2007

A tale of two templates: Automatically resolving double traces has many applications, including efficient PCR-based elucidation of alternative splices

Aaron E. Tenney
Washington University in St. Louis

Jia Qian Wu
Yale University

Laura Langton
Washington University in St Louis

Paul Klueh
Washington University in St Louis

Ralph Quatrano
Washington University in St Louis

See next page for additional authors

Recommended Citation

Tenney, Aaron E.; Wu, Jia Qian; Langton, Laura; Klueh, Paul; Quatrano, Ralph; and Brent, Michael R., ,"A tale of two templates: Automatically resolving double traces has many applications, including efficient PCR-based elucidation of alternative splices." Genome Research. 17,. 212-218. (2007).
https://digitalcommons.wustl.edu/open_access_pubs/2012

Authors

Aaron E. Tenney, Jia Qian Wu, Laura Langton, Paul Klueh, Ralph Quatrano, and Michael R. Brent

CSH PRESS GENOME RESEARCH

A tale of two templates: Automatically resolving double traces has many applications, including efficient PCR-based elucidation of alternative splices

Aaron E. Tenney, Jia Qian Wu, Laura Langton, et al.

Genome Res. 2007 17: 212-218 originally published online January 8, 2007
Access the most recent version at doi:10.1101/gr.5661407

A tale of two templates: Automatically resolving double traces has many applications, including efficient PCR-based elucidation of alternative splices

Aaron E. Tenney,[1,4] Jia Qian Wu,[2,4] Laura Langton,[1] Paul Klueh,[3] Ralph Quatrano,[3] and Michael R. Brent[1,5]

[1]Laboratory for Computational Genomics and Department of Computer Science, Washington University, St. Louis, Missouri 63130, USA; [2]Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, Connecticut 06620-8103, USA; [3]Department of Biology, Washington University, St. Louis, Missouri 63130, USA

Trace Recalling is a novel method for deconvoluting double traces that result from simultaneously sequencing two DNA templates. Trace Recalling identifies up to two bases at each position of such a trace. The resulting *ambiguity sequence* is aligned to the genome, identifying one template sequence. A second template sequence is then inferred from this alignment. This technique makes possible many exciting biological applications. Here we present two such applications, alternate splice finding and elucidation of multiple insertion sites in a random insertional mutagenesis library. Our results demonstrate that RT–PCR followed by Trace Recalling is a more efficient and cost effective way to find alternate splices than traditional methods. We also present a method for mapping double-insertion events in a random insertional-mutagenesis library.

[Supplemental material is available online at www.genome.org. Alternate splice forms discovered during this work have been deposited in GenBank under accession nos. EB71062–EB710342.]

During normal Sanger sequencing (Sanger et al. 1977), a collection of fluorescently labeled DNA fragments representing all initial subfragments of the target template are separated by length. A sequencing machine performs gel electrophoresis and scans the separated products for florescence, producing a chromatogram. Ideally, when a single template is sequenced, the chromatogram consists of a clearly defined set of regularly spaced peaks of similar height, representing the sequence of the DNA template (Fig. 1A). Effective base-calling programs have been written to extract the DNA sequence from such a chromatogram (Ewing et al. 1998). Not all traces, however, are this simple. In Figure 1B, for example, peaks are regularly spaced and of similar height, but, clearly, two bases are represented by each pair of overlapping peaks, whereas there is only one base for each peak location in Figure 1A. Such *double traces* are generated when polymerization occurs simultaneously on two DNA templates. A double trace is the superposition of the traces that would have been generated had the templates been sequenced in separate reactions. The essential problem faced when analyzing a double trace is that two bases are represented at each position, and it is impossible to tell from which template each base came by examining only the trace. Thus, the number of possible pairs of templates that could give rise to a particular double trace increases exponentially with the length of the double trace.

Double traces occur in a number of biological and biotechnological applications and have been observed since the early days of fluorescent dye sequencing (Gibbs et al. 1990). For example, a double trace is generated when an alternatively spliced region of a transcript is amplified by RT–PCR and sequenced directly—i.e., without cloning (Wu et al. 2004). Currently, such traces are often discarded as uninterpretable, which reduces the success rate in testing gene predictions by RT–PCR. If the sequences of the two templates could be deconvoluted (separated) computationally, these failures could be turned into successes. Furthermore, double traces would yield the sequences of both isoforms for the price of one. This method can also be used to check a particular region of a gene for alternative splices. Currently, the most reliable method for sequencing both isoforms is to ligate the PCR product into a vector, transform into competent cells, and sequence multiple clonal colonies. If one isoform is five times less abundant than the other, 10–20 clones must be sequenced in order to be reasonably certain of getting both forms. This is an expensive and time-consuming procedure, and it will yield no additional information if the targeted transcript region is not alternatively spliced. The somewhat less-expensive alternative, gel purification of the PCR product, followed by eluting and sequencing DNA from each gel band, still requires multiple sequencing reactions. Furthermore, it is less sensitive, since a substantial fraction of products that can be sequenced are not visible on ordinary agarose gels (Wu et al. 2004).

Double traces have also been observed in random insertional mutagenesis experiments. Recently a collection of 127,760 knockout *Arabidopsis thaliana* lines was created by using Agrobacterium T-DNA (Alonso et al. 2003). Localization of the insertion event to determine which, if any, gene has been disrupted involves amplifying and sequencing a short segment of genomic DNA that is adjacent to the insertion from universal primers. Ideally, there is only one insertion event and a clean single trace is obtained. If two insertion events occur, however, a double trace is generated. In these cases, the mutant strain is unusable since one or both insertion sites cannot be efficiently recovered from the double trace. Ecker and colleagues (Alonso et al. 2003) esti-

[4]These two authors contributed equally to this work.
[5]Corresponding author.
E-mail brent@cse.wustl.edu; fax (314) 935-7302.

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.5661407.

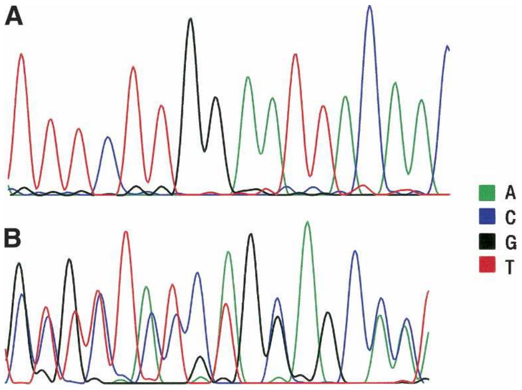

Figure 1. Examples of chromatograms produced by sequencing (*A*) one and (*B*) two templates.

mate that, on average, there are ~1.5 insertion events per line in this library. This means that many of their lines are unusable, and there is no way to identify these until the final sequencing step. Similar random insertional mutagenesis techniques have been used for other organisms (Lee et al. 1999; Garsin et al. 2004).

We have developed a method to analyze double traces that we call Trace Recalling. Trace Recalling works by recasting the de novo base-calling problem as a database search and alignment problem. Current base-calling programs are designed for the de novo sequencing of completely unknown sequences such as are encountered in a whole-genome sequencing project. Since the completion of sequencing projects for *Homo sapiens* (Lander et al. 2001) and numerous model organisms (e.g., Waterston et al. 2002; Mikkelsen et al. 2005), however, an increasing fraction of sequencing capacity is devoted to resequencing cDNA or genomic DNA from organisms for which a reference genome sequence exists. Trace Recalling works by calling either one or two bases at each position in a double trace. The resulting sequence of two-place ambiguity codes and bases is called an *ambiguity sequence* (Fig. 2). This ambiguity sequence is aligned to all or part of an assembled genome sequence by an alignment algorithm designed to handle ambiguity codes. The genomic sequence that best aligns to the ambiguity sequence is assumed to be the sequence of one template present in the sequencing reaction. The bases from this template are "subtracted" from the ambiguity sequence, resulting in a second sequence.

Results

Trace Recalling algorithm

The inputs to Trace Recalling are a chromatogram and a reference genomic sequence. The chromatogram is processed with the base caller PHRED (Ewing et al. 1998) using default parameters and the –d option. This option causes PHRED to output the two best bases corresponding to each base normally called if there is a good secondary peak. The next step is to construct an ambiguity sequence for the chromatogram by analyzing each base or pair of bases called by PHRED. Our algorithm retains the secondary base if the ratio of the areas of the smaller to larger peak exceeds a threshold (*peak area ratio threshold*). If two bases are detected, the corresponding character in the ambiguity sequence is set to the IUPAC two-place ambiguity code for the observed bases (Fig. 2). If only one base is observed, then an unambiguous DNA symbol (A, C, G, or T) is used. The ambiguity sequence is aligned to the genome with a modified version of the cDNA-to-genome alignment program EST_GENOME (Mott 1997). This version treats an alignment between an unambiguous genomic base and a compatible ambiguous base as a match. The result is called the *primary alignment*. The genomic sequence that aligns to the ambiguity sequence in the primary alignment is assumed to be the sequence of one of the templates. The sequence of the other template (*recalled sequence*) is inferred from the primary alignment by considering the five cases shown in Figure 3. The final step of Trace Recalling is to align the recalled sequence to the genome to determine whether it results from polymerization on two templates or simply noise in the chromatogram. In the former case, the recalled sequence should align to the genomic source of the second template; in the latter, it should have no high-scoring alignments to the genome.

Automatic detection of alternate splice double traces

One application of Trace Recalling is screening traces from direct sequencing of RT–PCR products for evidence of alternative splicing. Prior knowledge of the boundaries of alternate splices is not required for this application; they are discovered through the primary and secondary alignments. Our software compares the primary and secondary alignments obtained from each trace. If the primary and secondary alignments are identical, there is no evidence of alternative splicing. If there is no positive scoring secondary alignment, the recalled sequence is most likely noise and there is no evidence of alternative splicing. If the two alignments overlap, but differ in their internal exon–intron structure,

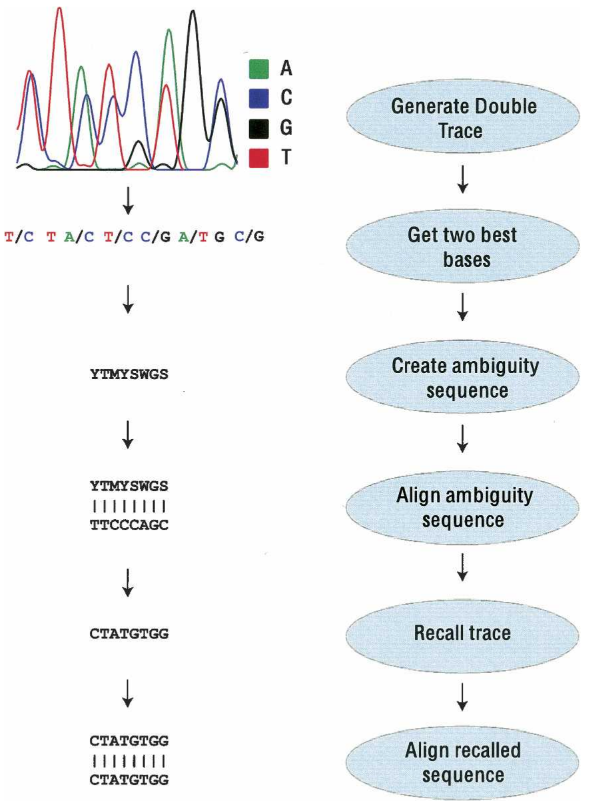

Figure 2. Short example and block diagram of the Trace Recalling algorithm.

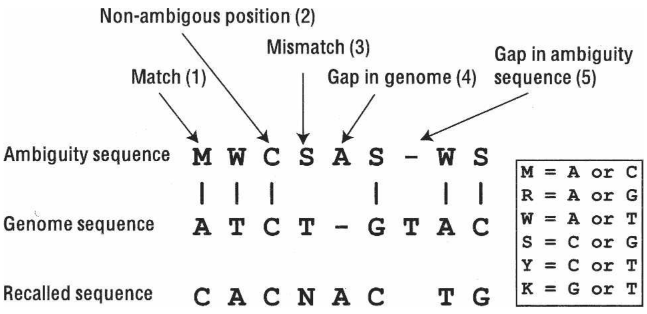
Non-ambigous position (2)
Mismatch (3)
Gap in ambiguity sequence (5)
Match (1)
Gap in genome (4)
Ambiguity sequence M W C S A S - W S
Genome sequence A T C T - G T A C
Recalled sequence C A C N A C T G
M = A or C
R = A or G
W = A or T
S = C or G
Y = C or T
K = G or T

Figure 3. Detailed example of cases that arise in the recalling step of Trace Recalling. (1) When an ambiguity symbol from the trace matches a base from the genome, the genomic base is subtracted from the ambiguity symbol. For example, a C is called when an M in the ambiguity sequence matches an A in the genomic sequence (M = A or C) (Gibbs et al. 1990). In a good double trace, this is the most frequently seen case and represents a position where different bases were sequenced at the same position in the trace. (2) In the case of a match between two unambiguous bases, that base is called in the recalled sequence. This means that there is only one peak observed and both templates have the same base at that position. In a double trace, we expect this to happen in about a quarter of the positions. (3) When there is a mismatch, we call an N in the corresponding position of the recalled sequence. Clearly, some base is present at this position, but we have no way of knowing what it is. (4) When there is a gap in the genome sequence, the ambiguity symbol aligned to the gap is left in the recalled sequence. This case could represent an insertion in the primary template sequence with respect to the genomic sequence. (5) In the case of a gap in the ambiguity sequence, nothing is called in the recalled sequence. This could represent a deletion in the primary template with respect to the reference sequence.

they are further analyzed for several types of alternative splices: alternate exon, clean alternate exon or cassette exon (an alternate exon in which the boundaries of the adjacent exons are identical in both alignments), alternate 5′ or 3′ splice sites, or retained intron. Details of the method used to compare the two alignments are presented in the Methods section. If none of these alternate splice forms are found and the alignments are not identical, there is no evidence of an alternate splice.

The peak area ratio threshold is used to discriminate between peaks from a second template and noise peaks. If an alternate splice is present and the threshold is set too high, the alternate splice is not detected because there are not enough peaks from the secondary template that pass the threshold, so the recalled sequence does not align properly to the secondary isoform of the gene. If it is set too low, noise peaks obscure the portion of the trace that is the same in both templates, including the portion between the sequencing primer and the locus where the two templates diverge. As a result, that portion of the trace fails to align, causing the alternate splice to not be detected. Because setting the threshold either too high or too low results in failure to detect true alternate splices rather than erroneous splice detection, there is no harm in trying several thresholds. Thus, to find alternate splices, we run Trace Recalling with several different thresholds (1/2, 1/3, . . .1/20) and analyze the results as a group. If the results are consistent with the above-described pattern, the gene is marked as alternately spliced. A visualization of this procedure is presented in Supplemental Figure 8 for several traces in the experiment described below.

Alternate splicing experiment

We designed an experiment to test the ability of our system to deconvolute double traces generated by sequencing RT–PCR products from alternately spliced genes. As described in the Methods, we selected 48 human genes containing an optional exon based on the presence of two RefSeq genes, one containing the target exon and one omitting it. PCR primers were designed to amplify two segments of each gene, one containing the clean alternate exon and one containing only constitutive splices. The constitutive targets serve both as negative controls for alternate splices and verification that the targeted gene is present. RT–PCR reactions were carried out in an mRNA pool from 20 human tissues and sequenced as described in the online methods. Two traces were created per target, one from each PCR primer. Alternately spliced targets are expected to yield a double trace. The traces were examined with Trace Recalling using the genomic sequence spanned by the whole RefSeq gene as the reference genomic sequence. Targets were characterized as either containing or not containing a clean alternate exon. We also attempted to sequence 12 cloned plasmids containing PCR products for each target to compare the efficacy of these two alternate splice-finding methods. Results of this experiment are presented in the Venn diagram in Figure 4. Trace Recalling identified many more of the targets thought to be alternately spliced based on RefSeq evidence than cloning did. It also predicted an additional optional exon that appears to have no mRNA or EST support.

We analyzed another set of possible alternate splice-derived double traces that were generated as part of the Mammalian Gene Collection (MGC) project (Gerhard et al. 2004). These traces are similar to the ones used in the controlled alternate splice experiment. They represent sequences of short (500–800 bp) amplicons RT–PCRed from a pool of human mRNA as described in the online methods. The unambiguous sequence of each trace as called by PHRED in default mode was aligned to the whole human genome with BLAT (Kent 2002). The sequence of the best hit along with 1000 bp of flanking sequence was extracted and used as the reference genomic sequence for Trace Recalling. Trace Recalling was allowed to search for all types of alternate splices described in the previous subsection (not just clean alternate exons). A total of 6106 MGC traces were examined without prior screening to detect double traces. Of these traces, 4068 aligned to the genome in the first alignment stage, of which 2165 were considered high-quality spliced alignment by the traditional MGC analysis. Trace Recalling detected evidence

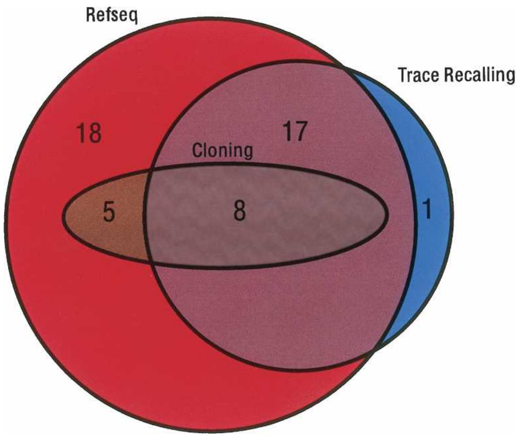

Figure 4. Results of the alternate splicing experiment. (Red circle) A priori expected alternately spliced targets; (blue circle) targets flagged as containing clean alternate exons by Trace Recalling; (dark oval) targets identified by cloning experiment.

for clean alternate exons in 155 traces, alternate exons with other changes in 96 traces, alternate splice sites in 73 traces, and retained introns in 58 traces. Examples are shown in Supplemental Figures 1 and 4–7. Of the 382 traces detected as containing alternate splices, 51 were not identified as high-quality spliced alignments by the traditional analysis pipeline. So, in 331 instances, Trace Recalling added one isoform, and in 51 instances, it added two isoforms not previously seen. This means that Trace Recalling increased the number of high-quality spliced alignments by 20%. Sequences representing an alternate form of a clean alternate exon not previously detected have been deposited in GenBank (accession nos. EB71062–EB710342). In cases where neither form had been previously identified, both sequences were deposited.

Random insertional mutagenesis experiment

We tested the ability of Trace Recalling to recover the locations of multiple insertion sites in a random insertional mutagenesis experiment. A diagram of the analysis pipeline is presented in Figure 5. When we applied this analysis pipeline to a set of 38,033 traces, we obtained the following results. A total of 15,986 traces do not align in the first BLAST step. These may represent cases where there was no insertion event. This yield was expected, given that in the original published analysis, only 88,000 of the 150,000 lines were found to contain at least one insert (Alonso et al. 2003). In our analysis, another 18,728 are classified as single traces because they have no double-trace segments (6592), or their recalled sequences do not align well to the genome in the second BLAST step (12,136). This leaves 3319 traces for which the cleaned recalled sequence aligns well to the genome. Of these, 1609 traces predict two insertion events on different chromosomes representing likely double-insertion events.

We carried out an experiment to examine a subset of the 1609 *Arabidopsis* lines, in which we predict two T-DNA insertion events on different chromosomes. Of these, 66 were selected for validation. In all but one of these, one of the two predicted insertion sites coincided with the insertion site predicted by the previously published method of BLASTing the original read (without ambiguity codes) against the genome (Alonso et al. 2003), even though this was not a criterion for selection. For each plant line tested, seeds were obtained from the *Arabidopsis* Biological Resource Center (http://www.biosci.ohio-state.edu/pcmb/Facilities/abrc/abrchome.htm). Plants were grown and whole-genomic DNA extracted as described in the Methods. PCR primers were designed spanning the left border of both predicted T-DNA insertion events in each mutant line, with one primer complementary to the T-DNA sequence and the other to the flanking genomic sequence within 500 bp of the predicted insertion site. PCR was carried out on whole-genomic DNA, and PCR products were sequenced using the PCR primers as sequencing primers.

Amplification of a predicted insertion site alone provides evidence that the prediction was correct, since the primer in the flanking genomic sequence was chosen to be unique in the *Arabidopsis* genome. Sequencing of the PCR product and alignment of that sequence to the predicted locus provides stronger evidence that the prediction was correct. Therefore, in our analysis we define two types of confirmation, confirmation by amplification and confirmation by alignment. In the former, a T-DNA/genome junction is amplified, but the genomic sequence does not align well to the genome. In the latter, the genomic sequence does align well to the predicted genome location. Of the 132 individual insertion events tested, 59 were confirmed by alignment, 18 were confirmed by amplification, 14 aligned to an unexpected locus, and another 41 showed no evidence of amplification. In 17 of the plant lines tested, both predicted insertion events were verified by alignment. In another seven lines, the primary insertion event was verified by alignment and the secondary insertion event was verified by amplification. Thus, we were able to verify 36% of the tested double-insertion events. An analysis of predicted multiple insertions that were not verified is presented in Table 1.

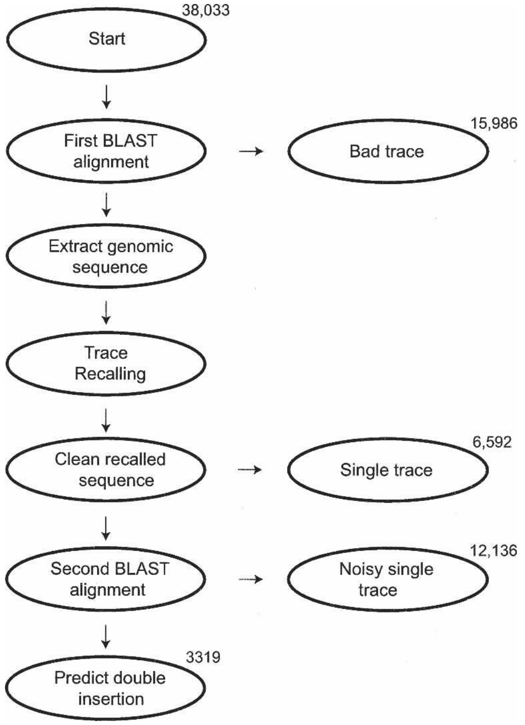

Figure 5. Diagram of pipeline used to analyze the *Arabidopsis* random insertional mutagenesis experiment. Numbers of traces present at key stages of the pipeline are noted. The trace is first called without ambiguity codes using PHRED with default parameters. This sequence is aligned to the whole *Arabidopsis* genome using BLAST. If there is no significant alignment, the trace is discarded. Otherwise, the aligned genomic sequence plus 1000 bases of flanking sequence on either side of the alignment are extracted. This is assumed to be the locus of one insertion event. Trace Recalling is applied to this extracted genomic segment and the trace. If the trace is a double trace resulting from two insertion events, the recalled sequence is a chimera of the T-DNA sequence, the genomic sequence flanking the second insertion, and the single-trace portion of the sequence flanking the first insertion. Therefore, in the next step we remove single-trace segments of the sequence by removing any subsequences of the recalled sequence that align well to the originally called sequence. This is called cleaning the recalled sequence. If this step removes all of the recalled sequence, the trace is classified as a single-insertion event and removed from the pipeline. Otherwise, the remaining recalled sequence is aligned to the genome with BLAST. If this alignment is not significant, we assume the recalled sequence represents noise and the trace is classified as a noisy single trace. However, if the alignment is significant, we call the trace as a double trace representing two insertion events and predict the locus of the second insertion as the location of the second BLAST hit.

Discussion

We have demonstrated that Trace Recalling can find alternate splices in single-pass reads of RT–PCR products. In a controlled

Table 1. Explanation of predicted multiple insertion predictions that were not verified

Times observed	Prediction A	Prediction B
12	Confirmed by alignment	No usable sequence
8	Confirmed by alignment	Hit unexpected locus
7	Confirmed by amplification	No usable sequence
2	Confirmed by amplification	Hit unexpected locus
2	Hit unexpected locus	No usable sequence
1	Hit unexpected locus	Hit unexpected locus
10	No usable sequence	No usable sequence

"Confirm by alignment" and "confirm by amplification" are defined in the text. "Hit unexpected locus" means a genomic sequence other than the predicted one was observed. "No usable sequence" means that after quality trimming no sequence was left, possibly indicating either failure to amplify the target due to an incorrect prediction or sporadic PCR failure.

experiment, we tested the ability of Trace Recalling to deconvolute double traces created by sequencing the RT–PCR products of two isoforms of a gene simultaneously. By examining only two reads per target, Trace Recalling was able to correctly identify both isoforms in a majority (25) of the targets that were previously known to be alternately spliced. In contrast, the cloning and sequencing strategy required much greater effort (cloning and sequencing multiple inserts per target), yet it identified less than half as many known alternately spliced targets. Trace Recalling also predicts a potentially novel optional exon in this experiment. RT–PCR products were also run on a gel in an attempt to quantify the number of templates present after amplification (Supplemental Fig. 2). In six of the missed targets, only one band appeared on the gel. These likely represented cases in which the second isoform was not expressed in the tissues tested. In seven of the missed targets, there were three or more bands. The excess bands may be nonspecific amplification or additional isoforms; in either case, it is likely that Trace Recalling was confused by the extra peaks. Finally, there were 10 targets in which exactly two bands were present and Trace Recalling failed to detect alternate isoforms. Examination of the traces indicates that in six cases there is a very low concentration of the secondary template. In another two cases, two sets of peaks are visible but they are shifted with respect to each other, and so the secondary template is lost. Finally, in two other cases, three sets of peaks are clearly present even though only two bands on the gel are visible. The most serious limitation of Trace Recalling appears to be secondary template peaks getting lost in noisy traces.

Trace Recalling makes possible a protocol in which most alternate splices are elucidated by one or two sequencing reactions followed by a targeted experiment to confirm the alternate splice. Our experiments suggest that to get similar results by blindly cloning and sequencing from a pool of RT–PCR products would require many more sequencing reactions. Sequencing full-length cDNA clones is still required for determining the global structures of the isoforms when more than one region shows alternative splicing, but Trace Recalling can determine the local structure with greater sensitivity and at a much lower cost. We also demonstrated the application of this method to a high throughput RT–PCR project. We have identified a large set of MGC targets that are likely to be alternately spliced, enhancing the value of those experiments.

We have also demonstrated that Trace Recalling can be used to screen for multiple insertion sites in a random insertional mutagenesis library. We found that a substantial fraction of the hypotheses generated for double-insertion sites can be verified by PCR and sequencing. Trace Recalling was used to screen a set of 38,033 traces generated as part of an *Arabidopsis* mutagenesis library yielding 1609 traces thought to represent double-insertion events. We experimentally tested 66 of these lines, and by our most stringent classification (confirmation by alignment), 17 lines were shown to contain inserts in the predicted locations. By a less-stringent classification (confirmation by amplification), another seven lines were shown to contain inserts at the predicted locations. Experimental verification of predicted multiple insertion sites is required, but Trace Recalling provides a method to design these verification experiments. The only other available method for identifying multiple insertion sites is the one used by the authors of the *Arabidopsis* study—blindly resequencing the tag sequences used for mapping. Sometimes this results in the secondary tag being more pronounced in a different trace, allowing the identification of the secondary insertion site.

Trace Recalling has difficulty with several types of traces, including those in which the secondary template signal is near the level of the noise in the trace. Trace Recalling ignores secondary peaks <1/20th the area of the primary peak, since in practice these very small secondary peaks are almost always noise. While this threshold makes it impossible to detect very weak secondary templates, it significantly reduces problems associated with the background noise present in all traces. This does not appear to be a limiting factor in the analysis of the MGC traces (Supplemental Fig. 9). Another difficulty occurs when slight differences in mobility rates of DNA molecules cause the peaks to become off register or out of phase. This interferes with Trace Recalling in two ways. First, it becomes difficult for PHRED to correctly identify secondary peaks. Second, it causes many single base-pair gaps in both alignment stages. Such gaps are heavily penalized to reduce spurious alignments. This often results in the loss or premature truncation of the alignment. Another problem arises when more than two templates of nearly equal concentration are simultaneously sequenced. If this occurs, peaks from the secondary and tertiary templates become interleaved in the recalled sequence, interfering with the secondary alignment step.

In addition to the applications explored here, Trace Recalling may be useful in other applications where double traces are encountered. For example, direct sequencing of genomic amplicons from an individual heterozygous for an indel polymorphism results in a double trace. This is because the chromosome sequences are shifted with respect to each other by the length of the indel. The ABI KB base caller is designed to handle this circumstance for short indels by attempting to shift the sequence represented in the trace with respect to itself looking for matches between the shifted sequences. This strategy, however, is limited to analyzing indels no longer than 15 bp (ABI representative, pers. comm.). Trace Recalling has no length restriction, since the reference genome sequence could be used to deconvolute such traces. As evidence of this, the alternate splicing work presented here demonstrates that Trace Recalling can handle gaps the size of introns that are often many kilobases in length. Another potential future application for Trace Recalling involves whole-genome shotgun resequencing. Double traces could be generated by sequencing two pooled clones in the same reaction, and the reference genome sequence could be used to deconvolute them, yielding two reads per sequencing reaction.

Methods

Computational methods

Automatic identification of traces resulting from alternate splicing

Method for detecting alternate splices at a single threshold

GTF (a standard format for recording genome annotations) files are created from the primary and secondary alignments of Trace Recalling, representing the coordinates of exons in these alignments in the coordinate system of the reference genomic sequence. These GTF files are used to create an "indicator string" for the pair of alignments. For each base in the genomic reference sequence, if it is contained in an exon from both alignments, the corresponding position of the indicator string is set to "2". If the position is covered by an exon in only one alignment, the indicator string position is set to "1". Finally, if the position is not covered by exons in either alignment, the indicator string position is set to "0". Alternate splices appear as matches to certain regular expressions in the indicator string. The Perl regular expressions used are: Clean alternate exon, 2+0+1+0+2+; Alternate exon, 2+1+0+1+0+2+ or 2+0+1+0+1+2+; Alternate splice site, 2+1+0+2+ or 2+0+1+2+; Retained intron 0+2+1+2+1*$ or ^1*2+1+2+0+ or ^1*2+1+2+1*$ or 0+2+1+2+0+.

Method for calling alternate splices using different thresholds

Trace Recalling is run at several different peak-area ratio thresholds. In the implementation presented here, the thresholds 1/2, 1/3, . . . , 1/20 are used. Alternate splices are detected for each individual threshold by the method described in the previous section. Next, we verify that the same alternate splice is seen in each threshold by examining the coordinates of the alternately spliced regions. The lowest threshold at which the alternate splice is seen is referred to as the critical threshold. Below the critical threshold, enough noise peaks are allowed into the single-trace section of the recalled trace to disrupt that part of the alignment. This is tested by looking for the absence of the exon flanking the alternate splice on the single-trace side and retention of the exon flanking the trace on the double-trace side. If this pattern is observed, the trace is called as alternately spliced.

Controlled alternate splice experiment

Target selection procedure

We began with 402 genes containing cassette exons for which both isoforms could be found in the RefSeq database. For each gene we looked at the tissues in which the two isoforms were expressed by examining human mRNAs from the UCSC Genome Browser. The gene was retained only if the different isoforms had been detected in one of the 29 commonly available tissues. This left us with a set of 138 genes. Since we are testing the ability of our system to deconvolute double traces, we wanted to make sure that there were no other alternate splices in the vicinity of the cassette exon. To this end, each of the genes was visually inspected along with its RefSeq annotations and all aligned human mRNAs in the UCSC Genome Browser. Any gene for which there was mRNA evidence of other alternate splices near the cassette exon was thrown out. This left 85 candidates, of which 48 were randomly selected for testing.

Alignment parameters used for Trace Recalling

Primary alignment parameters: Match score, 2; Mismatch penalty, −6; Gap penalty, −6; Splice penalty, −20; Intron penalty, −40.

Secondary alignment parameters: Match score, 1; Mismatch penalty, −1; Gap penalty, −2; Splice penalty, −20; Intron penalty, −40.

MGC experiment

Alignment parameters used for Trace Recalling

Same as parameters used in the controlled alternate splice experiment.

Random insertional mutagenesis experiment

BLAST alignment parameters

First BLAST the PHRED-called trace against whole genome using the parameter string: wordmask = seg lcmask –cpus 1 –kap –wink 1 –hspmax 0 W 6 Q 20 R 20. Default match, mismatch, and gap penalties were used.

Second BLAST the cleaned recalled sequence against whole genome using the parameter string: –cpus 1 –kap –wink 1 –hspmax 0 W 6. Default match, mismatch, and gap penalties were used.

EST_GENOME alignment parameters used to clean recalled sequence

The recalled sequence was cleaned of single-trace regions by aligning the default PHRED-called sequence against the recalled sequence from Trace Recalling by using EST_GENOME with default parameters.

Alignment parameters used for Trace Recalling

Primary alignment parameters: Match score, 2; Mismatch penalty, −6; Gap penalty, −6; Splice penalty, −100,000; Intron penalty, −100,000.

Secondary alignment parameters: Match score, 1; Mismatch penalty, −1; Gap penalty, −2; Splice penalty, −100,000; Intron penalty, −100,000.

Biological methods

Plant germination and DNA extraction for random insertional mutagenesis experiment

To each tube of seeds, 400 µL of 95% ethanol was added, mixed, and left to stand in a sterile hood for 10 min. The ethanol was decanted off and the tubes were left to dry completely, ~1 h.

For each sample, ~30 seeds were shaken into Murashige & Skoog agar plates, the plates sealed with micropore tape, and plates placed in a 4°C refrigerator for 48 h.

After the 48 h, the plates were placed into a 25°C, 16-h light incubator for 5 d.

After 5 d, a single leaf was cut from each surviving seedling. Leaves from each plate were pooled into a single tube (one per plate) and immediately immersed into liquid nitrogen.

Using a small pestle, each collection was ground within liquid nitrogen and after evaporation, further ground in the genomic filter extraction buffer (0.2 M Tris-HCL at pH 9.0, 0.4 LiCl, 25 mM EDTA, and 1% SDS). The samples were then centrifuged at high in a tabletop centrifuge and the supernate added to an equal part isopropanol, mixed, and again centrifuged at high to pellet the DNA. The liquid was decanted and the tubes allowed to dry.

DNA was resuspended in 400 µL of TE buffer and 2 µL used for initial PCR.

PCR and sequencing for alternate splice and MGC experiments

Equal amounts of total RNA were pooled from 20 human tissues including adrenal gland, bone marrow, cerebellum, brain (whole), fetal brain, fetal liver, heart, kidney, liver, lung, placenta, prostate, salivary gland, skeletal muscle, spleen, testis, thymus, thyroid gland, trachea, uterus (Human Total RNA master panel II, BD Biosciences Clontech). Pooled total RNA was reverse transcribed using Superscript II or III reverse transcriptase with Oligo dT primer according to the manufacturer's instructions (Invitrogen). RT was followed by PCR amplification using either the Clontech Advantage 2 PCR Enzyme System (controlled alternate splice finding experiment) or Phusion high-fidelity DNA polymerase (New England Biolabs; MGC and random insertional mutagenesis experiments). All experiments were performed using touchdown PCR (Don et al. 1991) with differing cycle parameters (see Supplemental material). In the controlled alternate splice finding experiment, PCR products were both cloned and directly sequenced. In the MGC and random insertional mutagenesis experiments, PCR products were directly sequenced.

Acknowledgments

We thank Joseph Ecker for providing traces for the *Arabidopsis* random insertional mutagenesis experiment. We also thank Randall Brown and Jeremy Buhler for many helpful discussions about the Trace Recalling algorithm and Beth Frazier for her valuable biological insights. We are grateful to Richard Gibbs and John McPherson for encouragement, interesting discussions, and supervision of the alternate splice experiment. M.R.B. was supported in part by R01 HG02278 and in part by the National Cancer Institute for the Mammalian Gene Collection project under Contract No. N01-CO-12400. A.T. was supported in part by T32 grant HG000045. P.K. was supported in part from funds provided by Washington University, Department of Biology. The methods here are covered under patent application PCT/VS05/003681.

References

Alonso, J.M., Stepanova, A.N., Leisse, T.J., Kim, C.J., Chen, H., Shinn, P., Stevenson, D.K., Zimmerman, J., Barajas, P., Cheuk, R., et al. 2003. Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* **301:** 653–657.

Don, R.H., Cox, P.T., Wainwright, B.J., Baker, K., and Mattick, J.S. 1991. 'Touchdown' PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Res.* **19:** 4008.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8:** 175–185.

Garsin, D.A., Urbach, J., Huguet-Tapia, J.C., Peters, J.E., and Ausubel, F.M. 2004. Construction of an *Enterococcus faecalis* Tn917-mediated-gene-disruption library offers insight into Tn917 insertion patterns. *J. Bacteriol.* **186:** 7280–7289.

Gerhard, D.S.L., Wagner, E.A., Feingold, C.M., Shenmen, L.H., Grouse, G., Schuler, S.L., Klein, S., Old, R., Rasooly, P., Good, M., et al. 2004. The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res.* **14:** 2121–2127.

Gibbs, R.A., Nguyen, P.N., Edwards, A., Civitello, A.B., and Caskey, C.T. 1990. Multiplex DNA deletion detection and exon sequencing of the hypoxanthine phosphoribosyltransferase gene in Lesch-Nyhan families. *Genomics* **7:** 235–244.

Kent, W.J. 2002. BLAT–the BLAST-like alignment tool. *Genome Res.* **12:** 656–664.

Lander, E.S.L.M., Linton, B., Birren, C., Nusbaum, M.C., Zody, J., Baldwin, K., Devon, K., Dewar, M., Doyle, W., FitzHugh, R., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Lee, M.S., Dougherty, B.A., Madeo, A.C., and Morrison, D.A. 1999. Construction and analysis of a library for random insertional mutagenesis in *Streptococcus pneumoniae:* Use for recovery of mutants defective in genetic transformation and for identification of essential genes. *Appl. Environ. Microbiol.* **65:** 1883–1890.

Mikkelsen, T.S., Hillier, L.W., Eichler, E.E., Zody, M.C., Jaffe, D.B., Yang, S.-P., Enard, W., Hellmann, I., Lindblad-Toh, K., Altheide, T.K., et al. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437:** 69–87.

Mott, R. 1997. EST_GENOME: A program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.* **13:** 477–478.

Sanger, F., Nicklen, S., and Coulson, A.R. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* **74:** 5463–5467.

Waterston, R.H.K., Lindblad-Toh, E., Birney, J., Rogers, J.F., Abril, P., Agarwal, R., Agarwala, R., Ainscough, M., Alexandersson, P., An, S.E., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Wu, J.Q., Shteynberg, D., Arumugam, M., Gibbs, R.A., and Brent, M.R. 2004. Identification of rat genes by TWINSCAN gene prediction, RT-PCR, and direct sequencing. *Genome Res.* **14:** 665–671.

Received June 19, 2006; accepted in revised form November 22, 2006.