14 Levin, M. (2003) Motor protein control of ion flux is an early step in embryonic left–right asymmetry. *BioEssays* 25, 1002–1010
15 Raya, A. *et al.* (2004) Notch activity acts as a sensor for extracellular calcium during vertebrate left–right determination. *Nature* 427, 121–128
16 Rand, M.D. *et al.* (1997) Calcium binding to tandem repeats of EGF-like modules. Expression and characterization of the EGF-like modules of human Notch-1 implicated in receptor-ligand interactions. *Protein Sci.* 6, 2059–2071
17 Yu, J.K. *et al.* (2002) An amphioxus nodal gene (AmphiNodal) with early symmetrical expression in the organizer and mesoderm and later asymmetrical expression associated with left–right axis formation. *Evol. Dev.* 4, 418–425
18 Yasui, K. *et al.* (2000) Left–right asymmetric expression of *BbPtx*, a Ptx-related gene, in a lancelet species and the developmental left-sidedness in deuterostomes. *Development* 127, 187–195
19 Morokuma, J. *et al.* (2002) *HrNodal*, the ascidian *nodal*-related gene, is expressed in the left side of the epidermis, and lies upstream of *HrPitx*. *Dev. Genes Evol.* 212, 439–446
20 Boorman, C.J. and Shimeld, S.M. (2002) The evolution of left–right asymmetry in chordates. *BioEssays* 24, 1004–1011
21 Boorman, C.J. and Shimeld, S.M. (2002) Pitx homeobox genes in *Ciona* and amphioxus show left–right asymmetry is a conserved chordate character and define the ascidian adenohypophysis. *Evol. Dev.* 4, 354–365
22 Mazet, F. *et al.* (2003) Phylogenetic relationships of the Fox (Forkhead) gene family in the Bilateria. *Gene* 316, 79–89
23 Kortschak, R.D. *et al.* (2003) EST analysis of the cnidarian Acropora millepora reveals extensive gene loss and rapid sequence divergence in the model invertebrates. *Curr. Biol.* 13, 2190–2195
24 Hermann, G.J. *et al.* (2000) Left–right asymmetry in *C. elegans* intestine organogenesis involves a LIN-12/Notch signalling pathway. *Development* 127, 3429–3440
25 Delattre, M. and Felix, M.A. (2001) Development and evolution of a variable left–right asymmetry in nematodes: the handedness of P11/P12 migration. *Dev. Biol.* 232, 362–371

0168-9525/$ - see front matter
doi:10.1016/j.tig.2004.04.010

BlastDigester – a web-based program for efficient CAPS marker design

Katica Ilic*, Thomas Berleth and Nicholas J. Provart

Department of Botany, 25 Willcocks Street, University of Toronto, Toronto, ON M5S 3B2, Canada

Single nucleotide polymorphisms (SNPs) and insertion and deletions (InDels), the most common type of intra-specific sequence polymorphism, are rich resources for creating DNA markers for genotyping. In this article, we present a web-based program for generating cleaved amplified polymorphic sequence (CAPS) markers, PCR-based markers that use polymorphism in restriction endonuclease digestion patterns (snip-SNPs). BlastDigester computationally 'digests' the sequence alignments that are returned in a Blast search, identifies snip-SNPs and simplifies PCR-primer generation through dynamic links to Primer3, a primer design program.

DNA sequences from different varieties or accessions of a given species are becoming available through several sequencing projects. Single nucleotide polymorphisms (SNPs) together with insertion and deletions (InDels) are the most common type of polymorphism in the genomes studied to date. Large sets of predicted SNPs are publicly available for the human genome (SNP consortium, http://snp.cshl.org) and for some genetic model organisms including *Caenorhabditis elegans* [1], *Drosophila melanogaster* [2] and *Arabidopsis thaliana* [3]. Approximately 30–40% of SNPs alter restriction endonuclease recognition sites and these are commonly referred to as snip-SNPs. Restriction enzyme digestion patterns that are polymorphic might be used to create cleaved amplified polymorphic sequence (CAPS) markers, which are codominant molecular markers that amplify a short genomic sequence around the polymorphic endonuclease restriction site [4]. They are detected easily by agarose gel electrophoresis. The use of CAPS is thus affordable and practical for genotyping in positional or map-based cloning projects [4–6].

With the increasing availability of parallel genomic sequences from different varieties or accessions of genetic model species, there is a need for a web-based, user-friendly program that facilitates snip-SNP-based CAPS marker design. Currently, there is no free software tool available for the rapid detection of the restriction site polymorphisms in aligned sequences. Some related applications are used for conceptually different markers (SNAPER [7]) or are not freely available for bench scientists through a web interface (autoSNP [8] and SNP2CAPS [9]), have a limited sequence size input and no sequence alignment option (dCAPS Finder [10,11]) or do not integrate primer design (autoSNP [8] and SNP2CAPS [9]).

To overcome these limitations we created BlastDigester, a web-based program intended for bench scientists, which identifies differential endonuclease restriction sites in the pairwise sequence alignments of a standard Blast output. Dynamic links to Primer3 enable rapid CAPS marker design with the ease of a

* Current address: The *Arabidopsis* Information Resource, Carnegie Institution of Washington, Department of Plant Biology, 260 Panama St, Stanford, CA 94305, USA

Corresponding author: Nicholas J. Provart (provart@botany.utoronto.ca).
Available online 28 May 2004

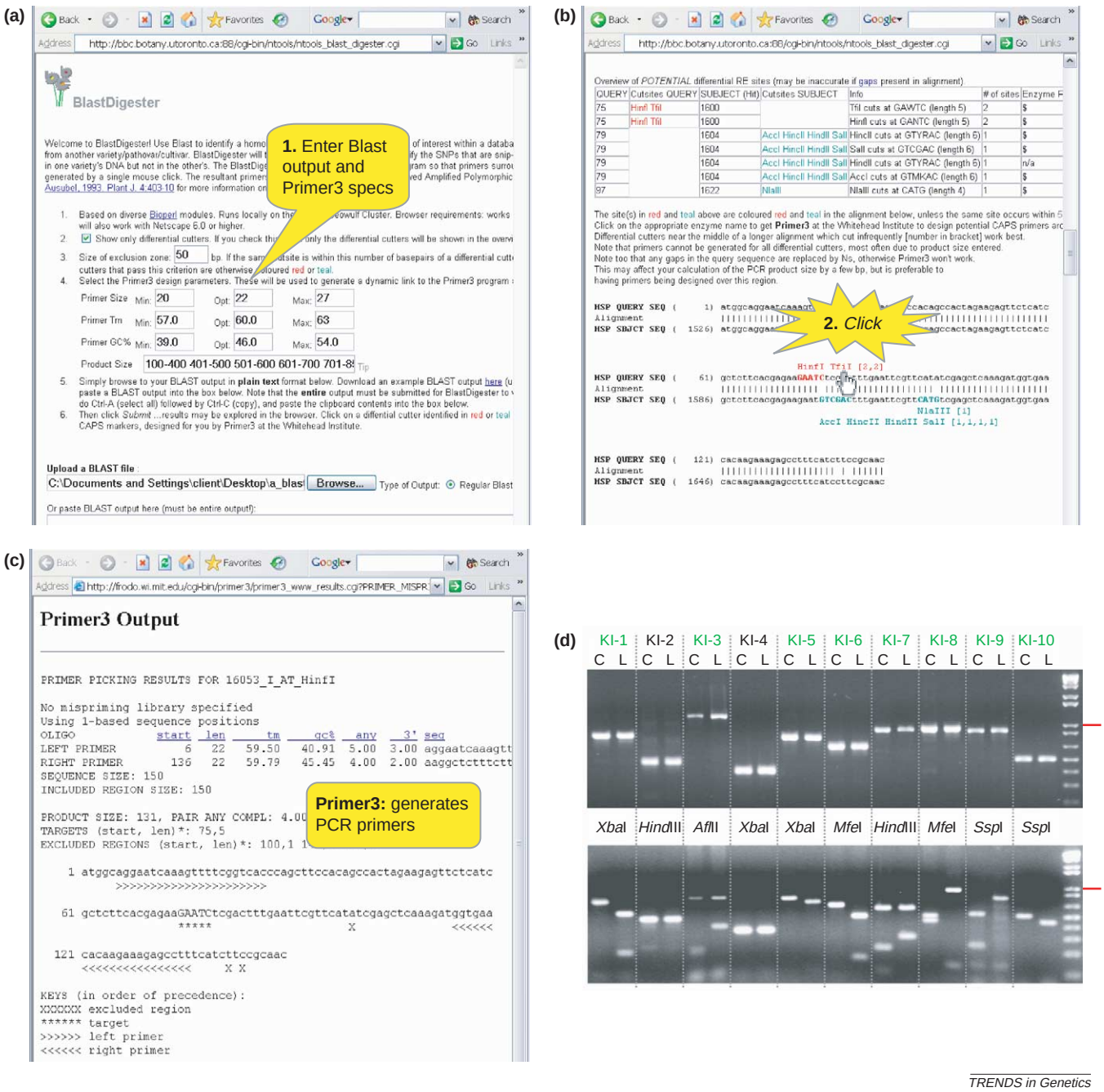

Figure 1. Generating cleaved amplified polymorphic sequence (CAPS) markers with BlastDigester. **(a)** Step one: enter the output of a Blast search and the Primer3 specifications into the BlastDigester page. **(b)** Step 2: click on a desired differential cutter identified by the BlastDigester program. **(c)** The user is directed to the Primer3 program, which generates oligonucleotides suitable for PCR. **(d)** Agarose gel electrophoresis results of ten CAPS primers that were designed based on BLAST alignments. The top and bottom image in this panel show PCR products before and after restriction digestion, respectively. The restriction endonuclease used for cleavage is indicated between the two gel images. The lane on the right-hand side contains the DNA ladder and the red line indicates the 1kb fragment. Abbreviations: C, Columbia-0 genomic DNA; L, Landsberg *erecta* genomic DNA.

mouse click. Using a sample dataset, we show that the software tool is experimentally robust, even with lower quality sequence at 1 × coverage.

Creating CAPS markers with BlastDigester

Using BlastDigester, CAPS markers can be generated in a simple, three-step procedure (Figure 1).

The input for the BlastDigester is the local sequence alignments found in the output of Blastn [12] or Blast2Seq [13]. There are two parts for each sequence alignment in the Blast output: (i) a table with the restriction enzymes listed for both sequences; and (ii) the aligned sequence with the exact position of the differential restriction sites (Figure 1b). Clicking on a selected polymorphic restriction site activates the Primer3 program hosted by the Whitehead Institute (http://www-genome.wi.mit.edu/genome_software/other/primer3.html) and primers flanking a specified snip-SNP site are automatically generated (Figure 1c). Several parameters for primer design can be defined on the input page of BlastDigester (i.e. primer T_m, GC content,

Box 1. The pipeline for CAPS marker generation using BlastDigester

- Selected coding or non-coding, non-repetitive sequences from two accessions are aligned with Blast. Selection depends on the level of polymorphism between the two sequences. The Blastn or Blast2Seq algorithms can be used for alignment. It is important to use a web-based Blast server that offers the option of saving or displaying the Blast output as plain text, although copied output from NCBI's Blast2Seq server will also work with the BlastDigester program.
- Each Blast output in plain text format is uploaded in its entirety to the BlastDigester website and the parameters for primer design are defined (i.e. primer T_m, GC content, primer size and product size range). For reasons relating to PCR parameters and gel electrophoresis, we recommend that the PCR product size be between 400–1000 bp. It is also possible to define an exclusion interval (set as default to 50 bp). Potential snip-SNP restriction enzymes (SNPs that alter restriction endonuclease recognition sites) that also cut elsewhere in the sequence within this interval will be flagged in gray in the output.
- The outputs of the BlastDigester program are examined for appropriate snip-SNPs. BlastDigester counts the number of times a potential snip-SNP enzyme cuts elsewhere in the sequence and displays this information, both in tabular format and in an alignment view, therefore, restriction enzymes that cut less frequently or cut uniquely can be chosen. In addition, if the potential snip-SNP restriction enzyme also cuts within the exclusion interval defined on the input page, it is grayed-out. Otherwise, potential snip-SNPs are coloured red or teal and are hyperlinked to Primer3. Clicking on a snip-SNP transfers the appropriate information for primer design to Primer3.
- After clicking on an appropriate snip-SNP restriction enzyme to activate Primer3, the cleaved amplified polymorphic sequence (CAPS) primers are obtained from the Primer3 output page. The dynamic link from BlastDigester contains information that restricts CAPS primer design to regions that are identical in both sequences (i.e. primer design is prohibited in regions of mismatch).

primer size and product size range; Figure 1a). To generate a sufficient fragment size difference between candidate markers, an exclusion window can be defined: if the potential snip-SNP restriction enzyme also cuts within the exclusion zone, it is 'grayed-out' in the output, eliminating 'imperfect' CAPS markers (Box 1). In cases where two sequences have a higher rate of polymorphism, the BlastDigester is designed to restrict primer selection by Primer3 to regions that do not have nucleotide mismatch (Figure 1c).

BlastDigester was implemented in Perl using several of the packages available in the Bioperl distribution (http://www.bioperl.org) [14]. The program is offered freely to non-commercial users through the Internet (http://bbc.botany.utoronto.ca:88). The source code for BlastDigester is also available upon request for local installation and supports a command line option.

Generating CAPS markers in *Arabidopsis* with BlastDigester

To test the efficacy of BlastDigester on a sample dataset, ten candidate CAPS markers were identified using BlastDigester for the *Arabidopsis* accessions Columbia-0 (Col-0) and Landsberg *erecta* (L*er*), which are found between the markers nga129 and m558A on chromosome 5. The Col-0 sequence is complete and is >99.99% accurate [15], whereas the Cereon *Arabidopsis* L*er* sequence collection consists of ~80 000 short contigs and sequences at ~1 × coverage [16]. The DNA sequences of Col-0 and L*er* are available at the TAIR website (http://www.arabidopsis.org) [17]. Selected 3′ untranslated regions (UTRs) and intergenic regions up to 5 kb in length of the Col-0 sequence were used to search for homologous L*er* contigs using Blast. A given Blast output was then analyzed with BlastDigester, which identified polymorphic restriction enzyme sites in the aligned sequences. PCR primers flanking selected sites were generated by linking to Primer3. The candidate CAPS-PCR products were digested with the appropriate enzymes and the results are shown in Figure 1d (the positive controls for restriction enzyme digestions are not shown). Eight out of the ten markers tested yielded differential restriction enzyme digestion patterns. The failed markers were due to errors in the L*er* sequence and were confirmed by sequencing the PCR products (data not shown). Primer sequences for the ten CAPS markers used were deposited in the TAIR database and a list of the CAPS markers with TAIR accession numbers is available in Table 1 of the supplementary data on the BlastDigester website (http://bbc.botany.utoronto.ca:88).

The specific example of CAPS marker design using *Arabidopsis* genomic sequence data that we present here demonstrates the usefulness and intuitiveness of the BlastDigester program. However, this tool is applicable to snip-SNP analysis in any organism where parallel sequence data of two varieties (or accessions) are available.

General considerations for generating CAPS markers with BlastDigester

BlastDigester integrates efficiently the three main components that are necessary for CAPS marker generation: (i) a local sequence alignment; (ii) the search for existing differential restriction endonuclease enzyme cutsites; and (iii) primer design.

The input size of the aligned sequence is not limited in the current version of BlastDigester. However, the program is not designed for whole-genome high-throughput snip-SNP analysis because of the nature of the Blast algorithm itself. Instead, appropriately selected, shorter DNA sequence segments (up to several kilobases) are recommended for generating Blast alignments. The identification of polymorphic restriction enzyme cutsites by BlastDigester depends on the frequency, distribution and the types of existing intra-specific sequence variation. Because SNPs and InDels are not distributed evenly throughout a given genome, the selection of the appropriate genomic segment of the reference sequence to be used for marker development is important. For instance, in the *Arabidopsis* Col-0 and L*er* accessions, exons have low levels of polymorphism; therefore, 3′ UTRs and intergenic regions are ideal locations to create CAPS markers. Logically, exon regions might also be used if the necessary polymorphisms are present. For genomes such as maize, where intergenic regions contain blocks of repetitive DNA, untranslated genic regions (introns) could represent potential spots for snip-SNP mining and CAPS marker design.

The success of BlastDigester depends on the accuracy of the sequences used for the alignments. As seen in the *Arabidopsis* example, sequence errors in the L*er* ecotype

resulted in the generation of two 'false' CAPS. This can be bypassed by developing markers from higher quality sequence data (i.e. with higher levels of redundancy) while avoiding single-read sequences. In practice, however, an 80% marker success rate is more than acceptable because of the low cost of primer synthesis and primer testing and the abundance of snip-SNPs between the genomes of different accessions.

Finally, common sense must be followed when using BlastDigester. The sequence used for the input Blast should, if possible, be unique and not align with paralogous sequences in the genome. This will avoid problems during PCR amplification, namely the generation of multiple PCR products.

Concluding remarks

With its all-in-one design, this program will have immediate application in positional cloning in plant and animal species for which sequence information is available. Although most useful for gene mapping, BlastDigester will also be a useful tool in a broad range of genetic studies [e.g. genotyping, phylogenetics, quantitative trait loci (QTL) mapping and RT–PCR-based expression studies of multigene families]. Considering the large number of genome sequencing projects that are currently underway or in the planning stages, BlastDigester will have immediate appeal for diverse groups of bench scientists in the genetic and genomic communities.

Acknowledgements

We thank Katrien Devos and Sean Cutler for critically reading and commenting on this manuscript. This work was funded by NSERC and Genome Canada.

References

1 Wicks, S.R. *et al.* (2001) Rapid gene mapping in *Caenorhabditis elegans* using a high density polymorphism map. *Nat. Genet.* 28, 160–164
2 Hoskins, R.A. *et al.* (2001) Single nucleotide polymorphism markers for genetic mapping in *Drosophila melanogaster*. *Genome Res.* 11, 1100–1113
3 Torjek, O. *et al.* (2003) Establishment of a high-efficiency SNP-based framework marker set for *Arabidopsis*. *Plant J.* 36, 122–140
4 Konieczny, A. and Ausubel, F.M. (1993) A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-based markers. *Plant J.* 4, 403–410
5 Glazebrook, J. *et al.* (1998) Use of cleaved amplified polymorphic sequences (CAPS) as genetic markers in *Arabidopsis* thaliana. *Methods Mol. Biol.* 82, 173–182
6 Lukowitz, W. *et al.* (2000) Positional cloning in *Arabidopsis*. Why it feels good to have a genome initiative working for you. *Plant Physiol.* 123, 795–805
7 Drenkard, E. *et al.* (2000) A simple procedure for the analysis of single nucleotide polymorphisms facilitates map-based cloning in *Arabidopsis*. *Plant Physiol.* 124, 1483–1492
8 Barker, G. *et al.* (2003) Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics* 19, 421–422
9 Thiel, T. *et al.* (2004) SNP2CAPS: a SNP and INDEL analysis tool for CAPS marker development. *Nucleic Acids Res.* 32, e5
10 Neff, M.M. *et al.* (1998) dCAPS, a simple technique for the genetic analysis of single nucleotide polymorphisms: experimental applications in *Arabidopsis thaliana* genetics. *Plant J.* 14, 387–392
11 Neff, M.M. *et al.* (2002) Web-based primer design for single nucleotide polymorphism analysis. *Trends Genet.* 18, 613–615
12 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
13 Tatusova, T.A. and Madden, T.L. (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* 174, 247–250
14 Stajich, J.E. *et al.* (2002) The bioperl toolkit: perl modules for the life sciences. *Genome Res.* 12, 1611–1618
15 Arabidopsis Genome Initiative, (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815
16 Jander, G. *et al.* (2002) *Arabidopsis* map-based cloning in the post-genome era. *Plant Physiol.* 129, 440–450
17 Rhee, S.Y. *et al.* (2003) The *Arabidopsis* information resource (TAIR). A model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.* 31, 224–228

0168-9525/$ - see front matter
doi:10.1016/j.tig.2004.04.012

Genome Analysis

Conservation of sequence and function in the *Pax6* regulatory elements

Richard Morgan

Department of Basic Medical Sciences, St George's Hospital Medical School, Cranmer Terrace, London SW17 0RE, UK

The Pax6 transcription factor is a highly conserved regulator of eye and brain development. Its complex mode of regulation is shared between *Drosophila melanogaster* and the vertebrates, despite there being no significant sequence identity between their enhancer and promoter regions. In this article, a more detailed examination of the *Pax6* genomic sequence reveals a common set of putative *Pax6* and basic helix–loop–helix transcription factor binding sites.

Several proteins that are involved in early developmental events are highly conserved throughout the Metazoans, a particularly striking example is paired box gene 6 (*Pax6*). This encodes a transcription factor with both a paired box and a homeodomain that is expressed in several different

Corresponding author: Richard Morgan (rmorgan@sghms.ac.uk).
Available online 6 May 2004