Efficient screening of long terminal repeat retrotransposons that show high insertion polymorphism via high-throughput sequencing of the PBS site

Yuki Monden, Nobuyuki Fujii, Kentaro Yamaguchi, Kazuho Ikeo, Yoshiko Nakazawa, Takamitsu Waki, Keita Hirashima, Yosuke Uchimura, and Makoto Tahara

Y. Monden, K. Yamaguchi, and M. Tahara. Graduate School of Environmental and Life Science, Okayama University, 1-1-1 Tsushimanaka Kitaku, Okayama, Okayama 700-8530, Japan.
N. Fujii, and K. Ikeo. Center for Information Biology, National Institute of Genetics Research Organization of Information and Systems, Yata 1111, Mishima, Shizuoka 411-8540, Japan.
Y. Nakazawa, and T. Waki. Biotechology Division, Tochigi Prefectural Agricultural Experiment Station, 1080 Kawaraya-cho, Utsunomiya, Tochigi 320-0002, Japan.
K. Hirashima, and Y. Uchimura. Department of Research Plan and Strategy, Fukuoka Agricultural Research Center, 587 Yoshiki, Chikushino, Fukuoka 818-8549, Japan.

* **Corresponding author:** M. Tahara (e-mail: tahara@cc.okayama-u.ac.jp)

Abstract: Retrotransposons have been used frequently for the development of molecular markers by using their insertion polymorphisms among cultivars, because multiple copies of these elements are dispersed throughout the genome and inserted copies are inherited genetically. Although a large number of long terminal repeat (LTR) retrotransposon families exist in the higher eukaryotic genomes, the identification of families that show high insertion polymorphism has been challenging. Here, we performed an efficient screening of these retrotransposon families using an Illumina HiSeq2000 sequencing platform with comprehensive LTR library construction based on the primer binding site (PBS), which is located adjacent to the 5′ LTR and has a motif that is universal and conserved among LTR retrotransposon families. The paired-end sequencing library of the fragments containing a large number of LTR sequences and their insertion sites was sequenced for seven strawberry cultivars (*Fragaria* x *ananassa* Duchesne) and one diploid wild species (*F. vesca*). Among them, we screened 24 families with a "unique" insertion site that appeared only in one variety and not in any others, assuming that this type of insertion should have occurred quite recently. Finally, we confirmed experimentally the selected LTR families showed high insertion polymorphisms among closely related cultivars.

Key words: retrotransposon, primer binding site, high-throughput sequencing, polymorphism, molecular markers.

Introduction

Retrotransposons are the major component of eukaryotic genomes and have contributed to their evolution and diversification (Kumar and Bennetzen 1999; Feschotte et al. 2002; Wessler 2006). Among them, the LTR retrotransposon is ubiquitous and abundant in the genomes of higher plants (approximately ~80% in the barley genome and >70% in the maize genome) (Feschotte et al. 2002; Schnable et al. 2009; Mayer et al. 2012) and its integration system, which generates inheritable insertion without excision, is well known (Kumar and Bennetzen 1999; Havecker et al. 2004; Levin and Moran 2011). As insertions of these elements with high copy number are dispersed throughout the genome, their insertion polymorphisms among cultivars have been used as useful molecular markers in the map-based cloning of genes and in phylogenetic analyses, and/or to analyze genetic diversity (Kumar and Hirochika 2001; Kalendar 2011; Kalendar et al. 2011; Poczai et al. 2013). Several molecular markers based on retrotransposon insertion polymorphisms have been developed, such as the sequence-specific amplified polymorphism (S-SAP) (Waugh et al. 1997; Syed et al. 2005), inter-retrotransposon amplification polymorphism (IRAP) (Kalendar et al. 1999), retrotransposon microsatellite amplification polymorphism (REMAP) (Kalendar et al. 1999), and retrotransposon-based insertion polymorphism (RBIP) (Flavell et al. 1998). To use these molecular markers fully, LTR families with high insertion polymorphisms among cultivars need to be identified. However, the identification of these retrotransposon families within the genome has been challenging, because most retrotransposon families with high copy number transposed long before modern cultivars were developed, and they are transpositionally inactive in the present genome (Kumar and Bennetzen 1999; Slotkin and Martienssen 2007; Lisch 2009); in this case, they hardly show insertion polymorphisms among cultivated varieties.

A previous study showed that PCR using a conserved motif, namely the primer binding site (PBS), which is located adjacent to the 5′ LTR sequence, allowed the screening of all possible LTR sequences with a conserved PBS in the genome (Kalendar et al. 2010). This method is based on the nearly universal use of host tRNA as primers by both retroviruses and LTR retrotransposons to initiate reverse transcription during their replication cycle (with a few exceptions, such as *Tf1*/*sushi* in fungi and vertebrates and *Fourf* in maize, which are able to self-prime cDNA synthesis) (Marquet et al. 1995; Mak and Kleiman 1997; Kelly et al. 2003; Hizi 2008). Thus, primers designed to match the conserved regions of the PBS proved to be very effective for the isolation of a wide range of LTR retrotransposons, including the nonautonomous terminal repeat retrotransposons in miniature (TRIM) and large retrotransposon derivatives (LARDs), which have no internal coding regions (Kalendar et al. 2010).

The recent development of next-generation sequencing (NGS) technologies has allowed the generation of a vast amount of sequence data at low cost and in a short time. As multiple copies of retrotransposons are dispersed throughout eukaryotic genomes, the genome-wide screening of these elements and their characterization require the large volume of sequencing data that is considered to be achieved by these NGS resources (Xing et al. 2013). In fact, NGS resources have enabled the sequencing of a massive number of transposable element (TE) insertion sites of targeted families in several species (Iskow et al. 2010; Ewing and Kazazian 2011; Kofler et al. 2012; Urbański et al. 2012). Moreover, the genome-wide characterization of several repetitive elements was performed using 454 sequencing and subsequent clustering analysis of the reads (Macas et al. 2007; Novák et al. 2010; Pagán et al. 2012). However, no reports have focused on the identification of retrotransposon families that

show high insertion polymorphism among cultivars, which represent a remarkably small portion of the whole group of TE families, but are useful for genetic analysis using NGS platforms.

Here, we developed an efficient approach to identify LTR retrotransposon families showing diverse insertion patterns among cultivars using the Illumina HiSeq2000 sequencing platform. By exploiting a conserved PBS motif in PCR amplification for paired-end sequencing using a multiplex barcoding system, we were able to identify a large number of LTR sequences and their insertion sites in several strawberry cultivars. Out of these insertion sites, we extracted “unique” insertion sites that were present only in one cultivar and not in others, because the insertion at these sites has not yet been shared through sexual reproduction after the insertion event; thus, it should be relatively recent. The pooling and clustering of LTR sequences corresponding to the unique insertion sites led to the acquisition of 24 LTR sequence candidates. Finally, we confirmed that the LTR sequences identified showed high insertion polymorphisms among closely related cultivars by displaying and comparing the insertions.

Materials and methods

DNA samples

The plants of strawberry cultivars and its wild species *F. vesca* **(Tables S1 and S2)** were obtained from the Tochigi Prefectural Agricultural Experiment Station, Fukuoka Agricultural Research Center, and Kyushu Okinawa Agricultural Research Center of National Agriculture and Food Research Organization. Genomic DNA was extracted from young leaves using the DNeasy plant mini kit (QIAGEN) according to the manufacturer's protocol.

Selection of the PBS sequences and primer design

The LTR_STRUC application was used to screen PBS sequences in the strawberry genome (*F. vesca* v1.1 scaffolds). The output files (*.rpt.txt) from LTR_STRUC showed the PBS sequence as the 26 nt located after the identified 3′ end of the LTR at the 5′ side. However, in some cases, this sequence seemed incorrectly located. This program predicts LTR retrotransposons based on (1) the distance between an LTR pair and (2) the sequence similarity between the paired LTRs (McCarthy and McDonald 2003). The program selects regions with high similarity to an LTR sequence. As a result, a putative LTR sequence does not contain the whole LTR region; alternatively, a putative PBS sequence may contain the remaining 3′ end of the LTR sequence. To avoid the incorrect identification of the PBS sequence, we chose putative PBS sequences that fulfilled either of the following requirements: (1) an output of 26 bases containing "TGG" trinucleotides 0–5 bp from the 5′ end; or (2) an output of 26 bases containing "CA" dinucleotides and "TGG" trinucleotides starting at 0–5 bp after "CA". This TGG motif is complementary to the "CCA" 3′ terminal sequence of all tRNAs, which is added posttranscriptionally and does not appear tRNA genes. Thus, the

designing of PCR primers including the CCA sequence enabled us to amplify specifically products from the PBS sequence, and not from tRNA genes (Kalendar et al. 2010).

The extracted PBS sequences were aligned and clustered after trimming to a size of 12 bp, including the TGG motif (Table 1). These PBS sequences were searched in the Soybean and Maize tRNA database (Table 1). Using two types of PBS sequences that matched part of the sequence of the iMET and Asp tRNA genes, we designed PCR primers to amplify toward the LTR (Table 1 and **Table S3**).

Preparation of libraries for next-generation sequencing

Five microgram of DNA from eight cultivars were fragmented using g-TUBE (Covaris) according to the manufacturer's protocol. These DNA samples were end-repaired, modified to add 3′ A overhangs, and ligated to the forked adaptors. The ligation products underwent primary amplification using AP2 Type1-4 and PBS (iMET and Asp) primer combinations, and secondary amplification using AP3 Type1-4 and PBS (iMET and Asp) primer combinations. Two PBS primers were mixed and used according to their genomic configuration (91:12). These PCR products were size-selected (300–500 bp) by gel electrophoresis using Pippin Prep (Sage Science). Each library was qualified by Bioanalyzer (Agilent Technologies, Inc., Santa Clara, California, USA). The eight libraries were pooled (Fig. 1) into one sequencing sample. Paired-end sequencing reads were generated on an Illumina HiSeq2000 platform. The information on forked adaptors and PCR primer sequences is listed in **Table S3**.

Sequencing analysis pipeline

Paired-end reads of 101 bp were obtained in fastq format: the read from one side supposedly contained the sequence of the PBS and LTR junction, whereas the read

158 from the other side contained that of the insertion site or an upstream part of the same
159 LTR (Fig. 1). Sequencing data were handled on the analysis pipeline execution system
160 of the Cell Innovation program at the National Institute of Genetics (NIG), which is
161 named Maser (http://cell-innovation.nig.ac.jp/index_en.html). Read pairs were filtered
162 and assigned by PBS (iMET and Asp) and 7–8 bp barcode sequences of the cultivar.
163 When one read of a sequence pair was filtered for invalid PBS or barcode sequence,
164 the entire pair was discarded.

165 The filtered non-PBS reads in each variety were aligned to the *F. vesca* v1.1
166 scaffold sequences using the Burrows–Wheeler alignment tool (BWA) (Li and Durbin
167 2009) with default options after trimming the adaptor sequence using cutadapt
168 (https://code.google.com/p/cutadapt/). The resulting BAM files were processed with a
169 perl script designed to extract uniquely mapped reads. To determine insertion loci, we
170 computed the coverage of aligned reads on every 10 kb window using SAMtools
171 (version 0.1.18) (Li et al. 2009) and BEDtools (version 2.13.3) (Quinlan and Hall
172 2010), and merged eight files into one file using perl script. By processing this file
173 with DEGseq (R package in Bioconductor) (Wang et al. 2010) and a custom perl script,
174 the amount of aligned reads for each window was represented as a log 2 ratio of one
175 sample to all others after normalization using the total number of aligned reads. We
176 calculated the *P* values for each window with Fisher's exact test, to perform a
177 statistical comparison analysis. The resulting *P* values were modified with a false-
178 discovery rate for multiple testing (Storey 2002; Benjamini and Hochberg 2013). We
179 considered the loci at which one sample had a significantly higher number of reads
180 compared with others as "presence"; in contrast, loci at which one sample had a
181 significantly lower number of reads compared with others were considered as "absence"
182 (these were tested statistically using Fisher's exact test, $P < 0.001$). The list of

chromosomal positions of putative LTR insertions and their presence/absence genotype in each sample was obtained. The concentric circles shown in Fig. 2 were drawn using the circus program (Krzywinski et al. 2009).

The putative unique sites were determined according to their presence in only one sample and absence in others. PBS reads corresponding to those unique insertions were extracted. After trimming the PBS sequences using cutadapt, the reads were clustered into distinct LTR groups using SlideSort (version 2) (Shimizu and Tsuda 2011) and aligned with MAFFT (Kato et al. 2002).

Sequence-specific amplified polymorphism (S-SAP)

Genomic DNA was digested using the *Mse*I **or *Rsa*I** restriction enzyme and ligated to forked adaptors. **Forked adaptors were prepared by annealing two different oligonucleotides (FA_*Mse*I and FA_cmpl for *Mse*I digested DNA and FA_*Rsa*I and FA_cmpl for *Rsa*I digested DNA).** We designed an LTR-specific primer that matched the end sequence of the LTR sequence identified. Primary PCR was performed using the AP2_Type1 and LTR-specific (Met_CL*_1st) primer sets , and nested PCR was performed using AP3_Type1 and LTR-specific (Met_CL*_2nd) primer combinations. The PCR products were loaded on the Applied Biosystems 3500 Genetic Analyzer (Life Technologies) for DNA fragment analysis. The results were visualized using the GeneMapper software (Life Technologies). The information regarding primer and **adaptor** sequences is listed in **Table S3**.

Results

Determination of the PBS sequence

The PBS is located adjacent to the 5′ LTR, and its sequence is complementary to those located at the 3′ end of the specific host tRNAs and is highly conserved among LTR families (Marquet et al. 1995; Mak and Kleiman 1997; Kelly et al. 2003; Hizi 2008). To identify this sequence, we scanned the 256 scaffold sequences covering 223 Mb of the *F. vesca* genome v1.1 assembly (http://www.rosaceae.org/projects/strawberry_genome/v1.1/assembly) using the LTR_STRUC software, which searches and identifies LTR retrotransposons (McCarthy and McDonald 2003). A total of 375 types of PBS sequences with a length of 26 nucleotides (nt) were identified in this search.

Among the 26-nt sequences output, we extracted those that fulfilled either one of the following conditions: (1) the output contained the "TGG" sequence starting at the 0–5 nt positions of the 5′ end, or (2) the output contained the "CA" sequence followed by the "TGG" sequence with a 0–5 nt interval, because the PBS sequence is adjacent to the 3′ end sequence, "CA", of the 5' LTR with 0–5 intervening nts and the LTR_STRUC searches often result in shorter LTR sequences at the 3′ end (Fig. 1). A total of 210 PBS sequences fulfilled one of the conditions; each sequence was then trimmed to a length of 12 nt with the starting sequence of "TGG". The simple alignment of these 12-nt sequences resulted in 61 putative PBS sequence groups, with the largest and second-largest groups containing 91 and 12 sequences, respectively (Table 1). The sequence of the largest group was homologous to that of the iMET tRNA, whereas the sequence of the second group was homologous to that of the Asp tRNA based on the search of tRNA genes of maize and soybean in a genomic tRNA

database (http://gtrnadb.ucsc.edu/). Considering the counts of iMET and Asp PBS sequences, these tRNAs are used most frequently as the reverse transcription initiation primers of LTR retrotransposons in the strawberry genome. Therefore, we used the 3′ terminal sequences of the iMET and Asp tRNAs as PCR primers to screen the major LTR sequences in the strawberry genome (**Table S3**).

Illumina NGS library construction

We constructed a sequencing library by PCR amplification of mechanically fragmented DNAs ligated with a forked adaptor at both ends using two PBS primers corresponding to iMET or Asp tRNAs and an adaptor primer for eight cultivars (Fig. 1, **Table S1 and Table S3**). The iMET and Asp PBS primers were mixed at a 7:1 ratio according to the LTR_STRUC results of the strawberry genome survey. PCR products for each cultivar were labeled with a unique 7–8 nt barcode sequence that was attached to the 5′ end of PBS primers (Fig. 1 and **Table S3**). We eluted DNA fragments of 300–500 bp, pooled an equal amount of the eluted DNA from each cultivar, and sequenced the fragments on a HiSeq2000 platform (Fig. 1). A total of 134,676,404 read pairs were obtained, 91.4% of which (123,106,589 read pairs) contained the expected PBS primer sequence at either end (102,967,889 and 20,138,700 read pairs of iMET PBS and Asp PBS, respectively) (Table 2), which indicated that the nested PCR amplified DNA fragments at the PBS sites of the strawberry genome specifically. After filtering those reads based on the primer sequence of the cultivar barcode and PBS, with the Q scores of all base calls being ≥ 30, 63.5% (total of 85,486,892 read pairs) of the reads remained (Table 2). The number of paired reads for iMET and Asp PBSs, and those that were further assigned to each cultivar, are shown in Tables 2 and 3, respectively.

Identification and comparison of LTR retrotransposon insertion sites

To identify insertion sites of the LTR sequence, the non-PBS read, i.e., the read opposite to the one through which the barcode and PBS sequences were filtered, were mapped to the *F. vesca* reference genome using the BWA software (Li and Durbin 2009). The mapping ratio of non-PBS reads ranged from 25.2% to 41.9% for Asp and iMET PBSs of the cultivated varieties; in contrast, the ratios were 76.2% and 84.1% for Asp and iMET PBSs of the wild species (*F. vesca*), respectively (Table 3). As expected, the mapping ratio of cultivated varieties was much lower than that of *F. vesca*. This is probably because cultivated species are allo-octoploid species that are derived from four different diploid ancestors, whereas the wild species *F. vesca* is diploid.

Although the non-PBS reads that mapped to a unique region should represent an insertion site sequence, those mapped to more than one site of the reference genome might be a part of an LTR sequence. The length of the LTR sequences varied according to family, and some of them were longer than the DNA fragments that were prepared for sequencing (300–500 bp). Thus, we extracted the regions in which non-PBS reads hit the reference genome uniquely. The ratio of unique hits to the reference genome over a total read in each sample ranged from 20.7% to 61.1% (Table 3). For each uniquely mapped region, we determined the “presence” or “absence” of insertions for each cultivar using DEGseq (R package in Bioconductor) (Wang et al. 2010) (Fig. 2). As a result, a total of 18,498 and 14,831 insertion sites were identified using the iMET and Asp PBSs, respectively (Fig. 2). Mapped insertion sites were visualized by the Integrative Genome Viewer (IGV) (Thorvaldsdóttir et al. 2013).

Pooling and clustering of LTR sequences corresponding to unique insertion sites

The "unique" insertion sites, which were detected in only one cultivar, were extracted from all identified insertion sites by DEGseq using Fisher's exact test with a significance level of 0.001 (Fig. 2). Some of these insertions must have occurred after cultivar divergence and were considered to be an evolutionarily recent event. The LTR family, which showed these unique insertions, is likely to have transpositional activity at the present time, or to have transposed in a recent past. We identified a total of 656 (representing 1,511,155 reads) and 114 (representing 440,977 reads) unique insertion sites using iMET and Asp PBSs, respectively (Fig. 2 and Table 4). We extracted the PBS reads corresponding to the unique insertion sites and pooled them. After trimming the cultivar barcode and PBS sequences from the reads, the LTR sequences adjacent to the PBS were obtained. We clustered the trimmed PBS reads based on their sequence similarities. The clustering analysis resulted in 18 and six clusters for iMET and Asp PBSs, respectively, which were composed of at least two putative unique insertion sites (Table 4). We aligned those reads in each cluster, some of which include the "TG" dinucleotide sequence at the 0–5 nt position in their alignment, which should correspond to the conserved end sequence of "CA" at the 3′ end of 5′ LTR adjacent to the PBS sequence (Fig. S1).

Experimental verification of insertion polymorphisms among cultivars

To investigate whether the LTR families identified have insertion polymorphisms among different cultivated varieties, we performed **S-SAP** analysis. This method visualizes insertion site patterns by amplifying specifically PCR fragments extending from the inserted **retrotransposon** copy to the nearest restriction endonuclease cutting site (**Waugh et al. 1997; Syed et al. 2005**). **We used mainly Japanese strawberry cultivars which are known to be genetically closely related.** As shown in **Fig. 3A ,**

299 **3B** and Fig. S2, a generally high degree of insertion polymorphism was detected
300 among cultivars. Some peaks were common to most cultivars, but some of them were
301 polymorphic (**Fig. 3A, 3B** and Fig. S2). Moreover, we detected unique peaks for many
302 of the cultivars tested, some of which were absent for both parents but present in
303 crossed offspring cultivars, indicating the occurrence of transposition during the
304 breeding process of cultivar development (**Fig. 3A, 3B** and Fig. S2). Considering their
305 close genetic relationships, the LTR retrotransposon families identified in this study
306 may have transpositional activity at the present time (or at least in a recent last).

307 ## Discussion

308 In this study, we screened for LTR retrotransposon families that show high
309 insertion polymorphisms among strawberry cultivars efficiently. This technique
310 comprises four major steps: (1) screening of all possible LTR sequences that share the
311 PBS sequences that are used predominantly in several cultivars using Illumina NGS
312 sequencing; (2) identification and comparison of LTR insertion sites among cultivars;
313 (3) clustering of LTR sequences with unique insertions among cultivars based on their
314 similarities; (4) experimental confirmation of the insertion polymorphisms of the LTR
315 sequences identified. We have shown that the LTR families identified using the
316 experimental steps described above had high insertion polymorphisms among closely
317 related cultivars (Fig. 3A, 3B and Fig. S2). Although various retrotransposon-based
318 marker systems have been developed (IRAP, RERAP, and SSAP), the successful
319 application of these systems depends largely on the availability of the LTR sequences.
320 Thus, our technique should provide LTR sequence information that is useful for the
321 groundbreaking development of these molecular markers, which will be used in

genetic analyses, the construction of linkage maps, and cultivar fingerprinting (Poczai et al. 2013).

Kalendar et al. (2010) reported that the iPBS method, in which sets of PBS primers are used in PCR amplification to visualize retrotransposon insertion polymorphisms directly, was a powerful DNA-fingerprinting technology. This method has some great advantages for screening LTR retrotransposons over previous methods, which included the PCR amplification of conserved protein-coding domains using degenerate sequence primers, particularly reverse transcriptase and integrase, followed by DNA walking toward the LTR. Although the experimental procedure is tedious, the insertion polymorphisms among cultivars need to be studied in the cloned LTR. The iPBS method using only PBS sequence information led to the identification of not only full-length LTR elements, such as *Gypsy* and *Copia*, but also the nonautonomous TRIM and LARD elements, which contain no protein-coding domains. Other advantages of this method are that the universal PBS motif is adjacent to the LTR sequence, thus facilitating the cloning of LTR sequences. However, in the iPBS method, only insertions that are sufficiently close to one another in a head-to-head orientation can produce DNA fragments (Kalendar et al. 2010; Poczai et al. 2013). This implies that, among the LTR families that share the same PBS sequence, those that have a large copy number or are organized in clusters are more likely to be cloned. In the large LTR families, most copies are genetically silenced. Even if the families contain a few active members, the new insertion sites formed by these members are covered almost completely by preexisting copies and it is quite hard to recognize them.

In contrast, our method amplifies random DNA fragments located between the PBS and the mechanically broken-down point. Thus, in principle, this method should allow the screening of all possible LTR retrotransposons that share the PBS sequences that

347 are used predominantly in the species that include the nonautonomous TRIM and
348 LARD elements, regardless of the distance and orientation of their insertion sites. We
349 identified a total of 20,918 LTR retrotransposon family candidates by clustering all
350 LTR sequences screened in this study (Monden et al. unpublished data). This number
351 may represent the majority of LTR retrotransposon families present in the strawberry
352 genome, because we selected the two most frequently used tRNA sequences to
353 construct the sequencing library. However, at this time, we selected the 300–500 bp
354 PCR products that were amplified using a combination of PBS and adaptor primers for
355 HiSeq2000 sequencing, which implies that any families with an LTR sequence longer
356 than about 500 bp could not be screened, because only products within the LTR
357 sequence were obtained. LTR sequences vary from 100 to 5,000 bp; thus, this size
358 selection might limit the LTR sequence that could be identified here. However, in
359 recent years, a sequencing platform that can provide significantly longer reads, such as
360 those over 10 kb, has become available, which should enable the screening of these
361 longer LTR sequences and their insertion sites. To date, there have been no reports
362 describing the total number of LTR retrotransposon families in the strawberry genome.
363 Moreover, similar to their method, we identified a TRIM element, which was
364 represented in iMET_Cl3 based on its characteristics, such as short length, high copy
365 number, absence of coding domains, and presence of PBS and PPT sites (Antonius-
366 Klemola et al. 2006) (Fig. S4).
367 It is known that Japanese strawberry cultivars are closely related genetically,
368 because most of them were developed from a limited number of ancestral cultivars,
369 such as “Haward17”, “Fukuba”, and “Donner” (Isobe et al. 2013). Thus, it has been
370 challenging to identify genetic polymorphisms efficiently among those varieties (Isobe
371 et al. 2013). In recent years, several linkage maps have been constructed in strawberry

using SSR, SCAR, and AFLP markers (Rousseau-Gueutin et al. 2008; Sargent et al. 2009; Zorrilla-Fontanesi et al. 2011; Sargent et al. 2012); in particular, the valuable resources of the whole-genome sequence information of *F.vesca* wild species and EST databases have accelerated the generation of higher-density linkage maps (Sargent et al. 2012; Isobe et al. 2013). The highest-density linkage map was constructed quite recently, which contains 1,856 SSR loci on 28 linkage groups (Isobe et al. 2013). **In addition, the whole genome sequence information of octoploid *Fragaria* species were finally released (ftp://ftp.kazusa.or.jp/pub/strawberry/genome/) (Hirakawa et al. 2014). We aligned one of the identified LTR sequence, iMET_Cl3, to the reference genome (*F. vesca* and *F. ananassa*), to identify insertion loci within the genome. We searched all alignments using bowtie2 (-a mode). The iMET_Cl3 LTR sequence was mapped at 147 (*F. vesca*) and 582 (*F. ananassa*) loci, respectively. Thus, the cultivated (allo-octoploid) species have many more copies of this LTR sequence than do *F. vesca* wild species (diploid), because allo-polyploidization may trigger TE activation (Petit et al. 2010). Furthermore, the results of the S-SAP analysis showed that cultivated species had more peaks than did wild species, which supports the contention that those species tend to have a higher copy number of this element than do wild species (data not shown). In addition, their insertion sites mapped to *F. vesca* genome were dispersed throughout the genome, particularly in the genic regions (exons, introns, and untranslated regions) (26.5% of the total mapped loci) and within the 1 kb flanking regions of genes (14.3% in the 5´ and 9.5 % in the 3´ flanking regions, respectively) (Fig. S3). These results suggest that the LTR family identified has characteristics that are suitable for use as molecular markers (high copy number and preferential insertion into genic regions). Combining several types of**

molecular markers such as AFLP, SSR, SCAR and retrotransposon-based, the development of linkage maps and map-based cloning of the genes should be accelerated even in Japanese strawberry cultivars.

Finally, we demonstrated that our method based on the use of the PBS of the LTR families can identify systematically LTR retrotransposon members showing diverse insertion patterns in the strawberry genome. As shown above, the insertion sites of these members were highly polymorphic among cultivars, which leads us to assume that the comprehensive sequencing of those retrotransposon junctions should provide useful information for genetic analyses. **In our recent work on sweet potato cultivars, we sequenced a total of over 76,912 data points (2,024 insertion sites across 38 cultivars) of one LTR retrotransposon family that showed high insertion polymorphism among cultivars in just one run of an Illumina MiSeq sequencing (Monden et al., in press).** Our results demonstrated that the method described here was useful for the efficient identification of the LTR retrotransposon families that show high insertion polymorphisms and can be applied not only to crop species, but also to animal and fungal species, as long as whole-genome sequence data is available. This technique should contribute to the development of molecular markers and to the construction of linkage maps using the insertion polymorphisms of these LTR retrotransposon families.

Acknowledgements

This work was supported by the Research and Development Projects for Application in Promoting New Policy of Agriculture Forestry and Fisheries grant from the Ministry of Agriculture, Forestry and Fisheries of Japan and the Program to

420 Disseminate Tenure Tracking System, from the Ministry of Education, Culture, Sports,
421 Science and Technology (MEXT), Japan (to Y.M.).

References

Antonius-Klemola, K., Kalendar, R., and Schulman, A.H. 2006. TRIM retrotransposons occur in apple and are polymorphic between varieties but not sports. Theor. Appl. Genet. **112**(6): 999–1008. doi:10.1007/s00122-005-0203-0. PMID:16404583.

Benjamini, Y., and Hochberg, Y. 2013. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. **57**(1): 289–300. Available from http://www.jstor.org/stable/2346101.

Ewing, A.D., and Kazazian, H.H. 2011. Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. Genome Res. **21**(6): 985–990. doi:10.1101/gr.114777.110. PMID: 20980553.

Flavell, A.J., Knox, M.R., Pearce, S.R., and Ellis, T.H. 1998. Retrotransposon-based insertion polymorphisms (RBIP) for high throughput marker analysis. Plant J. **16**(5): 643–650. PMID: 10036780.

Feschotte, C., Jiang, N., and Wessler, S.R. 2002. Plant transposable elements: where genetics meets genomics. Nat. Rev. Genet. **3**(5): 329–341. doi:10.1038/nrg793. PMID: 11988759.

Fragaria vesca Genome v1.1 Assembly. Available from http://www.rosaceae.org/species/fragaria/fragaria_vesca/genome_v1.1 [accessed 26 December 2010].

Havecker, E.R., Gao, X., and Voytas, D.F. 2004. The diversity of LTR retrotransposons. Genome Biol. **5**(6): 225. doi:10.1186/gb-2004-5-6-225. PMID: 15186483.

Hirakawa, H., Shirasawa, K., Kosugi, S., Tashiro, K., Nakayama, S., Yamada, M., et al. 2014. Dissection of the octoploid strawberry genome by deep sequencing

of the genomes of *Fragaria* species. DNA Res. 21(2): 169–181. doi:10.1093/dnares/dst049. PMID: 24282021.

Hizi, A. 2008. The reverse transcriptase of the Tf1 retrotransposon has a specific novel activity for generating the RNA self-primer that is functional in cDNA synthesis. J. Virol. **82**(21): 10906–10910. doi:10.1128/JVI.01370-08. PMID:18753200.

Iskow, R.C., McCabe, M.T., Mills, R.E., Torene, S., Pittard, W.S., Neuwald, A.F., et al. 2010. Natural mutagenesis of human genomes by endogenous retrotransposons. Cell, **141**(7): 1253–1261. doi:10.1016/j.cell.2010.05.020. PMID:20603005.

Isobe, S.N., Hirakawa, H., Sato, S., Maeda, F., Ishikawa, M., Mori, T., et al. 2013. Construction of an integrated high density simple sequence repeat linkage map in cultivated strawberry (*Fragaria* × *ananassa*) and its applicability. DNA Res. **20**(1): 79–92. doi:10.1093/dnares/dss035. PMID:23248204.

Levin, H.L., and Moran, J.V. 2011. Dynamic interactions between transposable elements and their hosts. Nat. Rev. Genet. **12**(9): 615–627. doi:10.1038/nrg3030.

Kalendar, R. 2011. The use of retrotransposon-based molecular markers to analyze genetic diversity. Ratar. Povrt. **48**(2): 261–274. doi:10.5937/ratpov1102261K.

Kalendar, R., Grob, T., Regina, M., Suoniemi, A., and Schulman, A. 1999. IRAP and REMAP : two new retrotransposon-based DNA fingerprinting techniques. Theor. Appl. Genet. **98**: 704–711. doi:10.1007/s001220051124.

Kalendar, R., Antonius, K., Smýkal, P., and Schulman, A.H. 2010. iPBS: a universal method for DNA fingerprinting and retrotransposon isolation. Theor. Appl. Genet. **121**(8): 1419–1430. doi:10.1007/s00122-010-1398-2. PMID:20623102.

Kalendar, R., Flavell, A.J., Ellis, T.H.N., Sjakste, T., Moisy, C., et al. 2011. Analysis of plant diversity with retrotransposon-based molecular markers. Heredity, **106**(4): 520–530. doi:10.1038/hdy.2010.93. PMID: 20683483.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. **30**(14): 3059–3066. PMID:12136088.

Kelly, N.J., Palmer, M.T., and Morrow, C.D. 2003. Selection of retroviral reverse transcription primer is coordinated with tRNA biogenesis. J. Virol. **77**(16): 8695–8701. doi:10.1128/JVI.77.16.8695.

Kofler, R., Betancourt, A.J., and Schlötterer, C. 2012. Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in Drosophila melanogaster. PLoS Genet. **8**(1): e1002487. doi:10.1371/journal.pgen.1002487. PMID:22291611.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. 2009. Circos: an information aesthetic for comparative genomics. Genome Res. **19**(9): 1639–1645. doi:10.1101/gr.092759.109. PMID: 19541911.

Kumar, A., and Bennetzen, J.L. 1999. Plant retrotransposons. Annu. Rev. Genet. **33**: 479–532. doi:10.1146/annurev.genet.33.1.479. PMID: 10690416.

Kumar, A., and Hirochika, H. 2001. Applications of retrotransposons as genetic tools in plant biology. Trends Plant Sci. **6**(3): 127–134. PMID: 11239612.

Li, H., and Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, **25**(14): 1754–1760. doi:10.1093/bioinformatics/btp324. PMID:19451168.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics, **25**(16): 2078–2079. doi:10.1093/bioinformatics/btp352. PMID:19505943.

Lisch, D. 2009. Epigenetic Regulation of Transposable Elements in Plants. Annu. Rev. Plant Biol. **60**:43–66. doi:10.1146/annurev.arplant.59.032607.092744.

497 Macas, J., Neumann, P., and Navrátilová, A. 2007. Repetitive DNA in the pea (*Pisum*
498 *sativum* L.) genome: comprehensive characterization using 454 sequencing and
499 comparison to soybean and Medicago truncatula. BMC Genomics, **8**: 427.
500 doi:10.1186/1471-2164-8-427. PMID:18031571.
501 Mak, J., and Kleiman, L. 1997. Primer tRNAs for reverse transcription. J. Virol.
502 **71**(11): 8087–8095. PMID: 9343157.
503 Marquet, R., Isel, C., Ehresmann, C., and Ehresmann, B. 1995. tRNAs as primer of
504 reverse transcriptases. Biochimie, **77**: 113–124. PMID:7541250.
505 Mayer, K.F.X., Waugh, R., Brown, J.W.S., Schulman, A., Langridge, P., Platzer, M.,
506 et al. 2012. A physical, genetic and functional sequence assembly of the barley
507 genome. Nature, **491**(7426): 711–716. doi:10.1038/nature11543. PMID: 23075845.
508 McCarthy, E.M., and McDonald, J.F. 2003. LTR_STRUC: a novel search and
509 identification program for LTR retrotransposons. Bioinformatics, **19**(3): 362–367.
510 doi:10.1093/bioinformatics/btf878.
511 **Monden, Y., Yamamoto, A., Shindo, A., and Tahara, M. (In press). Efficient DNA**
512 **fingerprinting based on the targeted sequencing of active retrotransposon**
513 **insertion sites using a bench-top high-throughput sequencing platform. DNA**
514 **Res.**
515 Novák, P., Neumann, P., and Macas, J. 2010. Graph-based clustering and
516 characterization of repetitive sequences in next-generation sequencing data. BMC
517 Bioinformatics, **11**: 378. doi:10.1186/1471-2105-11-378. PMID:20633259.
518 Pagán, H.J.T., Macas, J., Novák, P., McCulloch, E.S., Stevens, R.D., and Ray, D.A.
519 2012. Survey sequencing reveals elevated DNA transposon activity, novel
520 elements, and variation in repetitive landscapes among vesper bats. Genome Biol.
521 Evol. **4**(4): 575–585. doi:10.1093/gbe/evs038. PMID: 22491057.

Petit, M., Guidat, C., Daniel, J., Denis, E., Montoriol, E., Bui, QT., et al. 2010. Mobilization of retrotransposons in synthetic allotetraploid tobacco. New Phytol. **186**(1): 135–147. doi:10.1111/j.1469-8137.2009.03140.x. PMID: 20074093.

Poczai, P., Varga, I., Laos, M., Cseh, A., Bell, N., Valkonen, J.P., et al. 2013. Advances in plant gene-targeted and functional markers: a review. Plant Methods, 2013, **9**(1): 6. doi:10.1186/1746-4811-9-6. PMID:23406322.

Quinlan, A.R., and Hall, I.M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, **26**(6): 841–842. doi:10.1093/bioinformatics/btq033. PMID:20110278.

Rousseau-Gueutin, M., Lerceteau-Köhler, E., Barrot, L., Sargent, D.J., Monfort, A., Simpson, D., et al. 2008. Comparative genetic mapping between octoploid and diploid Fragaria species reveals a high level of colinearity between their genomes and the essentially disomic behavior of the cultivated octoploid strawberry. Genetics, **179**(4): 2045–2060. doi:10.1534/genetics.107.083840. PMID:18660542.

Sargent, D.J., Fernandéz-Fernandéz, F., Ruiz-Roja, J.J., Sutherland, B.G., Passey, A., Whitehouse, A.B., and Simpson, D.W. 2009. A genetic linkage map of the cultivated strawberry (*Fragaria* × *ananassa*) and its comparison to the diploid *Fragaria* reference map. Mol. Breed. **24**(3): 293–303. doi:10.1007/s11032-009-9292-9.

Sargent, D.J., Passey, T., Surbanovski, N., Lopez Girona, E., Kuchta, P., Davik, J., et al. 2012. A microsatellite linkage map for the cultivated strawberry (*Fragaria* × *ananassa*) suggests extensive regions of homozygosity in the genome that may have resulted from breeding and selection. Theor. Appl. Genet. **124**(7): 1229–1240. doi:10.1007/s00122-011-1782-6. PMID:22218676.

546 Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., et al. 2009.
547 The B73 maize genome: complexity, diversity, and dynamics. Science, **326**(5956):
548 1112–1115. doi:10.1126/science.1178534. PMID:19965430.
549 Shimizu, K., and Tsuda, K. 2011. SlideSort: all pairs similarity search for short reads.
550 Bioinformatics, **27**(4): 464–470. doi:10.1093/bioinformatics/btq677. PMID:
551 21148542.
552 Slotkin, R.K., and Martienssen, R. 2007. Transposable elements and the epigenetic
553 regulation of the genome. Nat. Rev. Genet. **8**(4): 272–285. doi:10.1038/nrg2072.
554 PMID:17363976.
555 Syed, N.H., Sureshsundar, S., Wilkinson, M.J., Bhau, B.S., Cavalcanti, J.J.V., and
556 Flavell, A.J. 2005. Ty1-copia retrotransposon-based SSAP marker development in
557 cashew (*Anacardium occidentale* L.). Theor. Appl. Genet. **110**(7): 1195–1202.
558 doi:10.1007/s00122-005-1948-1. PMID:15761718.
559 **Strawberry Genome And Resource Database Entry (Strawberry GARDEN).**
560 **Available from http://strawberry-garden.kazusa.or.jp/.**
561 Storey, J.D. 2002. A direct approach to false discovery rates. J. R. Stat. Soc. **64**(3):
562 479–498. doi:10.1111/1467-9868.00346.
563 Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. 2013. Integrative Genomics
564 Viewer (IGV): high-performance genomics data visualization and exploration.
565 Brief Bioinform. **14**(2): 178–192. doi:10.1093/bib/bbs017. PMID:22517427.
566 Urbański, D.F., Małolepszy, A., Stougaard, J., and Andersen, S.U. 2012. Genome-wide
567 *LORE*1 retrotransposon mutagenesis and high-throughput insertion detection in
568 *Lotus japonicus*. Plant J. **69**(4): 731–741. doi:10.1111/j.1365-313X.2011.04827.x.
569 PMID: 22014280.

570 Wang, L., Feng, Z., Wang, X., Wang, X., and Zhang, X. 2010. DEGseq: an R package
571 for identifying differentially expressed genes from RNA-seq data. Bioinformatics,
572 **26**(1): 136–138. doi:10.1093/bioinformatics/btp612. PMID:19855105.
573 Waugh, R., McLean, K., Flavell, A.J., Pearce, S.R., Kumar, A., Thomas, B.B., et al.
574 1997. Genetic distribution of Bare-1-like retrotransposable elements in the barley
575 genome revealed by sequence-specific amplification polymorphisms (S-SAP).
576 Mol. Gen. Genet. **253**(6): 687–694. PMID:9079879.
577 Wessler, S.R. 2006. Transposable elements and the evolution of eukaryotic genomes.
578 Proc. Natl. Acad. Sci. U. S. A. **103**(47): 17600–17601.
579 doi:10.1073/pnas.0607612103. PMID:17101965.
580 Xing, J., Witherspoon, D.J., and Jorde, L.B. 2013. Mobile element biology: new
581 possibilities with high-throughput sequencing. Trends Genet. **29**(5): 280–289.
582 doi:10.1016/j.tig.2012.12.002. PMID:23312846.
583 Zorrilla-Fontanesi, Y., Cabeza, A., Domínguez, P., Medina, J.J., Valpuesta, V.,
584 Denoyes-Rothan, B., et al. 2011. Quantitative trait loci and underlying candidate
585 genes controlling agronomical and fruit quality traits in octoploid strawberry
586 (*Fragaria* × *ananassa*). Theor. Appl. Genet. **123**(5): 755–778.
587 doi:10.1007/s00122-011-1624-6. PMID:21667037.
588

589 **Tables**

590 **Table 1.** PBS sequences in the strawberry genome.

Putative PBS sequences	Number of appearances	Type of tRNA
TGGTATCAGAGC	91	iMET
TGGCGCCGTCTG	12	Asp
TGGTACCAGAGC	9	
TGGCACGCCCAG	6	
TGGCTCCCCCTT	6	
TGGTATCAAGAG	6	
TGGCGCTAGAAG	5	
TGGTATCAAAGC	5	
TGGCATCAGAGC	4	
TGGTATTAGAGC	4	
TGGCGCCGTTTG	3	
TGGTAATCAGAG	3	
TGGTATCAGAGT	3	
TGGTATCCAGAG	3	
TGGCACGCCTAG	2	
TGGTATCAGCCT	2	
TGGTATCTAGAG	2	
Others*	44	
Total	210	

*There were 44 different sequences, each appeared just once.

591

Table 2. Summary of paired-end sequence reads.

Classification	No. of read pairs	% of total
Total read pairs	134,676,404	100.0
PBS sequence identified	123,106,589	91.4
(a) iMET PBS	102,967,889	76.5
(b) Asp PBS	20,138,700	15.0
PBS and Barcode filtered*	85,485,943	63.5
(a) iMET PBS and barcode	72,200,783	53.6
(b) Asp PBS and barcode	13,285,160	9.9

* The PBS and cultivar barcode sequences with quality scores of all base calls ≥ 30

592

593

Table 3. Read mapping information.

PBS type	Variety	Total reads	Mapped read number	Hit Ratio (%)	Uniquely mapped read number	Uniquely hit ratio (%)
Asp	Hinoshizuku	1797668	453438	25.2	385525	21.4
	Amaou	1056503	269213	25.5	218278	20.7
	Fukuba	972319	281111	28.9	236830	24.4
	Kotoka	4678389	1611315	34.4	1395512	29.8
	Tochiotome	1663097	543240	32.7	466708	28.1
	Natsuotome	2092534	748777	35.8	612657	29.3
	Donner	982549	291057	29.6	241792	24.6
	Fragaria vesca	42101	32099	76.2	25719	61.1
iMET	Amaou	3009803	1004224	33.4	632714	21
	Hinoshizuku	10794257	3806400	35.3	2349967	21.8
	Fukuba	6097634	2350180	38.5	1370934	22.5
	Kotoka	24354637	9265125	38	5823274	23.9

Tochiotome	16476654	5892011	35.8	3689183	22.4
Natsuotome	9041762	3403783	37.6	2221733	24.6
Donner	1550995	649939	41.9	320803	20.7
Fragaria vesca	875041	735950	84.1	533211	60.9

Table 4. Extracting and clustering of LTR sequences corresponding to unique insertion sites.

Types of library	No. of unique insertions	No. of reads	No. of clusters*
Asp PBS	114	440,977	6
iMET PBS	656	1,511,155	18

* Extracted cluster contains ≥ 2 reliable unique insertions.

596

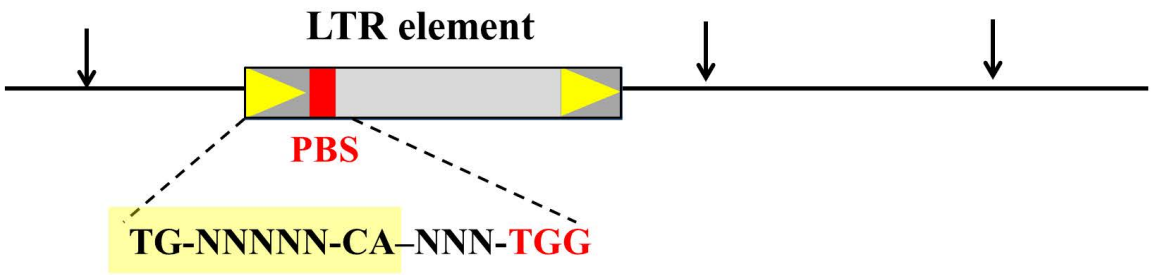
Figure captions

Fig. 1. Schematic representation of the preparation and sequencing of the Illumina NGS library. (A) Genomic DNAs are fragmented by g-TUBE. An LTR retrotransposon is described (gray rectangle, LTR element; yellow triangle, LTR region; red box, PBS site). The sequence of the LTR region starts at "TG" and ends at "CA". The PBS site starts at "TGG" (note in red), which is 0–5 bp away from the 3′ end of the LTR. The black arrow represents the cutting site. (B) Fragment ends are repaired, 3′ A overhangs are added, and forked adaptors are ligated onto the ends. The green boxes represent adaptors. (C) PCR amplification is performed using a PBS primer (red triangle) carrying a cultivar-specific barcode sequence (orange, purple, and blue boxes) and an adaptor primer (green triangle). PCR products with a length of 300–500 bp are selected by gel electrophoresis. (D) Multiple barcoded samples pooled for subsequent Illumina paired-end sequencing. One read of 101 bp (red arrow) contains the 7–8 bp barcode sequence (orange, purple, and blue lines), followed by the PBS sequence and the LTR sequence. The read of the other side (representing a non-PBS read of 101 bp (green arrow)) contains a genomic sequence.

Fig. 2. Graphical view of insertion sites among cultivars. The outermost concentric circle (A) shows the eight scaffolds of the *Fragaria vesca* genome (v1.1). The inner circle (B) indicates the gene density from a gene count per 10 kb sliding window. The remaining inner circles (C–J) display the distribution of non-PBS (iMET) mapped regions as eight colored histograms. The putative unique insertion sites were determined using the modified DEG seq via Fisher's exact test ($P < 0.001$) and are indicated by the yellow line.

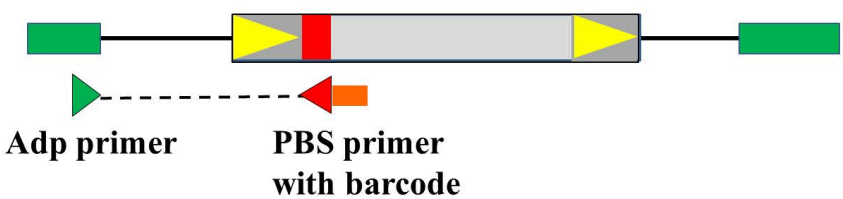
Fig. 3. S-SAP of the selected LTR sequence. S-SAP of an identified LTR sequence (iMET_Cl3) on 17 selected cultivars. (A) shows the peak pattern using *Mse*I and (B) shows that using *Rsa*I. X-axis: size of DNA fragments. Y-axis: height of peaks. The red arrow indicates the cultivar-specific peaks. 1, Pechika; 2, Natsuakari; 3, Summer Candy; 4, Summer Tiara; 5, Natsuotome; 6, Tochihitomi; 7, Aptos; 8, Celine; 9, Kitanokagayaki; 10, Moikko; 11, Otomegokoro; 12, Echigohime; 13, Yayoihime; 14, Nyoho; 15, Shinnyoho; 16, Skyberry; 17, Fukuba.

(A) Genomic DNA fragmentation

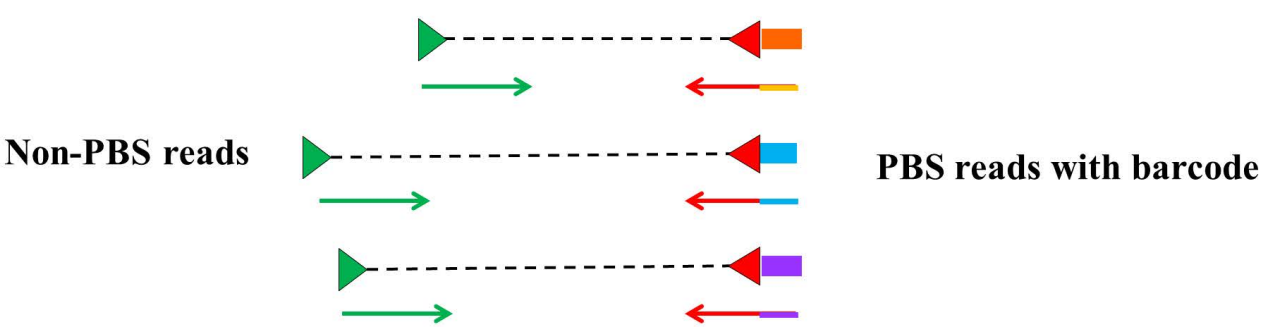


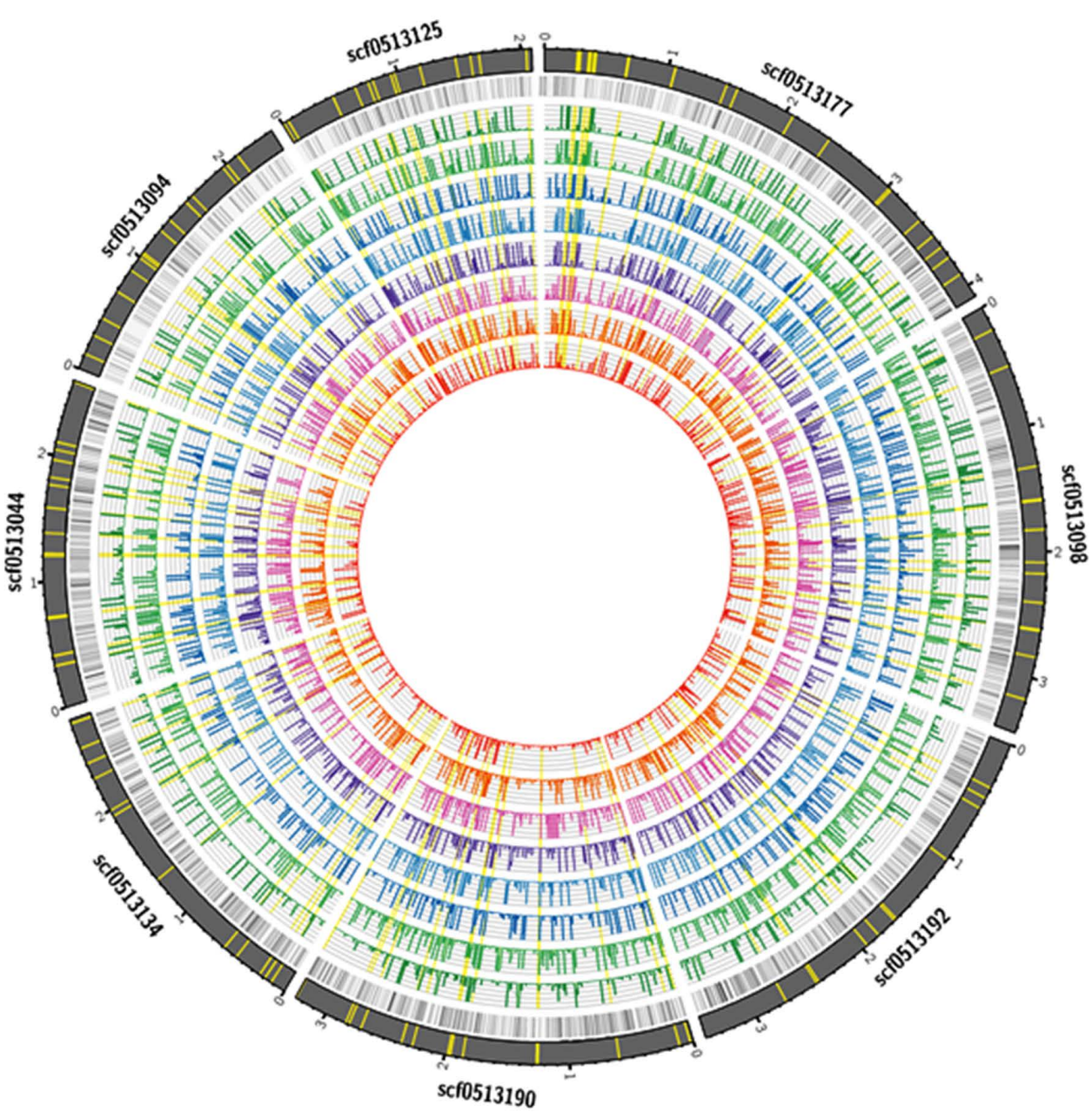
(B) End repair of DNA fragment and adaptor ligation

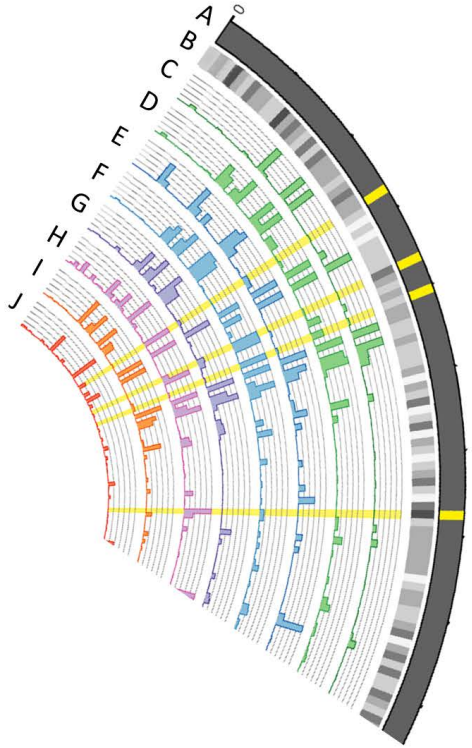
(C) PCR amplification using PBS and adaptor primers



(D) Pooling individually barcoded libraries and Paired-end sequencing







B; Gene density (genes/10kb)
C-J; Insertion sites (inserts/10kb)
C; Amaou
D; Hinoshizuku
E; Fukuba
F; Kotoka
G; Tochiotome
H; Natsuotome
I; Donner
J; *Fragaria vesca*

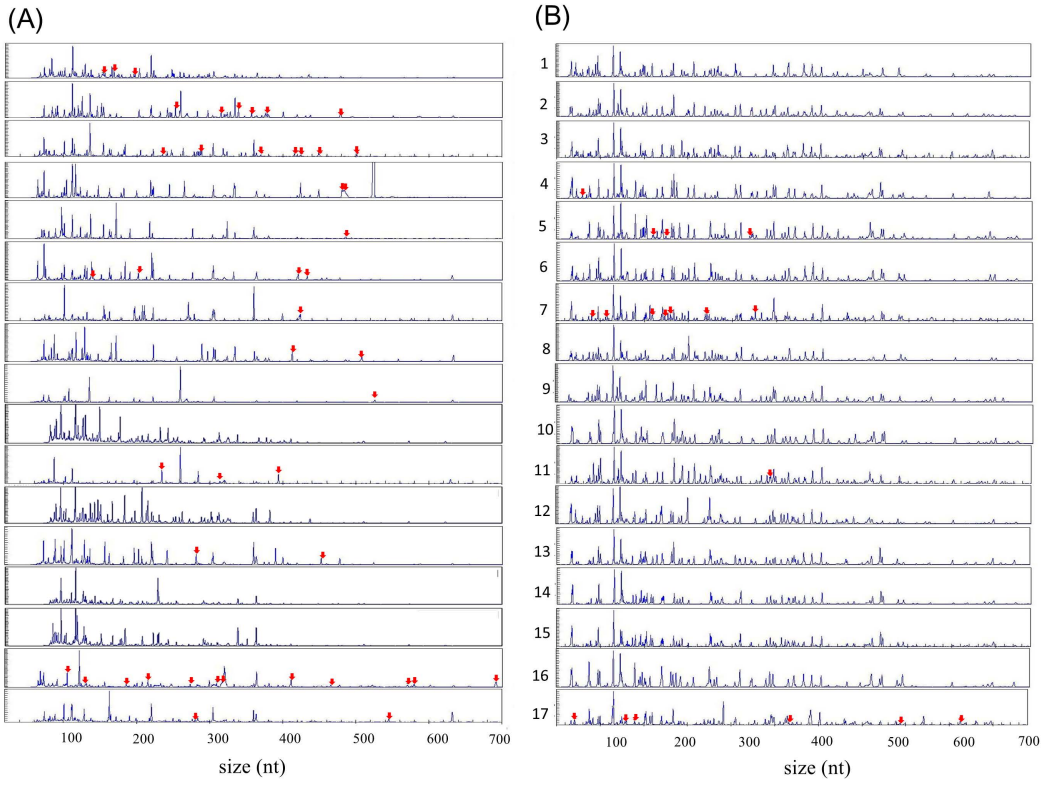

(A)
(B)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
100
200
300
400
500
600
700
size (nt)
size (nt)

Fig. S1. Alignment of trimmed PBS reads in the selected cluster (iMET_Cl4). The black arrows indicate a “TG” dinucleotide motif that is the reverse complementary sequence of the 3′ end of the LTR. The sequence shows the 1 bp spacer and the leading to the LTR sequence.

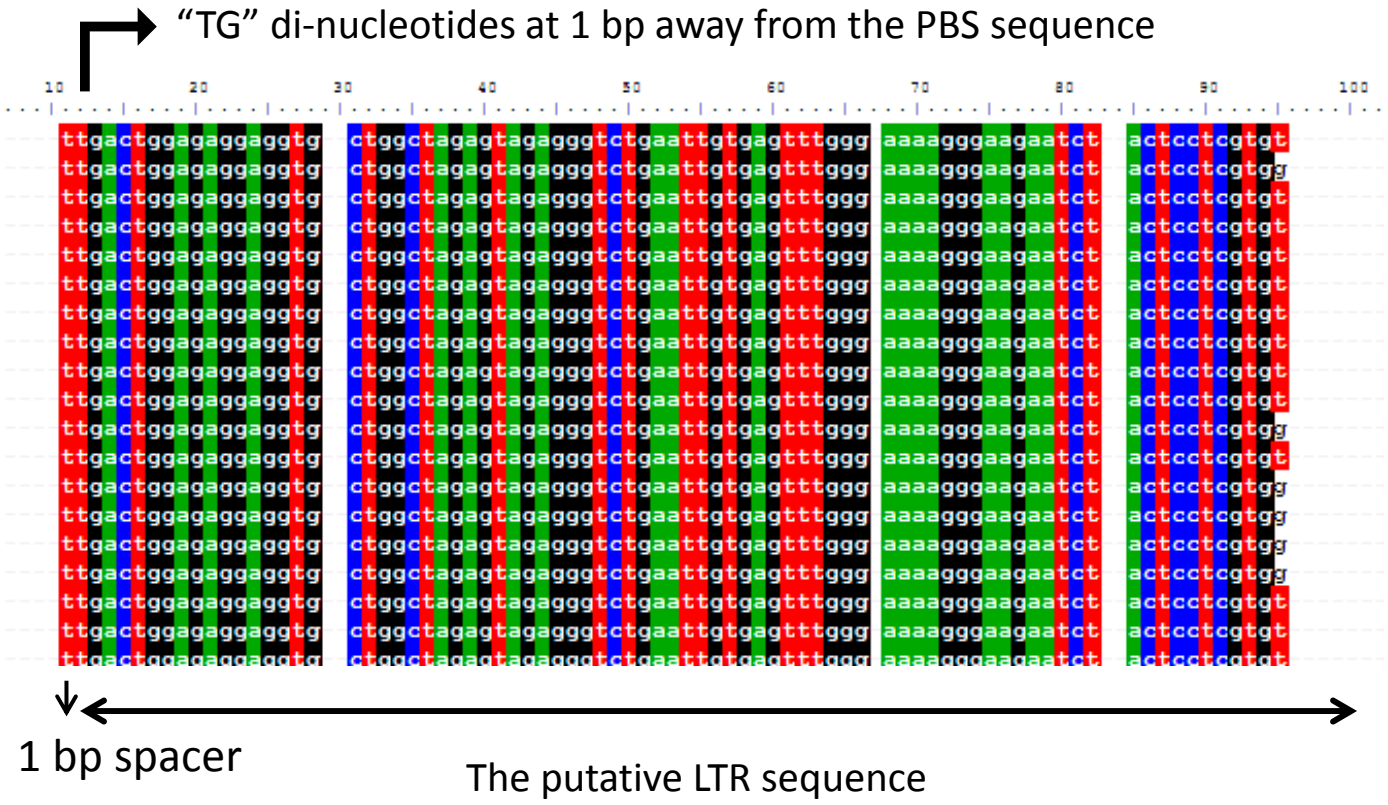
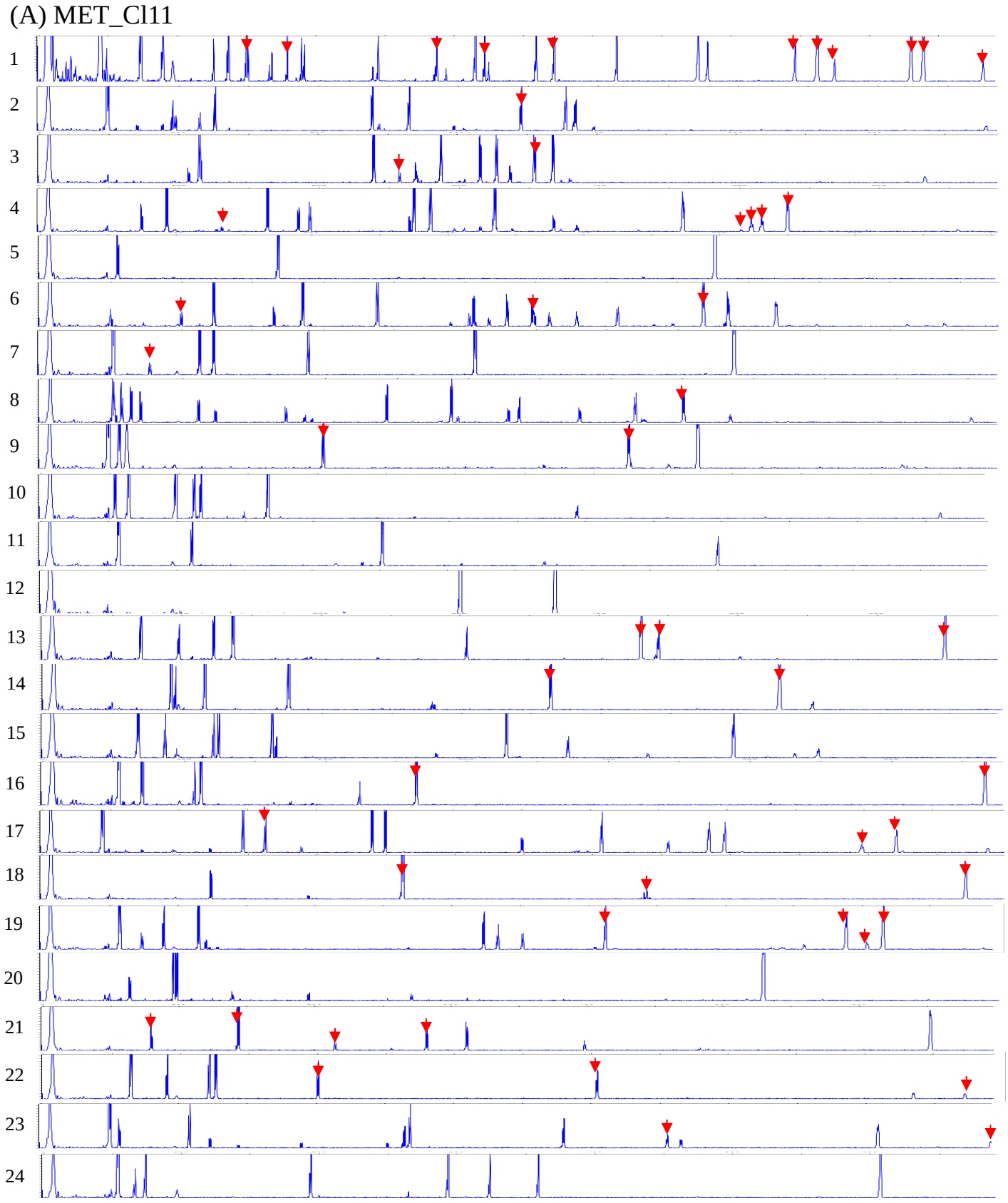
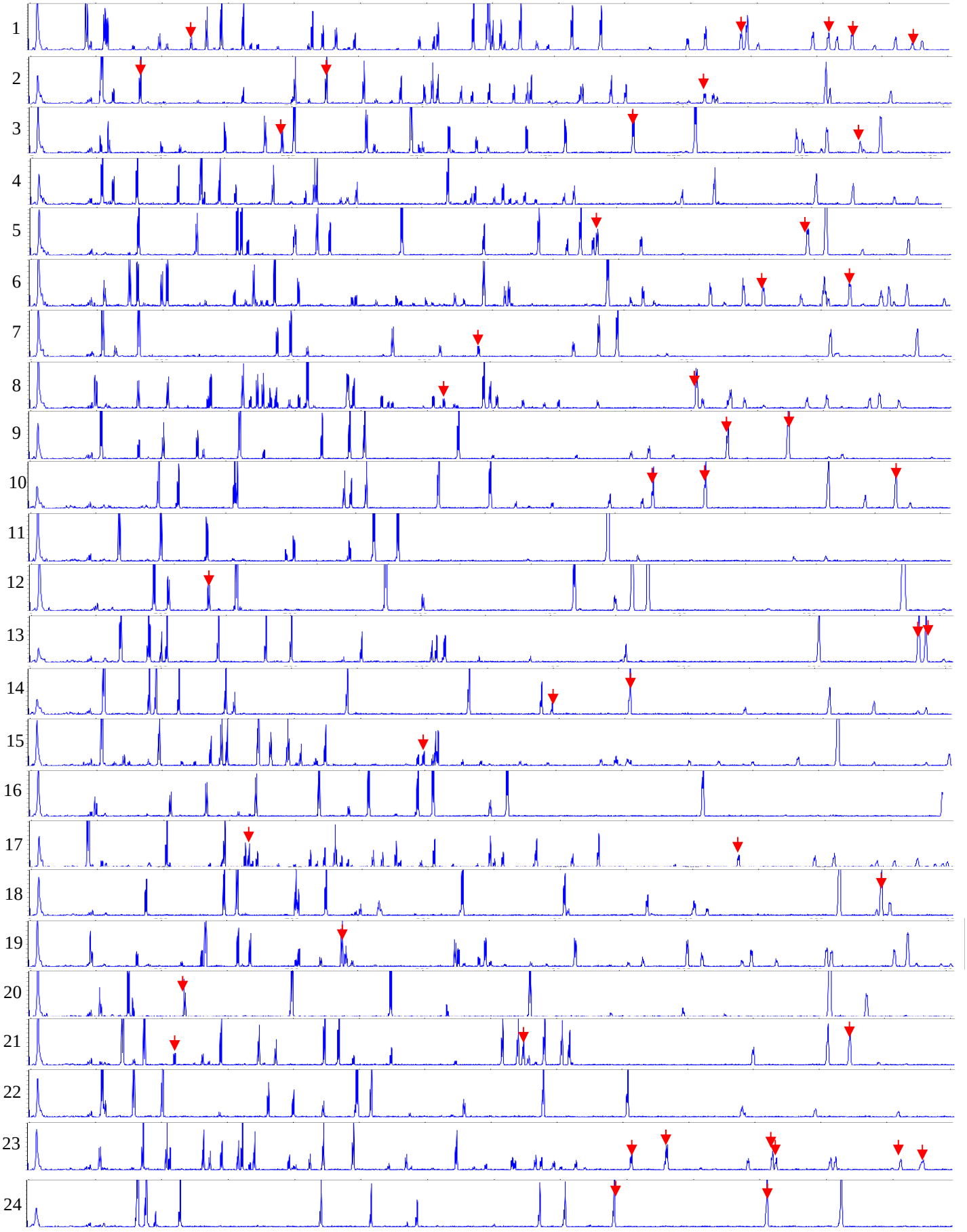


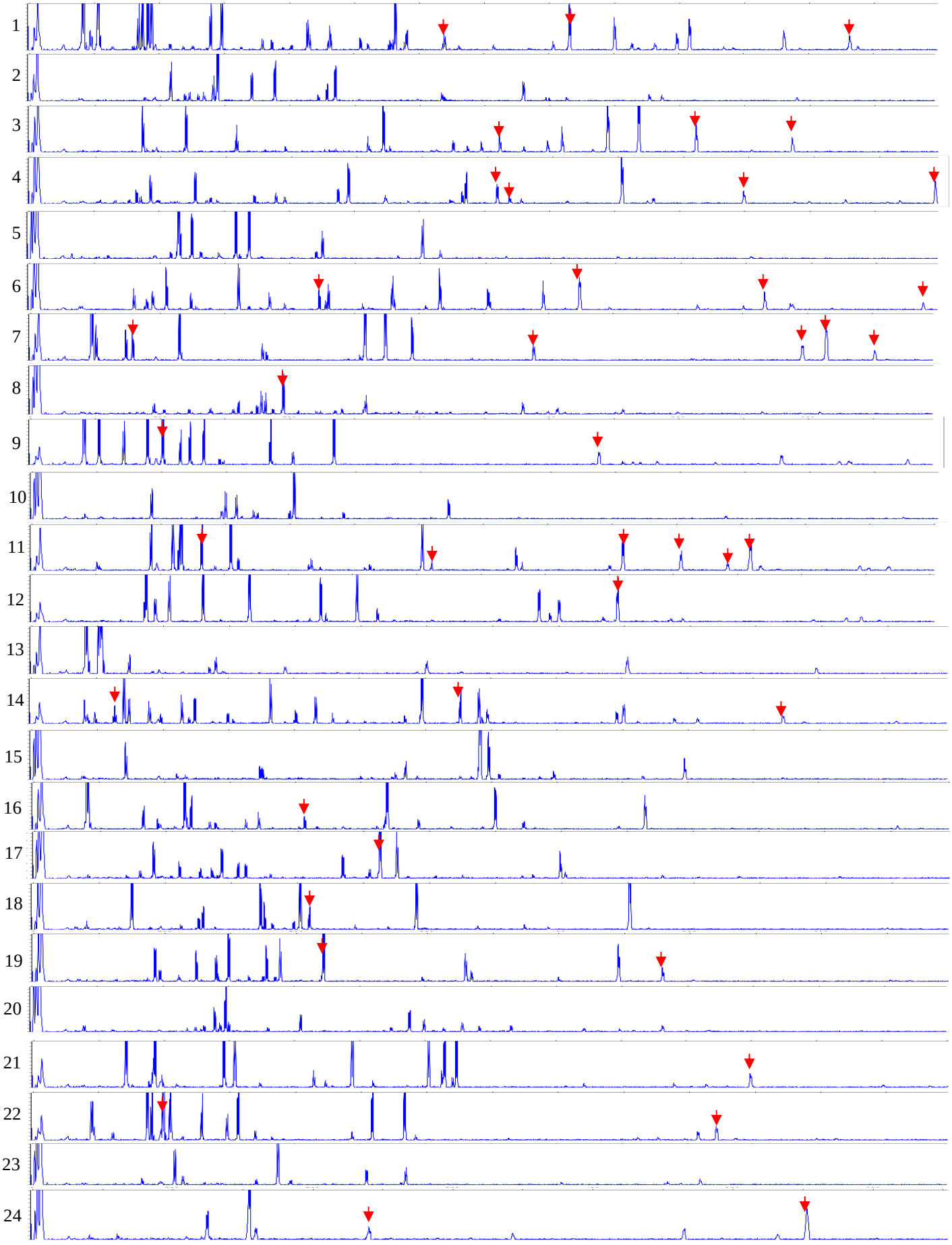
Fig. S2. Results of the transposon display analysis of the LTR sequences identified. The red arrows indicate the cultivar-specific peaks. (A) iMET_Cl11; (B) iMET_Cl20; (C) iMET_Cl28; (D) iMET_Cl76. 1, Benihoppe; 2, Yumenoka; 3, Kaorino; 4, Kotoka; 5, Marihime; 6, Amaotome; 7, Amaou; 8, Sagahonoka; 9, Hinoshizuku; 10, Miyazakinatsuharuka; 11, Satsumaotome; 12, Ohkimi; 13, Tochiotome; 14, Sachinoka; 15, Akihime; 16, Aiberry; 17, Fukuba; 18, Donner; 19, Hokowase; 20, 06A-184; 21, Elsanta; 22, Floridabelle; 23, Red Pearl; 24, Toyonoka.



(B) MET_Cl20



(C) MET_Cl28



(D) MET_Cl76

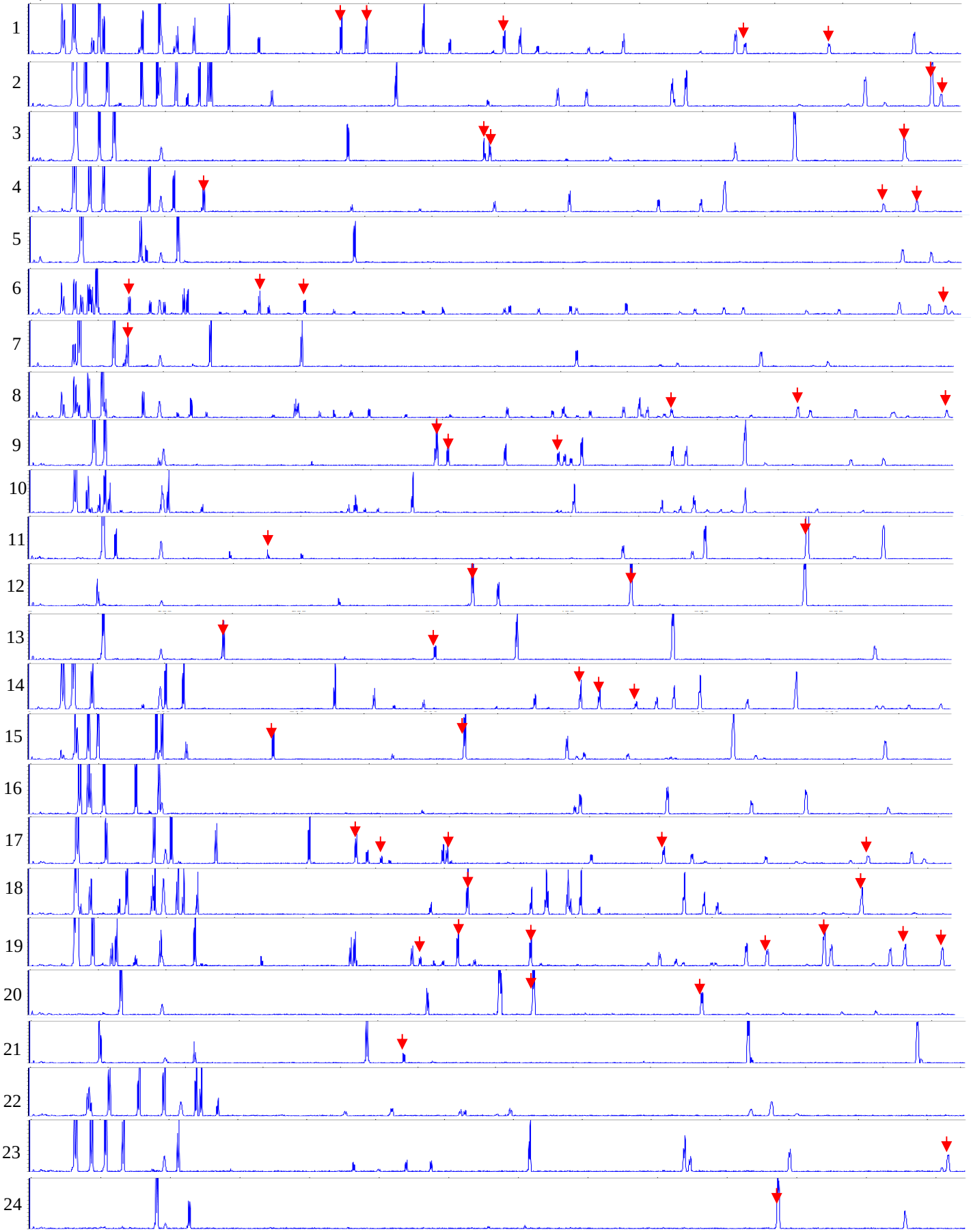


Fig. S3. Distribution of an LTR element (iMET_Cl3) identified in the *Fragaria vesca* v1.1 genome. The x-axis shows the number of iMET_Cl3 mapped sites. Among all the mapped sites, 26.5% were located within genic regions, such as the UTR, exons, and introns, whereas 14.3% and 9.5% were located within the 1 kb flanking region located at the 5′ and 3′ ends, respectively.

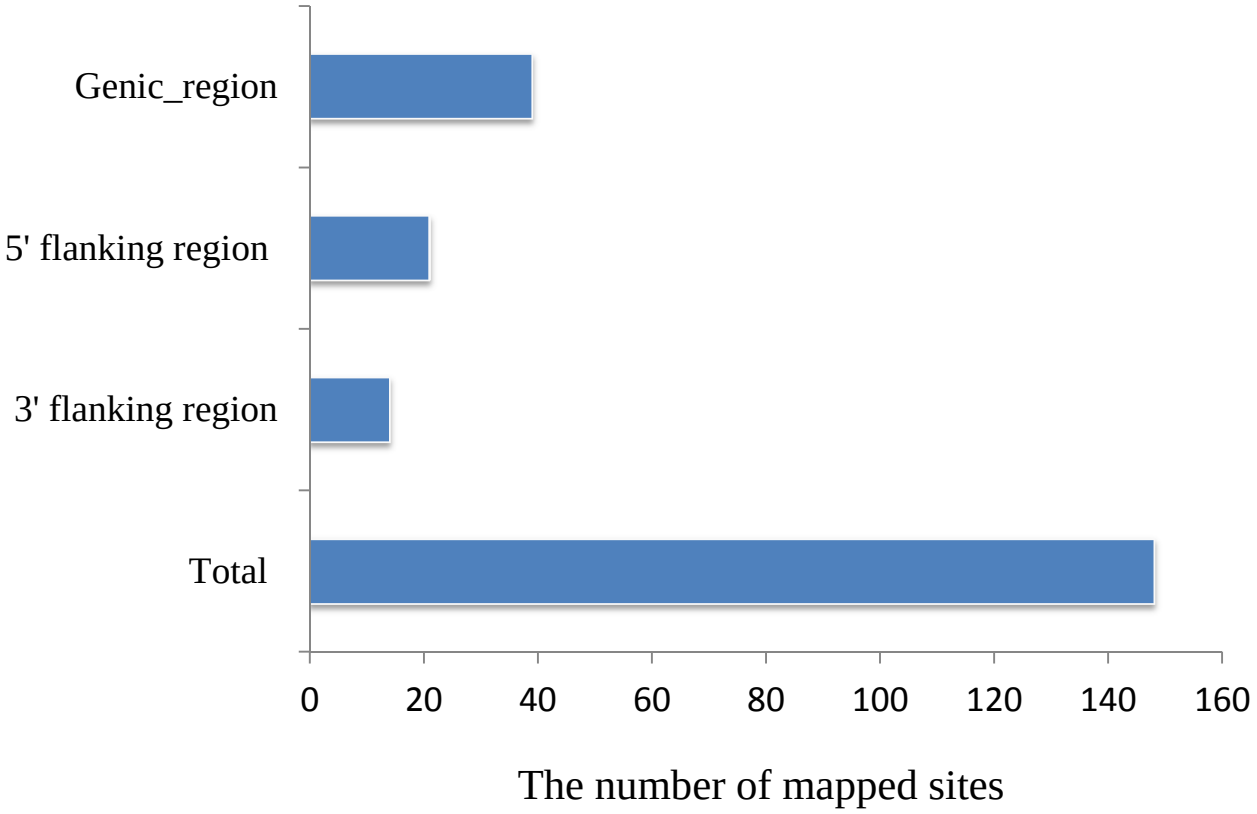
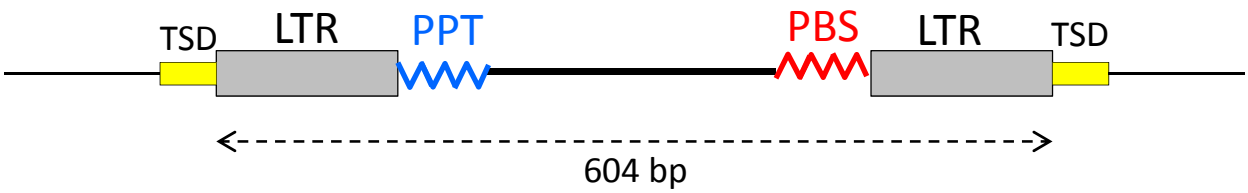


Fig. S4. Structure and sequence of a TRIM element (iMET_Cl3) identified in this study. The target site duplication (TSD) of 5 nt (represented by the yellow box) is located in the flanking region of the element. The grey box represents the LTR region. In the internal region, although the protein-coding region was not observed, the putative PPT (blue wavy line) and PBS (red wavy line) sequences were detected. The entire length was estimated at 604 bp.



ATCAAGTTTCTGTCAAATAGCGACAAGCGTGGCTCGTGGATCATCCTTGTACCGAT
ATTGTCCCAAATTAACCACCTTTTAAGGTGTTGGGTTTTAATCACAAAAGACCTCG
GTACAATTAGGAATGATCCACCCACTTATAAGTTATATTTTATTTGTCACTTTTCCAA
TGTGGGATCTCTCCTCTCCAACACTCCCCCTCACGTGCAACCTAACTTTTAGGTCT
GCACGTGAAACTGATTAACAAACTAAACCCGGGCCCCATGATAACCGAAAGAACC
CCAAACTCCCAGACTTCCAGTGACAACTGAAAGAACCAAACCAGGCTGAATTAA
CTCCAAACTCGAAAGAACCAAACCAGGGCCATGACCCCTGAAAGAACCAACCCG
GGCTTATGGAGACGCAACTGGAGAGGGACCCGCTCTGATACCATGTCAAACAGC
GACAAGCGTGGCTCGTGGATCATCCTTGTACCGATATTGTCCCAAATTAACCACCT
TTAAGGTGTTGGGTTTTAATCACAAAAGGCCTCGGTACAATTAGGAATGATCCAC
CCACTTATAAGTTATATTTTATTTGTCACTTTTCCAATGTGGGATCTCTCCTCTCCAA
CAGTTTCTACTT

Table S1. Summary of strawberry cultivars used for sequencing library constrcution.

Cultivar	Species	Counytry of Origin	Parentage	Accession code	Reference[a)]	Note
Hinoshizuku (Kumaken I 548)	*Fragaria ananassa* Duch.	Japan	98-30 (Sachinoka x Tochinomine) x 98-20-3 (Kurume 54gou x Tochinomine)	13882	1	http://www.pref.kumamoto.jp/uploaded/attachment/64129.pdf
Amaou (Fukuoka S6gou)	*Fragaria ananassa* Duch.	Japan	Kurume 53gou x 92-46 (Kurume 49gou x Sachinoka)	12572	1	http://farc.pref.fukuoka.jp/farc/kenpo/kenpo-22/22-13.htm
Fukuba	*Fragaria ananassa* Duch.	Japan	Seedling from General Chanzy	PI 231088	3	
Kotoka	*Fragaria ananassa* Duch.	Japan	7-3-1 x Benihoppe	21164	1	http://www.pref.nara.jp/secure/73720/41-1-10.pdf
Tochiotome	*Fragaria ananassa* Duch.	Japan	Kurume 49gou x Tochinomine	PI 617008	3	
Natsuotome	*Fragaria ananassa* Duch.	Japan	Tochigi 24gou x 00-25-1	20766	1	
Donner	*Fragaria ananassa* Duch.	USA	CAL 222 x CAL 145.52	ESP138-0303	2	
	Fragaria ananassa Duch.	USA	US-634 x Blakemore	FRA207-5041	2	
F. vesca	*Fravarie vesca* L.	Spain		ESP138-*[b)]	2	Wild species

[a)] 1: PVP (Plant Variety Plotectin) Office at MAFF, JAPAN (http://www.hinsyu.maff.go.jp/en/en_top.html), 2: GenBerry Database (http://www.bordeaux.inra.fr/eustrawberrydb/), 3: USDA ARS (Agricultural Research Service) (http://www.usda.gov/wps/portal/usda/usdahome)

[b)]*: 0030, 0005, 0326, 0013, 0006, 0007, 0016, 0010, 0011, 0017, 0188, 0192, 0189, 0325, 0190, 0191, 0196, 0012, 0015, 0651, 0655, 0656, 0596, 0597, 0599, 0600, 001

Table S2. Summary of strawberry cultivars used for S-SAP experiment.

Cultivar	Species	Counytry of Origin	Parentage
Pechika	*Fragaria ananassa* Duch.	Japan	Oishishikinari 2gou x Summer Berry
Natsuakari	*Fragaria ananassa* Duch.	Japan	Summer Berry x Kitanokagayaki
Summer Candy	*Fragaria ananassa* Duch.	Japan	Tochiotome x (Summer Berry × Morioka 26gou)
Summer Tiara	*Fragaria ananassa* Duch.	Japan	Selva x Benihoppe
Tochihitomi	*Fragaria ananassa* Duch.	Japan	Celine x Sachinoka
Aptos	*Fragaria ananassa* Duch.	USA	Tufts x CAL 65.63-601
Celine	*Fragaria ananassa* Duch.	Japan	Oishikinari x Kaho
Kitanokagayaki	*Fragaria ananassa* Duch.	Japan	Bell Rouge x Pajaro
Moikko	*Fragaria ananassa* Duch.	Japan	Sachinoka x ?
Otomegokoro	*Fragaria ananassa* Duch.	Japan	Sakyu S2gou x Kitanokagayaki
Echigohime	*Fragaria ananassa* Duch.	Japan	(Bell Rouge x Nyoho) x Toyonoka
Yayoihime	*Fragaria ananassa* Duch.	Japan	(Tone-hoppe x Tochiotome) x Tone-ho
Nyoho	*Fragaria ananassa* Duch.	Japan	Kei 210 x Reiko
Shinnyoho	*Fragaria ananassa* Duch.	Japan	Mutant line from Nyoho
Skyberry (Tochigi I 27gou)	*Fragaria ananassa* Duch.	Japan	00-24-1 x Tochigi 20 gou
Benihoppe	*Fragaria ananassa* Duch.	Japan	Akihime x Sachinoka
Yumenoka	*Fragaria ananassa* Duch.	Japan	Kurume 55gou x Kei 531
Kaorino	*Fragaria ananassa* Duch.	Japan	unknown
Marihime	*Fragaria ananassa* Duch.	Japan	Akihime x Sachinoka
Amaotome	*Fragaria ananassa* Duch.	Japan	Tochiotome x Sagahonoka
Sagahonoka	*Fragaria ananassa* Duch.	Japan	Onishiki x Toyonoka
Miyazakinatsuharuka	*Fragaria ananassa* Duch.	Japan	Sweet charmy x ?
Satsumaotome	*Fragaria ananassa* Duch.	Japan	8821-11 x Kurume 52gou
Ohkimi	*Fragaria ananassa* Duch.	Japan	Satsumaotome x Ichigochukanbohon-Nou 1gou
Sachinoka	*Fragaria ananassa* Duch.	Japan	Toyonoka x Aiberry
Akihime	*Fragaria ananassa* Duch.	Japan	Kunowase x Nyoho
Aiberry	*Fragaria ananassa* Duch.	Japan	Reiko x Hokowase
Hokowase	*Fragaria ananassa* Duch.	Japan	Kogyoku x Tahoe
06A-184	*Fragaria ananassa* Duch.	Japan	Amaou x Sanchigo
Elsanta	*Fragaria ananassa* Duch.	Netherland	Gorella x Holiday
Floridabelle	*Fragaria ananassa* Duch.	USA	Sequoia x Ealibelle
Red Pearl	*Fragaria ananassa* Duch.	Japan	Aiberry x Toyonoka
Toyonoka	*Fragaria ananassa* Duch.	Japan	Himiko x Harunoka

a) 1: PVP (Plant Variety Plotectin) Office at MAFF, JAPAN (http://www.hinsyu.maff.go.jp/en/en_top.htn
ARS (Agricultural Research Service) (http://www.usda.gov/wps/portal/usda/usdahome)
b) Application Number

Accession code	Reference[a)]	Note
4293	1	
15540	1	https://www.jstage.jst.go.jp/article/hrj/10/1/10_1_121/_pdf
16153	1	http://www.pref.miyagi.jp/soshiki/res_center/n 79.html
20497	1	
17158	1	http://www.pref.tochigi.lg.jp/g61/seika/docume ts/kp_058_05.pdf
PI 616761	2	
3754	1	
7649	1	http://agriknowledge.affrc.go.jp/RN/2010560 5.pdf
16154	1	
14187	1	
5196	1	
12576	1	
716	1	http://www.agrinet.pref.tochigi.lg.jp/81_area desaki/10_nousi/04_kenkyuuseika/g31_seika01/ seika/kenhou/kp_031/kp_031_03.pdf
2048	1	
26477[b)]	1	
10371	1	
15261	1	http://www.pref.aichi.jp/nososi/seika/hokoku/h koku37/37-17s.pdf
19529	1	
19473	1	
17391	1	
8839	1	
19203	1	
9654	1	
20810	1	
7650	1	
2991	1	
ESP138-0032	2	
PI 617007	3	
-	Unpublished	
PI 551579	3	
PI 551396	3	
3755	1	
615	1	

nl), 2: GenBerry Database ://www.bordeaux.inra.fr/eustrawberrydb/), 3: USD

Table S3. Sequences of the adaptors and primers used in this study

Primer name	Sequence (5′ <- 3′)
For sequencing library construction	
Forked_Type1	AATAGGGCTCGAGCGGCAGCTATTAATAGTACT
Forked_Type2	AATAGGGCAGCTGCGGCAGCTATTAATAGTACT
Forked_Type3	AATAGGGCGATGGCGGCAGCTATTAATAGTACT
Forked_Type4	AATAGGGCCTACGCGGCAGCTATTAATAGTACT
Forked_Com	GTACTATTAATAGCATCTTCGTTCGTCGAT
AP2_Type1	AATAGGGCTCGAGCGGC
AP2_Type2	AATAGGGCAGCTGCGGC
AP2_Type3	AATAGGGCGATGGCGGC
AP2_Type4	AATAGGGCCTACGCGGC
AP3_Type1	TCGAGCGGCAGCTATTAATAGTACT
AP3_Type2	AGCTGCGGCAGCTATTAATAGTACT
AP3_Type3	GATGGCGGCAGCTATTAATAGTACT
AP3_Type4	CTACGCGGCAGCTATTAATAGTACT
Fr_Mal_iMET_1	AGACTGCNNGCTCTGATACCA
Fr_Mal_iMET_2	ATGATCGCNNGCTCTGATACCA
Fr_Mal_iMET_3	CGTCCAANNGCTCTGATACCA
Fr_Mal_iMET_4	CTTGACCNNGCTCTGATACCA
Fr_Mal_iMET_5	GACTAGTCNNGCTCTGATACCA
Fr_Mal_iMET_6	GAGTGTGNNGCTCTGATACCA
Fr_Mal_iMET_7	TCAGCTAGNNGCTCTGATACCA
Fr_Mal_iMET_8	TCCAGATGNNGCTCTGATACCA
Fr_Mal_Asp_1	AGACTGCAGACGGCGCCA
Fr_Mal_Asp_2	ATGATCGAGACGGCGCCA
Fr_Mal_Asp_3	CGTCCAAAGACGGCGCCA
Fr_Mal_Asp_4	CTTGACCAGACGGCGCCA
Fr_Mal_Asp_5	GACTAGTAGACGGCGCCA
Fr_Mal_Asp_6	GAGTGTGAGACGGCGCCA
Fr_Mal_Asp_7	TCAGCTAAGACGGCGCCA
Fr_Mal_Asp_8	TCCAGATAGACGGCGCCA
For S-SAP	
FA_*Mse* I	TAAGTACTATTAATAGCATCTTCGTTCGTCGAT
FA_*Rsa* I	AGTACTATTAATAGCATCTTCGTTCGTCGAT
FA_cmpl	AATAGGGCTCGAGCGGCAGCTATTAATAGTACT
Met_CL_3_1st	CCCGCTCTGATACCATGTC

Met_CL_11_1st	GCTCTGATACCAAACTTATCCATCC
Met_CL_20_1st	GCTCTGATACCAGGCCAATG
Met_CL_28_1st	GCTCTGATACCAGTTATTAGTACTGG
Met_CL_76_1st	GCTCTGATACCACCGCAATC
Met_CL_3_2nd	GGGATCTCTCCTCTCCAACA
Met_CL_11_2nd	CACTATTTCTCTTCTTTCTGAACAACTC
Met_CL_20_2nd	GAACCATCTATTTTTCATATTGGCAGCC
Met_CL_28_2nd	CTCGAAGAACTGACAGAAAATTAACACAG
Met_CL_76_2nd	GCGCTTCGGGAGAGAAGTG