, Hernandez Wences A, Gurtowski J, Biggers E, Lee H, Kramer M, Antoniou E, Ghiban E, Wright MH, Chia JM, Ware D, McCouch SR, and McCombie WR. 2014. Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica. *Genome Biol* 15:506. 10.1186/PREACCEPT-2784872521277375

Stewart C, Kural D, Stromberg MP, Walker JA, Konkel MK, Stutz AM, Urban AE, Grubert F, Lam HY, Lee WP, Busby M, Indap AR, Garrison E, Huff C, Xing J, Snyder MP, Jorde LB, Batzer MA, Korbel JO, Marth GT, and Genomes P. 2011. A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet* 7:e1002236. 10.1371/journal.pgen.1002236

Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, and Wang T. 2014. Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res* 24:1963-1976. 10.1101/gr.168872.113

Xing J, Zhang Y, Han K, Salem AH, Sen SK, Huff CD, Zhou Q, Kirkness EF, Levy S, Batzer MA, and Jorde LB. 2009. Mobile elements create structural variation: analysis of a complete human genome. *Genome Res* 19:1516-1526. 10.1101/gr.091827.109

Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, Li J, He W, Zhang G, Zheng X, Zhang F, Li Y, Yu C, Kristiansen K, Zhang X, Wang J, Wright M, McCouch S, Nielsen R, Wang J, and Wang W. 2012. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat Biotechnol* 30:105-111. 10.1038/nbt.2050

Zhao Q, Zhang Y, Cheng Z, Chen M, Wang S, Feng Q, Huang Y, Li Y, Tang Y, Zhou B, Chen Z, Yu S, Zhu J, Hu X, Mu J, Ying K, Hao P, Zhang L, Lu Y, Zhang LS, Liu Y, Yu Z, Fan D, Weng 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 H, Wu M, Zhang R, Guan J, Zhu J, Fu G, Gu M, Hong G, Xue Y, Wing R, Jiang J, and Han B. 2002. A fine physical map of the rice chromosome 4. *Genome Res* 12:817-823. 10.1101/gr.48902

Zhuang J, Wang J, Theurkauf W, and Weng Z. 2014. TEMP: a computational method for analyzing transposable element polymorphism in populations. *Nucleic Acids Res* 42:6826-6838. 10.1093/nar/gku323

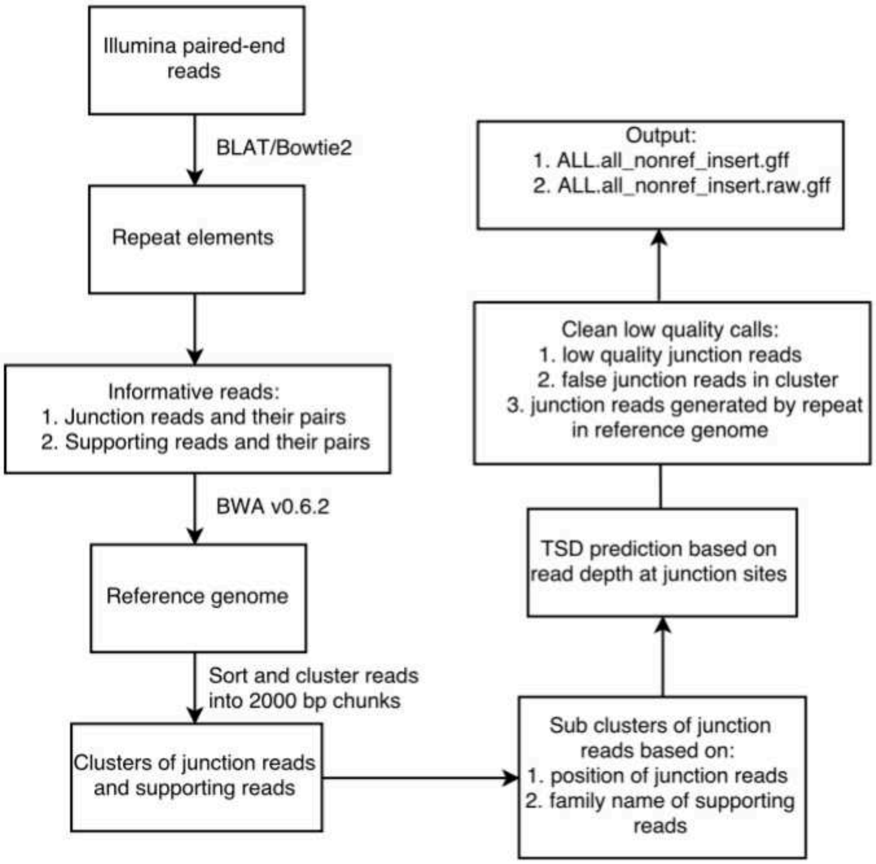

Figure 1. Workflow for identification of transposable element insertions in population resequencing data using Illumina paired-end reads.

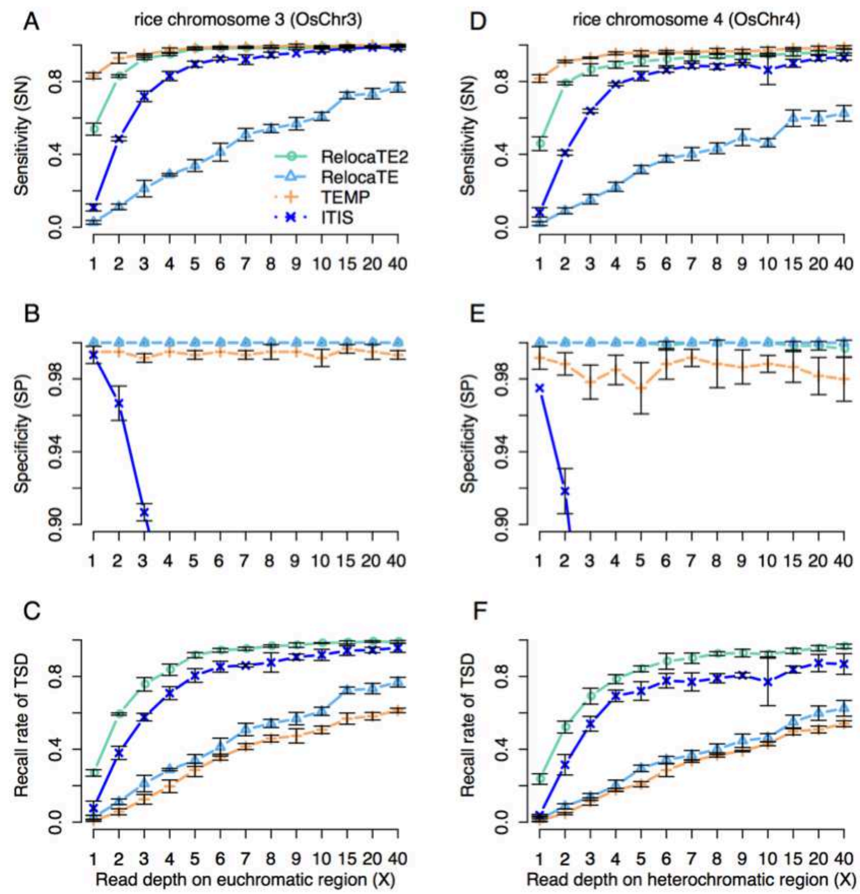

Figure 2. Performance of RelocaTE2, RelocaTE, TEMP and ITIS on simulated rice data. Simulations of 200 random transposable element (TE) insertions were generated for rice chromosome 3 (OsChr3) and rice chromosome 4 (OsChr4) with three replicates. A series of datasets with different sequence depths (from 1X to 40X) were generated for each simulation dataset. Sensitivity (SN), Specificity (SP) and Recall rate of target site duplication (TSD) of each tool were estimated for each of these datasets and plotted against sequence depth. The error bars show the standard deviation of three replicates with different sets of 200 random TE insertions. SN was defined as the percentage of calls within 100 base pairs of 200 random TE insertions. SP was defined as the percentage of calls not within 100 base pairs of 200 random TE insertions. Recall rate of TSD was defined as the percentage of true positive calls that correctly matched the simulated TSD of TE insertions. The results illustrate how tools perform on chromosomes which are primarily euchromatic or heterochromatic using OsChr3 and OsChr4 respectively.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2447v1 | CC BY 4.0 Open Access | rec: 13 Sep 2016, publ: 13 Sep 2016

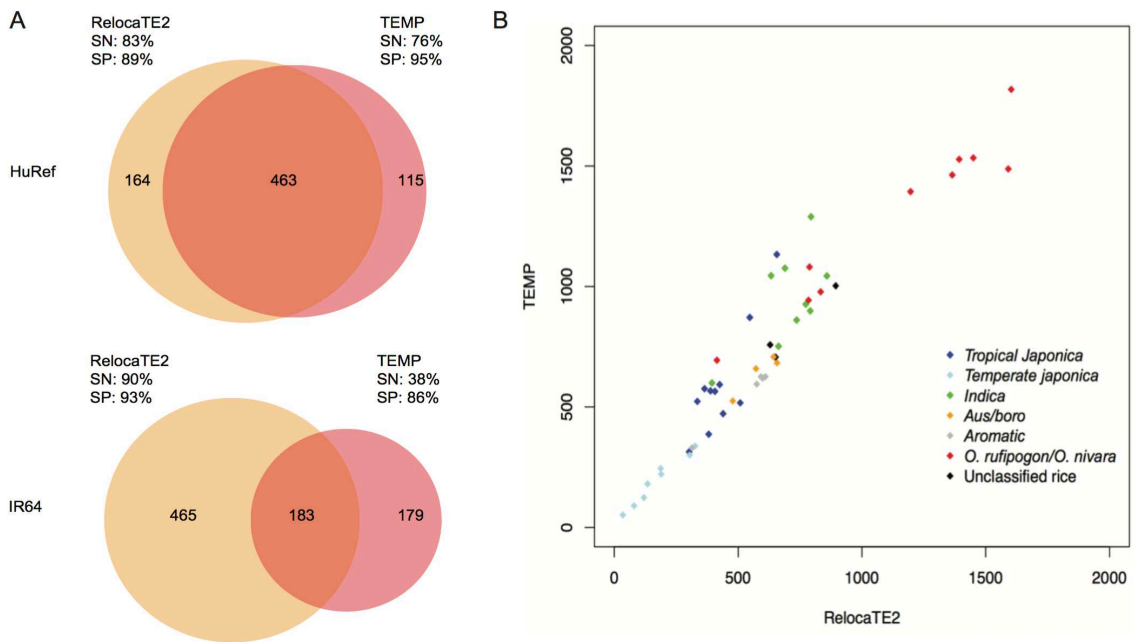

Figure 3. Performance of RelocaTE2 and TEMP on biological dataset in HuRef genome, IR64 genome, and 50 rice and wild rice strains. A. Venn diagram of the overlap in non-reference TE insertions identified in the HuRef genome and the rice IR64 genome using RelocaTE2 and TEMP. Sensitivity (SN) and Specificity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