Review

Restriction Enzyme Mining for SNPs in Genomes

LI-YEH CHUANG1*, CHENG-HONG YANG2*, KE-HUNG TSUI3*, YU-HUEI CHENG2, PHEI-LANG CHANG3, CHENG-HAO WEN4 and HSUEH-WEI CHANG4,5,6

1Department of Chemical Engineering, I-Shou University, Kaohsiung; 2Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung; 3Department of Surgery, Division of Geriatric Urology, Chang Gung Bioinformatics Center, Chang Gung Memorial Hospital, Chang Gung University, Taoyuan; 4Faculty of Biomedical Science and Environmental Biology, 5Graduate Institute of Natural Products, College of Pharmacy, and 6Center of Excellence for Environmental Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan, R.O.C.

Abstract. Many different single nucleotide polymorphisms (SNPs) genotyping methods have been developed recently. However, most of them are expensive. Using restriction enzymes for SNP genotyping is a cost-effective method. However, restriction enzyme mining for SNPs in a genome sequence is still challenging for researchers who do not have a background in genomics and bioinformatics. In this review, the basic bioinformatics tools used for restriction enzyme mining for SNP genotyping are summarized and described. The objectives of this paper include: i) the introduction of SNPs, genotyping and PCR-restriction fragment length polymorphism (RFLP); ii) a review of components for genotyping software, including tools for primer design only or restriction enzyme mining only; iii) a review of software providing the flanking sequence for primer design; iv) recent advances in PCR-RFLP tools and natural and mutagenic PCR-RFLP; v) highlighting the strategy for restriction enzyme mining for SNP genotyping; vi) a discussion of potential problems for multiple PCR-RFLP. The different implications for restriction enzymes on sense and antisense strands are also discussed. Our PCR-RFLP freeware, SNP-RFLPing, is included in this review to illustrate many characteristics of PCR-RFLP software design. Future developments will include further sophistication of PCR-RFLP software in order to provide better visualization and a more interactive environment for SNP genotyping and to integrate the software with other tools used in association studies.

SNPs and Genotyping

Single nucleotide polymorphisms (SNPs) are essential for the studies of malignancies (1, 2) and other diseases (3, 4) and in preventive medicine (5, 6), personalized medicine (7), forensics (8) and evolution (9). Therefore, SNP genotyping methods and related software tools are required. Many SNP genotyping methods have been reviewed (10-12), including PCR-restriction fragment length polymorphism (RFLP) analysis, DNA sequencing, the Taqman probe, SNP genotyping with fluorescence polarization detection (13), Invader assay (14), DNA microarrays (15), MALDI-TOF (matrix assisted laser desorption/ionization-time of flight) (16), pyrosequencing (17), etc. But most of these methods are currently prohibitively expensive.

PCR-RFLP

PCR-RFLP is one of the most common choices for SNP genotyping in regular laboratories for genetic association studies. PCR can be used to amplify very small amounts of DNA, usually in two hours, to the levels required for RFLP analysis. Therefore, many samples can be analyzed in a relatively short time. RFLP is a technique in which individual SNPs may be differentiated by analysis of patterns derived from cleavage of their amplified DNA (Figure 1A). If a sample with different nucleotides in the same SNP site differs in the distance between sites of cleavage of a particular restriction endonuclease, the length of the
fragments produced will differ when the DNA is digested with a restriction enzyme. Restriction endonucleases are enzymes that cleave DNA molecules when specific nucleotide sequences are recognized. Enzyme recognition sites are usually 4 to 6 base pairs in length. If molecules differ in their nucleotide sequence, as is the case with SNPs, fragments of different sizes may be generated. The fragments can be separated by gel electrophoresis. Accordingly, two sets of information are required for PCR-RFLP, namely the primers for the PCR reaction and restriction enzymes for RFLP. Both are described below.

Components of Genotyping Software

Primer design only or restriction enzyme mining only. Many kinds of primer design software, e.g., Primer 3 (18), PCR suite (19), Primer Design Assistant (PDA) (20), and MultiPrimer (21), have been developed. However, they do not provide the information needed for restriction enzymes for RFLP in SNP genotyping. In contrast, some databases or software, e.g., NEBCutter – a program to cleave DNA with restriction enzymes (22) and REBASE - restriction enzymes and DNA methyltransferases) (23), provide the restriction enzyme information for sequences without the primer design function. Actually, NEBCutter (22) provides the restriction enzyme information for input sequences using the REBASE database (23). However, it is not suitable for SNP related sequences for several reasons. Firstly, all the information for the restriction enzymes within a sequence and for RFLP genotyping, only the recognition sites containing the SNP are used for the experiment. Secondly, a SNP containing sequence, in which the SNP is always presented as an alternative nucleotide in the format of [dNTP1/dNTP2] or IUPAC (The International Union of Pure and Applied Chemistry) code, e.g., [A/G] or R.

SNP information software – flanking sequence for primer design. When the primer design function is coupled with the restriction enzyme database as described above, the source of the SNP containing sequence is still disputed. The SNP containing sequence, SNP_fasta, for each SNP is provided by The National Center for Biotechnology Information (NCBI) dbSNP (24). One Reference cluster ID (rs#) is frequently associated with many sequences of the NCBI assay ID (ss#) (http://www.ncbi.nlm.nih.gov/projects/SNP/). The submission (ss#) has the longest flanking sequence of all cluster members and is used to instantiate sequence for rs17886268 during BLAST analysis for the current build of dbSNP (24). For example, the rs17886268 contains the three events of NCBI assay ID ss32469404, ss35451533 and ss48294403 (http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?rs=17886268). The ID ss35451533 has the longest flanking sequence and is the best choice in NCBI dbSNP for a primer design template if needed. However, sometimes the longest flanking sequence in certain ss# is still too short to use for primer design. For example, the longest flanking sequence of rs8152226 was only 101bp in ss12543815 (http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?rs=8152226). Accordingly, software for providing SNP flanking sequences is required for primer design in PCR-RFLP. Some kinds of software, such as GeneWindow (25), FESD – a Functional Element SNPs Database in human (26) and our developed tool, SNPflanking as shown in Table I, can provide the SNP flanking sequence with an adjustable length of at least 1000bp. This long flanking sequence is suitable for primer design especially if the original flanking sequence is rather short like in rs8152226. Unfortunately, these tools cannot provide the necessary restriction enzyme information and primer design for PCR-PFLP.

PCR-RFLP Tools

Recently, some tools have been constructed by the combination of primer design and restriction enzyme mining from sequences and some of these tools are shown in Table II. PIRA PCR designer for restriction analysis of single nucleotide polymorphisms (27), in silico restriction software (28), SNPPicker (29), SNPcutter (30), and SNP-RFLPing (31) are all designed to yield a mutagenic primer (see below) and the restriction enzyme. Among these tools, only in silico software, SNP cutter and SNP-RFLPing provide the restriction enzymes for target SNPs within the input sequence. SNP-RFLPing provides the hyperlink to Primer 3 (18) for a regular primer (in contrast to a mutagenic primer) and SNP cutter (30) provides the plug-in Primer 3 function. In addition, SNP cutter supports online input, but output only via email, whereas in SNP-RFLPing, the input and output data are entirely provided online. SNP-RFLPing also accepts the input of SNPs – in IUPAC sequences or [dNTP1/dNTP2] format, as well as multiple mixed forms. Gene name, gene ID, and SNP ID rs# and ss# are also supported input formats.

Principle of PCR-RFLP – Natural and Mutagenic RFLP

PCR-RFLP includes two kinds of genotyping, i.e. RFLP with natural and mutagenic primers. The restriction enzyme used in natural PCR-RFLP (Figure 1A) only presents its recognition sequences containing the SNP. No other recognition sequences are allowed in the sequence of the PCR product. If they were allowed, the resulting RFLP pattern would be more complex, and thus make genotyping more difficult. When the length of the F (forward) primer to the SNP site is different from the length of the R (reverse) primer to the SNP site, there exist two cut bands in AG and GG in the example of rs8144801. If the length is equal, only
one cut band is observed. The PCR product of the uncut band retains its full length without enzyme digestion.

In Figure 1B, mutagenic RFLP is illustrated. Originally, for the sequence of the PCR product with only one nucleotide of the SNP a distinct restriction enzyme cannot be found. When the nucleotide of the sequence near the SNP is changed at a certain position, some restriction enzymes can be artificially created at only one of the alternative sequences for the SNP-containing sequence. The length difference between the cut and uncut bands is only limited by the length of the F primer minus the cut part at its 5’ end.

**Strategy for Restriction Enzyme Mining in SNP Genotyping**

The strategy for mining the restriction enzymes in SNP genotyping for both natural and mutagenic RFLP from the input sequence which we have developed, SNP-RFLPing (31), consists of several steps. First, all the restriction enzymes for each sequence with an alternative nucleotide of the same SNP are retrieved from a database such as REBASE (23) and listed. As an example, the rat SNP rs8144801, shown in Figure 1A and Figure 2A, and the rat SNP rs8143194 (Figure 2C) have an A/G genotype. This means that the two SNP-containing sequences are different only in the site of the SNP (A or G). Second, the system is designed to select the unique restriction enzymes in sequence (+) of Figure 2A and Figure 2C. The “+” symbol indicates the sense strand while the “−” symbol indicates the antisense strand, both of which are discussed in further detail below. The “0” and “1” symbols of Figure 2 represent the alternative nucleotides of the same SNP. Identical restriction enzymes are programmed to be omitted and therefore only distinguishable restriction enzymes around the SNP site show up in each sequence (two sequences for an alternative nucleotide of the same SNP) similar to “0” and “1” in Figure 2A and Figure 2C. The length of the input-flanking sequence does not influence the mining process. The reason for this is that the sequences outside the recognition site with SNP have the same enzymes that are filtered out by system.

**Sense and antisense strands have different implications for the restriction enzyme.** It is sometimes essential to transform the input (sense) sequence into an antisense sequence.
Although most endonucleases are panlindromic, the flanking sequences of the recognition site (5'→3' orientation) containing the SNP in the sense and antisense strands is different in Figure 2A and Figure 2C. Therefore, some recognition sites will be found exclusively only in the sense or antisense strands. For example, the rat SNP rs8144801 (the example used in Figure 1A and Figure 2A) is represented as the sequence TGGATGA CG GG CG GG CAC TGG GG C A [A/G]CATAG CCAAAGGGG C CT ACACAC TC. In Figure 2B, the sense (+) strand does provide a RF L P restriction enzyme for this SNP, however, the anti-sense (–) strand does not provide any recognition site for discriminating the SNP by RFLP.

Accordingly, the SNP containing sequences are programmed to be transformed into four separate sequences. The recognition site for BbvI, BseXI and BstVII is 5'-GCAGC-3' (Figure 2B). Please note that the sequence of the recognition site is directional (5'→3' orientation) for restriction enzymes and the same sequence with inverse orientation cannot be digested by the same restriction enzymes. The recognition site for BbvI, BseXI and BstVII typically should be written in the following format.

5'-GCAGC-3' (sense strand)
3'-CGTCG-5' (antisense strand)

However, the inverse orientation of the sense strand is 3'-GCAGC-5' (equal to 5'-GCAGC-3') but it is not the recognition site for BbvI, BseXI, and BstVII. Similarly, while 5'-GCAGC-3' is the recognition site for BbvI, BseXI and BstVII the anti-sense sequence 3'-CGTCG-5' (5'-GCTGC-3') is again not a recognition site. This is because the flanking sequence of the SNP is usually different between sense and anti-sense strands and the sequence of the recognition site is directional. In this case, both strands are required to be analyzed for mining restriction enzymes.

Sometimes, the sense and antisense strands provide the same restriction enzyme for RFLP genotyping. This is because the same recognition site is chosen in both strands (the recognition sites in both strands are complementary). In the example of Figure 2C and Figure 2D, the rat SNP rs8143194 is represented by the sequence TGGATGA CG GG CG GG CAC TGG GG C A [A/G]CATAG CCAAAGGGG C CT ACACAC TC. Its recognition site is 5'-GCAGC-3' for AspLEI, BstHHI, CfoI, Hhal, Hin6I, HinPII, and HspAI. This is the same sequence and orientation in both the sense (5'-GCAGC-3') and the anti-sense strand (3'-CGTCG-5', which is equivalent to 5'-GCTGC-3'). The sequence of the recognition site is complementary, too. Therefore, it does not
matter whether the sequence has been transformed into an anti-sense strand or not. In this example, the RFLP result is shown in one of the strands. The illustrated examples (rs8144801 and rs8143194) show that the SNP-RFLPing system (31) is designed to analyze four sequences of one SNP (Figure 2A and Figure 2B).

Figure 2. Restriction enzyme status for both sense and antisense strands. In both, the rat SNP rs8144801 (A, B) and SNP rs8143194 (C, D), the genotype is A/G in the sense strand (marked with “+”) and T/C in the antisense strand (marked with “−”). Part of the SNP flanking sequence for SNP rs8144801 and rs8143194 is shown in (A) and (C), respectively. The restriction enzyme information for RFLPs mined from (A, C) are shown in (B, D), respectively. The underlined section in both (A) and (C) indicates the recognition site of suitable restriction enzymes. The numbers “0” and “1” shown in (A, B, C and D) represent the alternative nucleotide of the same SNP site. Please note that the SNP flanking sequence in both sense and antisense strands are usually different. Sometimes its corresponding restriction enzymes are the same in both strands (C, D) and sometimes they are different (A, B).
SNP-RFLP for special types of SNP. In the SNP-RFLPing system, the SNP containing sequence in the format “A”, “T”, “G”, “C” should be accepted as it is contained in the developed program. Other ambiguous letters are recognised according to the IUPAC system. Upper and lower case is not significant and all other characters, including spaces and digits, are ignored.

For true SNPs (that is the change of a single nucleotide), the SNP-RFLPing software can provide highly confident information for RFLP genotyping. For indel (insertion/deletion) SNPs like rs10592063 (-/GGGA), the two recognition sites “ACTGGG” and “ACTGG” are recognized by BfiI, Bmrl, BseII, BseNI, BsrI and BsrSI in the sequence with GGGA, while in the sequence of rs10592063 without GGGA, the recognition site of CCTC is recognized by MnlI. According, this example explains that the central sequence for restriction enzyme mining for SNP genotyping is the sequence covering the SNP within the recognition site of the restriction enzyme. Therefore, SNP-RFLPing (31) is capable of dealing with true SNPs and non-true SNP variations (such as indel).

Like any other indel SNP, it is possible that one of the recombinated sequences may create the same restriction enzyme as in the sequence without the insertion (or with the deletion). However, it still is still highly probable that the RFLP enzyme can be found because the RFLP availability is dependent on subtracting the enzymes between sequences with alternative SNPs. When the same restriction enzyme is found, it is ignored. Only different enzymes for the target SNP will be shown. Of course, it could happen that no different RFLP enzymes can be provided and that the system indicates that the SNP cannot be distinguished. Basically, the success rate is very high except for cases where no available RFLP can be found. According to the subtracting strategy, mining errors are very small in SNP-RFLPing (31).

Potential Problems for Multiple RFLPing

For multiple RFLP genotyping of gene or genes, users can design suitable primers for PCR to extend the length of the product in order to cover the SNPs of interest. However, several problems are of concern: First, recognition sites for the restriction enzymes may be found outside of the SNP neighborhood. Then, the lengths and fragments of the RFLP products will change and it becomes more difficult to determine their SNP genotype. Second, different restriction enzymes may not be able to treat the PCR products at the same time due to different buffer systems. Third, it is currently accepted that there is about one SNP per 1000~1500bp. Sometimes, the distances between the SNPs are too long for multiple genotyping if only one primer pair is used. Accordingly, users have to perform restriction enzyme digestion individually when multiple RFLP genotyping is performed. In this case we suggest that the user performs single RFLP genotyping using the current software version of SNP-RFLPing (31).

Summary and Future Developments

In this paper, the basic background for restriction enzyme mining for SNP genotyping and the progress of PCR-RFLP related software. The development of PCR-RFLP software using our freeware SNP-RFLPing (31) as an example has been described in detail. Three steps are required for the construction of SNP-RFLPing. First, all the restriction enzymes for four sequences, including the two alternative nucleotides for both sense and antisense strands, belonging to one SNP are retrieved from a database like REBASE (23) and listed. Second, the unique restriction enzymes are programmed to substrate the same kind of enzymes existing in both sequences with one alternative nucleotide for both sense and antisense stands. This is the case in natural RFLP. Third, in cases where the natural restriction enzymes cannot be mined because no suitable restriction enzyme can distinguish the SNP site, the system changes a nearby nucleotide of the SNP by default to create an artificial recognition site and thereby design the mutagenic primer.

NCBI dbSNP (24) provides a comprehensive SNP database for many species and is a powerful support tool for computer scientists in the construction of useful algorithms in bioinformatics and genomics. Many different kinds of software have been designed for SNP genotyping using information from NCBI dbSNP. In Table I, FESD (26) and our unpublished software, SNPflanking, for example, provide a long distance for the SNP flanking sequence used in the primer design and will be included in the next version of SNP-RFLPing (31). We have also downloaded database information from Entrez Gene (32) which provides gene-centered information at NCBI. Using the gene information, the official HUGO (Human Genome Organisation) gene name and gene ID are cross-matched to our local SNP database obtained from NCBI dbSNP. Accordingly, SNP-RFLPing provides an innovative function for gene-centric RFLP restriction enzyme mining for SNPs within the input gene. Three kinds of user-friendly input formats are incorporated in the interface. NCBI dbSNP "rs" or "ss" IDs, NCBI Entrez gene ID and HUGO gene name and any format of SNP-in-sequence, are all compatible with SNP-RFLPing.

When a large number of SNPs are genotyped, the statistical analysis should be considered. The Hardy-Weinberg equilibrium, linkage disequilibrium (LD) and haplotype analysis are essential for association studies. Recently, JLIN (33) has developed the LD analysis tool, HelixTree™ (Golden Helix Inc, Montana, USA) packages providing both Hardy-Weinberg equilibrium and LD analysis and Haploview (34) has provided haplotype-based analysis and presentation for SNPs. The updated version of SNP-RFLPing will integrate these software tools, which will lead to a well-rounded and more comprehensive software product for association studies. Further SNP data for many other
species should also be incorporated in order to provide information for evolutionary studies.

In conclusion, SNP-RFLPing is potentially a very useful software package for association studies and evolutionary analysis. Besides input/output improvements over existing software tools, its GUI (Graphical User Interface) makes this software very easy to use. The capability of producing RFLP information for any given known genes is a particularly powerful feature, which has been lacking so far.

**Acknowledgements**

This work was partly supported by the National Science Council in Taiwan under grants 96-2221-E-214-050-MY3, NSC96-2311-B037-002, NSC96-2622-E214-004-CC3, NSC95-2221-E214-087, NSC95-2622-E-151-015-CC3, NSC95-2622-E-214-004-CC3 and the funds KMU-EM-97-2.1a and KMU-EM-97-1.1b.

**References**


Received March 4, 2008
Revised May 2, 2008
Accepted May 9, 2008

2007