SNP-RFLPing 2

an updated and integrated PCR-RFLP tool for SNP genotyping

User Manual

Contact:
changhw@kmu.edu.tw
yuhuei.cheng@gmail.com
chuang@isu.edu.tw
chyang@cc.kuas.edu.tw

Updated: 2009/11/03
# Table of content

1. **Introduction**
   - 1.1 Purpose
   - 1.2 System workflow
   - 1.3 Function overview

2. **Function: SNP ID for input**
   - 2.1 Overview
   - 2.2 Example: A complete operation using SNP ID to analyze the SNP-RFLPing
     - 2.2.1 Example (1): SNP IDs input
     - 2.2.2 Example (2): Query SNPs information
     - 2.2.3 Example (3): SNP-RFLP analysis
     - 2.2.4 Example (4): Natural primers information acquiring
     - 2.2.5 Example (5): Available restriction enzymes request
     - 2.2.6 Example (6): Further restriction enzyme information description
     - 2.2.7 Example (7): Mutagenic primer information acquiring
     - 2.2.8 Example (8): TaqMan information providing

3. **Function: SNP in fasta sequence for input**
   - 3.1 Overview
   - 3.2 Example: Use SNP in fasta sequence format to analyze

4. **Function: Multiple SNPs within one sequence for input**
   - 4.1 Overview
   - 4.2 Example: Multiple SNPs within one sequence input

5. **Function: GenBank accession no. for input**
   - 5.1 Overview
   - 5.2 Example: Use HUGO gene name ‘TP53’ to analyze

6. **Function: TagSNP from HapMap**
   - 6.1 Overview
   - 6.2 Example: Use HUGO gene name ‘BRCA2’ to analyze

7. **Function: Transcript ID/miRNA**
   - 7.1 Overview
   - 7.2 Example: Use transcript ID and miRNA number to analyze

8. **Function: Gene Ontology-based annotation for SNPs**
   - 8.1 Overview
   - 8.2 Example: Query Gene Ontology vocabulary term

9. **Function: File upload for input**
   - 9.1 Overview
   - 9.2 Example: Upload interface for SNP IDs and SNP fasta sequences

10. **Appendix**
   - 10.1 Definitions
   - 10.2 Related links
1. Introduction

1.1 Background

PCR-restriction fragment length polymorphism (RFLP) assay is a cost-effective method for SNP genotyping and mutation detection, but the manual mining of restriction enzymes is challenging and cumbersome. Three years after we constructed SNP-RFLPing, a freely accessible analysis tool for restriction enzyme mining of SNPs, significant improvements over the 2006 version have been made in the latest version, SNP-RFLPing 2. The primary aim of SNP-RFLPing 2 is to provide a comprehensive PCR-RFLP information for SNPs in multiple fields, such as SNPs for multiple species, multiple kinds (di-, tri, tetra- and indel), gene-centric search, HapMap tagSNPs, gene ontology-based search, miRNAs, and SNP500Cancer. The RFLP restriction enzymes and their corresponding PCR primers of natural and mutagenic types for each SNP are simultaneously analyzed. All the prices of the RFLP restriction enzymes are provided for selection. Furthermore, the previously encountered updating problems for most SNP related databases are resolved by an online retrieval system. The user interfaces for the functional SNP analyses have been substantially improved and integrated. SNP-RFLPing 2 offers a new, user-friendly interface for RFLP genotyping in association studies and is freely available at http://bio.kuas.edu.tw/snp-rflping2/rflpUI.jsp.

1.2 System workflow

The workflow of SNP-RFLPing 2 was illustrated in Figure 1 (see next page). The system was divided into nine modules, including: 1) input module, 2) data retrieve module, 3) sequence process module, 4) SNP-RFLP module, 5) data query module, 6) remote database module, 7) primer design module, 8) output module, and 9) SNP-RFLP database module according to MVC (Model-View-Controller) architecture. Input and Output modules are served with the graphical user interface (GUI) to operate and analyze.
1.3 Function overview

SNP-RFLPing 2 provides *eight major functions* to assist and analyze the RFLP restriction enzymes and its corresponding primers for SNPs in all SNP-related fields, including:

1. **SNP ID input**
2. **SNP in fasta sequence input**
3. **Multiple SNPs within one sequence**
4. **GenBank accession**
5. **TagSNP from HapMap**
6. **Transcript ID/miRNA**
7. **Gene Ontology-based annotation for SNPs**
8. **File upload**
2. Function: SNP ID input

2.1 Overview

SNP ID (rs# and ss#) is acceptable to query the SNP-RFLP information. Single ID, multiple rs#, multiple ss#, and mixed rs# and ss# are available for SNP ID input. For multiple SNP IDs, three symbols which are “comma”, “space”, and “line” can be used to separate SNPs. The maximal input SNP IDs is limited to fifty.

2.2 Example: A complete operation using SNP ID to analyze the SNP-RFLPing

2.2.1 Example (1): SNP IDs input

Fig. 2-1 and Fig. 2-2 show a single ID input for rs# and ss#, respectively.

Fig. 2-3, Fig. 2-4 and Fig. 2-5 show the multiple rs# input with the “comma” symbol, “space”, and “line” (press the “Enter key” in keyboard) to separate, respectively.

Fig. 2-6, Fig. 2-7 and Fig. 2-8 show the multiple ss# input with comma symbol, space, and line to separate, respectively.

Fig. 2-9, Fig. 2-10 and Fig. 2-11 show the multiple rs# and ss# input with comma symbol, space, and line to separate, respectively.

Fig. 2-1. Single ID input with rs#.
Fig. 2-2. Single ID input with ss#.

Fig. 2-3. Multiple rs# input with the comma symbols to separate.
Fig. 2-4. Multiple rs# input with space to separate.

Fig. 2-5. Multiple rs# input with lines (press the enter key of keyboard) to separate.
Fig. 2-6. Multiple ss# input with the comma symbols to separate.

Fig. 2-7. Multiple ss# input with space to separate.
Fig. 2-8. Multiple ss# input with lines (press the enter key of keyboard) to separate.

Fig. 2-9. Multiple rs# and ss# input with the comma symbols to separate.
Fig. 2-10. Multiple rs# and ss# input with space to separate.

Fig. 2-11. Multiple rs# and ss# mixed input with lines (press the enter key of keyboard) to separate.
2.2.2 Example (2): Query SNPs information

After inputting SNP ID, users can click the “Query” button to query SNPs information. Users will see the waiting page (Fig. 2-12) for a moment and the result for SNP information will appear (Fig. 2-13). The result for SNP information includes: reference SNP ID (rs#), organism, short SNP flanking sequence (52 bp in total), chromosome information, clinical/LSDB associated, PubMed citing, SNP gene, sequence viewer, SNP 3D, OMIM, heterozygosity frequency, validated and genotype data, such as NCBI Entrez SNP. Furthermore, HGVS Names are provided for *Homo sapiens*.

![Fig. 2-12. Waiting for the SNP information query.](image)

<table>
<thead>
<tr>
<th>Query SNPs Results: 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>All selected</td>
</tr>
<tr>
<td><strong>rs11540652</strong> [Homo sapiens]</td>
</tr>
<tr>
<td>![Checkmark]</td>
</tr>
</tbody>
</table>

![Fig. 2-13. The result for SNP information.](image)
2.2.3 Example (3): SNP-RFLP analysis

Subsequently, user can check SNP for further analysis. In this example, a SNP rs11540652 is chosen to check (Fig. 2-13) and users can click the “Analyze sequence” button to execute SNP-RFLP function. When the “Analyze sequence” button is clicked, a waiting page will appear (Fig. 2-14).

![Fig. 2-14. Waiting for the SNP-RFLP recognition.](image)

After the waiting page is disappeared, the result for SNP-RFLP recognition is appearing (Fig. 2-15). The result information of SNP-RFLP includes SNP ID (rs#), short SNP flanking sequence (52 bp in total), result for sense strand and anti-sense strand whether recognized by restriction enzymes or not, TaqMan, primer set, and enzymes information. The SNP ID is hyperlinked to NCBI Entrez SNP for further information. In this SNP (rs11540652), its TaqMan probe is not provided in both NCBI and SNP500Cancer and the natural primer set (i.e., the primers are the regular primers and the RFLP restriction enzymes are available) is available.

![Fig. 2-15. The result for SNP-RFLP recognition. + is the sense strand and – is the antisense strand. Smile symbol means that the RFLP enzymes are available in the SNP-RFLPing 2 analysis.](image)
2.2.4 Example (4): Natural primers information acquiring

Clicking the icon of “Natural”, the natural primer set information is appearing (Fig. 2-16). The natural primer set information contains SNP ID, strand orientation, recognized allele, restriction enzymes, natural primer pair information, and natural primers visualization.

**SNP ID: rs11540652**

<table>
<thead>
<tr>
<th>Strand orientation</th>
<th>Recognized allele</th>
<th>Restriction enzymes</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>A</td>
<td>BspCI (Rindl) StuI</td>
</tr>
<tr>
<td>+</td>
<td>G</td>
<td>Asp742</td>
</tr>
<tr>
<td>-</td>
<td>T</td>
<td>Acanry</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>AcII</td>
</tr>
</tbody>
</table>

Fig. 2-16. The restriction enzymes and natural primer set information for rs11540652. Both forward and reverse primers are underlined with red color and the SNP is indicated by an arrow line. + is the sense strand and – is the antisense strand.
2.2.5 Example (5): Available restriction enzymes request

In order to further get full enzymes information, users can click the “Enzymes Information” button. Fig. 2-17 shows only the available “sense” strand restriction enzymes information. When the checkbox for anti-sense strand is checked, the restriction enzymes information for both the available “sense” and “anti-sense” strands are shown in Fig. 2-18.

<table>
<thead>
<tr>
<th>Sequence ID</th>
<th>Recognition Sequence</th>
<th>Commercial Suppliers</th>
<th>Commercial NEB Price</th>
<th>Non-Commercial</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ sequence</td>
<td>1: 5'-CCAGA-3'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2: 5'-GAAACA-3'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3: 5'-CAGAG-3'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- sequence</td>
<td>5' - 5'-CACTTCGTCATGGGCCGCGTGAACC-3'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5' - 5'-CACTTCGTCATGGGCCGCGTGAACC-3'</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 2-17. The available sense strand restriction enzymes information for rs11540652. + is the sense strand and – is the antisense strand. 0 and 1 indicate the alternative nucleotide for the same SNP.

<table>
<thead>
<tr>
<th>Sequence ID</th>
<th>Recognition Sequence</th>
<th>Commercial Suppliers</th>
<th>Commercial NEB Price</th>
<th>Non-Commercial</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ sequence</td>
<td>1: 5'-CCGGA-3'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2: 5'-GAAACA-3'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3: 5'-CAGAG-3'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- sequence</td>
<td>5' - 5'-CACTTCGTCATGGGCCGCGTGAACC-3'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5' - 5'-CACTTCGTCATGGGCCGCGTGAACC-3'</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 2-18. The available sense and anti-sense strand restriction enzyme information for rs11540652 when the checkbox for anti-sense strand (indicated by a red arrow line) is checked. + is the sense strand and – is the antisense strand. 0 and 1 indicate the alternative nucleotide for the same SNP.
### 2.2.6 Example (6): Further restriction enzyme information description

The enzymes information contains: (1) “Sequence ID” for sense and anti-sense strands, (2) “Recognition Sequence”, and (3) “Commercial” and “Non-Commercial” enzymes.

For “Recognition Sequence”, this system provides the hyperlink for the same recognition sequence (Fig. 2-19). When the hyperlink for “Enzyme Name” is clicked, the search results for this enzyme name in NEW ENGLANE BioLabs (NEB) will be output (Fig. 2-20).

For “Commercial” enzymes, two options are available: one is ‘Suppliers’ and the other is ‘NEB Price’. When users select the ‘Suppliers’, the following enzyme hyperlinks will link to REBASE Suppliers (Fig. 2-21). When users select the ‘NEB Price’, the following enzymes hyperlink will link to price manager for Restriction Endonucleases which is our locally built-in database, all enzyme prices come from the price for NEB (Fig. 2-22 and Fig. 2-23). For “Non-Commercial” enzymes, enzymes hyperlink will be link to NCBI PubMed to search related documents (Fig. 2-24).

<table>
<thead>
<tr>
<th>ID</th>
<th>Enzyme Name</th>
<th>Microorganism</th>
<th>Source</th>
<th>Recognition Sequence</th>
<th>Methylation Site</th>
<th>Commercial Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alv261</td>
<td>Acinetobacter lwoffi RFL26</td>
<td>A. Janulaitis</td>
<td>GTTCG(15)</td>
<td>3(6), 4(6)</td>
<td>Fermentas International Inc</td>
</tr>
<tr>
<td>2</td>
<td>BglII</td>
<td>Bacteroides coprocola DSM 17136</td>
<td>DSM 17136</td>
<td>GTTCG(15)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>BscII</td>
<td>Bacillus species 4304</td>
<td>D. Clark</td>
<td>GTTCG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>BsmAI</td>
<td>Bacillus steinfieldeae DSM 17136</td>
<td>Z. Chen</td>
<td>GTTCG(15)</td>
<td>7(6)</td>
<td>New England Biolabs</td>
</tr>
<tr>
<td>5</td>
<td>BsmAI</td>
<td>Bacillus steinfieldeae MA</td>
<td>Sh. Degtyarev</td>
<td>GTTCG(15)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>BsmAI</td>
<td>Bacillus steinfieldeae MA</td>
<td>Sh. Degtyarev</td>
<td>GTTCG(15)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 2-19. The different enzyme names with the same recognition sequence ‘GTCTC’.
Fig. 2-20. The search results for Alw26I in NEW ENGLANE BioLabs.

Fig. 2-21. “Commercial” enzymes option is ‘Suppliers’. Clicking the BsiSI enzyme will link to REBASE Suppliers and show the suppliers for BsiSI.

Fig. 2-22. For enzyme BsiSI, no any price is provided.

Fig. 2-23. For enzyme HpaII, the price is provided from NEB.
Fig. 2.24. Clicking the non-Commercial enzymes will be sent to NCBI PubMed to search related documents.
2.2.7 Example (7): Mutagenic primer information acquiring

Some SNPs are not recognized by any restriction enzymes. Alternatively, the mutagenic primer may be designed by the SNP-RFLPing 2 system. When a SNP does not have any restriction enzymes, the icon of “Mutagenic” will appear after SNP-RFLP analysis (Fig. 2-25). Clicking the icon of “Mutagenic”, the mutagenic primer set information is appearing (Fig. 2-26). The mutagenic primer set information contains SNP ID, strand orientation, recognized allele, restriction enzymes, mutagenic primer pair information, and mutagenic primer visualization.

![Figure 2-25](image)

Fig. 2-25. No any available restriction enzymes for SNP rs45541837, the icon of “Mutagenic” appears in the result of SNP-RFLP analysis.

![Figure 2-26](image)

Fig. 2-26. The restriction enzymes and natural primer set information for rs45541837. The mutagenic position is underlined with green color in the F of primer pair field. Both forward and reverse primers are underlined with red color and the SNP is indicated by an arrow line. + is the sense strand and – is the antisense strand.
2.2.8 Example (8): TaqMan information providing

In above SNP (rs11540652), its TaqMan probe is not provided in both NCBI and SNP500Cancer (Fig. 2-15). However, some SNPs, such as rs1318703 which is shown on SNP ID example (2) (Fig. 2-27) and rs1202183 which is shown on SNP ID example (4) - ss1717832 (Fig. 2-28) provide TaqMan probe. When a SNP provide TaqMan probe, the icon of “TaqMan” will appear in the result of SNP-RFLP analysis.

![Fig. 2-27. The icon of TaqMan appears in rs1318703 which is in SNP ID example (2).](image1)

![Fig. 2-28. The icon of TaqMan appears in rs1202183 which is in SNP ID example (4) - ss1717832.](image2)

User can click the icon of “TaqMan”, and then TaqMan information for this SNP will be represented. Fig. 2-29 and Fig. 2-30 show the available TaqMan information for rs1318703 and rs1202183, respectively.

![Fig. 2-29. The TaqMan information for rs1318703 is shown after clicking the icon of “TaqMan” in Fig. 2-27.](image3)
Fig. 2-30. The TaqMan information for rs1202183 is shown after clicking the icon of “TaqMan” in Fig. 2-28.
3. Function: SNP in fasta sequence input

3.1 Overview

SNP in fasta sequence format is acceptable to query the SNP-RFLP information in SNP-RFLPing 2. Single fasta sequence, multiple fasta sequences with SNPs in [dNTP1/dNTP2] or IUPAC formats, and special sequences with SNP in the indel (insertion and deletion), tri-allelic and tetra-allelic formats are all acceptable.

Fig. 3-1 shows a regular single fasta sequence.

Fig. 3-2 shows a single fasta sequence with line (press the enter key of keyboard) which is ignored in the system.

Fig. 3-3 and Fig. 3-5 show regular multiple fasta sequences with SNPs in [dNTP1/dNTP2] and IUPAC format, respectively.

Fig. 3-4 shows multiple fasta sequences with SNPs in [dNTP1/dNTP2].

Fig. 3-6 shows the IUPAC format with lines (press the enter key of keyboard) which are ignored.

Fig. 3-7 to Fig. 3-9 are special sequence with SNP in the in-del, tri-allelic and tetra-allelic formats, respectively. In the SNP fasta sequence input, the uppercase or lowercase letters are acceptable.

3.2 Example: Use SNP in fasta sequence format to analyze

Fig. 3-1. SNP in a regular single fasta sequence.
Fig. 3-2. SNP in a single fasta sequence with line (press the enter key of keyboard) which is ignored in the system. When the sequence is too long, the part of the sequence will list in the next line automatically. The system is able to regard them as the single sequence input.

Fig. 3-3. Regular multiple fasta sequences with SNPs in [dNTP1/dNTP2] format.
Fig. 3-4. Multiple fasta sequences with SNPs in [dNTP1/dNTP2] format with line (press the enter key of keyboard) which are ignored in the system. When the sequence is too long, the part of the sequence will list in the next line automatically. The system is able to regard them as the single sequence input.

Fig. 3-5. Regular multiple fasta sequences with SNPs in IUPAC format.
Fig. 3-6. Multiple fasta sequences with SNPs in IUPAC format with line (press the enter key of keyboard) which is ignored in the system. R = G or A; S = G or C. When the sequence is too long, the part of the sequence will list in the next line automatically. The system is able to regard them as the single sequence input.

Fig. 3-7. Special sequence with SNP in the in-del format. When the sequence is too long, the part of the sequence will list in the next line automatically. The system is able to regard them as the single sequence input.
**Fig. 3-8. Special sequence with SNP in the tri-allelic format.** When the sequence is too long, the part of the sequence will list in the next line automatically. The system is able to regard them as the single sequence input.

**Fig. 3-9. Special sequence with SNP in the tetra-allelic format.** When the sequence is too long, the part of the sequence will list in the next line automatically. The system is able to regard them as the single sequence input.
After inputting the SNP fasta sequence as mentioned above, users can click the “Input” button, and all information which is not belong to ‘A’, ‘T’, ‘C’, ‘T’, \([\text{dNTP1}/\text{dNTP2}/…/\text{dNTPn}]\), or IUPAC will be filtered out. The results for SNP fasta sequence input are shown as Fig. 3-10, including: (1) “All selected”, (2) “Clear”, (3) “Analyze sequence”, (4) the description for sequence, and (5) SNP sequence. Users can check the desired SNPs and click the “Analyze sequence” to next step. If the all SNPs need be analyzed, the “All selected” button provides convenience to check all SNPs. The “Clear” button provides cancel all SNP checked.

![Input SNP Fasta Results](image)

**Fig. 3-10. The results for SNP fasta sequence input.**

When clicking the “Analyze sequence”, users need to wait for the analysis of SNP-RFLP recognition (Fig. 3-11). Then, the results for SNP-RFLP recognition will be presented (Fig. 3-12). Finally, users can select the interested SNPs as described above.

![SNP-RFLP Recognition](image)

**Fig. 3-11. Wait for SNP-RFLP recognition.**
When uses input a fasta sequence with SNP in the in-del (Insertion and Deletion), tri-allelic, and tetra-allelic formats, their results are described as follows.

For SNP in the in-del format, the results are similar to di-allelic SNP (Fig. 3-13 and Fig. 3-14). For SNP in the tri-allelic format, the results for restriction enzymes information will present three sequence IDs with their corresponding available restriction enzymes for three different alleles (Fig. 3-15).

For SNP in the tetra-allelic format, the results for restriction enzymes information will present four sequence IDs with their corresponding available restriction enzymes for four different alleles (Fig. 3-16).

## For input SNP in fasta sequence format, the results for SNP-RFLP recognition will show the natural or mutagenic design results, because the SNP flanking sequence is not provided by the user input sequence. ##

---

### Fig. 3-12. The results for SNP-RFLP recognition.###

<table>
<thead>
<tr>
<th>SNP ID</th>
<th>Sequence</th>
<th>Result (+)</th>
<th>Result (−)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs11540652</td>
<td>GATTTCTGC ATGCCTCGCA TGACCC</td>
<td>![Green Square] The SNP in sequence (+) can be distinguished.</td>
<td>![Green Square] The SNP in sequence (−) can be distinguished.</td>
</tr>
</tbody>
</table>

---

### Fig. 3-13. The result for SNP-RFLP recognition with SNP in the in-del format.###

<table>
<thead>
<tr>
<th>SNP ID</th>
<th>Sequence</th>
<th>Result (+)</th>
<th>Result (−)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs28934876</td>
<td>CCGTCAGCAA GATGGTTCTGCA CAACTG</td>
<td>![Green Square] The SNP in sequence (+) can be distinguished.</td>
<td>![Green Square] The SNP in sequence (−) can be distinguished.</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>SNP ID</th>
<th>Sequence</th>
<th>Result (+)</th>
<th>Result (−)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs45541837</td>
<td>CAGTCTTGCC CTCTAATTCA GAATTA</td>
<td>![Green Square] The SNP in sequence (+) can't be distinguished.</td>
<td>![Green Square] The SNP in sequence (−) can't be distinguished.</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>SNP ID</th>
<th>Sequence</th>
<th>Result (+)</th>
<th>Result (−)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs68134313</td>
<td>TGGGATGTTGG TGGCGCTGCC CTTCGA</td>
<td>![Green Square] The SNP in sequence (+) can be distinguished.</td>
<td>![Green Square] The SNP in sequence (−) can be distinguished.</td>
</tr>
</tbody>
</table>
Fig. 3-14. The available restriction enzyme information for the sense strand with SNP in the in-del format.

<table>
<thead>
<tr>
<th>Sequence ID</th>
<th>Recognition Sequence</th>
<th>Commercial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-IUPAC</td>
<td>1. G-CATG-3</td>
<td>Bsp111/ Cvi31</td>
</tr>
<tr>
<td></td>
<td>2. -CATG-3</td>
<td></td>
</tr>
<tr>
<td>IUPAC</td>
<td>1. CTGAGTCGATGATTAGAGGGCTCCGCCAGCTGCTGATTATTTTGGT-3'</td>
<td>BstI, BsaGI, BsaCI, BsaGI, BsaCI, BsaCI, BstI, BstI, BstI, BstI</td>
</tr>
<tr>
<td></td>
<td>2. CTGAGTCGATGATTAGAGGGCTCCGCCAGCTGCTGATTATTTTGGT-3'</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 3-15. The available restriction enzyme information for the sense strand with SNP in the tri-allelic format.

<table>
<thead>
<tr>
<th>Sequence ID</th>
<th>Recognition Sequence</th>
<th>Commercial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-IUPAC</td>
<td>1. G-CATG-3</td>
<td>MnII</td>
</tr>
<tr>
<td></td>
<td>2. CATG-3</td>
<td></td>
</tr>
<tr>
<td>IUPAC</td>
<td>1. CTGAGTCGATGATTAGAGGGCTCCGCCAGCTGCTGATTATTTTGGT-3'</td>
<td>BclI, BclI, BclI, BclI, BclI, BclI, BclI, BclI</td>
</tr>
<tr>
<td></td>
<td>2. CTGAGTCGATGATTAGAGGGCTCCGCCAGCTGCTGATTATTTTGGT-3'</td>
<td></td>
</tr>
</tbody>
</table>
**Fig. 3-16. The available restriction enzyme information for the sense strand with SNP in the tetra-allelic format.** In this case, the enzymes for the antisense strand are chosen (as indicated by red arrow in the top). In the left side, three red arrow lines indicate that three nucleotides are distinguished between each other. The sequence 1 is unable to find the suitable enzyme but it is still distinct to others because it is unable to cut by the enzymes listed for sequences 0, 2, and 3.

<table>
<thead>
<tr>
<th>Sequence ID</th>
<th>Recognition Sequence</th>
<th>Commercial</th>
<th>Non-Commercial</th>
</tr>
</thead>
<tbody>
<tr>
<td>(+) = 0</td>
<td></td>
<td>IUPAC</td>
<td>Nebill NebillHin88lTspDI</td>
</tr>
<tr>
<td></td>
<td>1:5'GCAC3'</td>
<td>MsiI NruClBa3II</td>
<td>...</td>
</tr>
<tr>
<td>(+) = 1</td>
<td>2:5'TIRAC3'</td>
<td>HaeIIHin88lHin99l</td>
<td>...</td>
</tr>
<tr>
<td>(+) = 2</td>
<td>3:5'GCACGAC3'</td>
<td>Sall TaaIII</td>
<td>Acsl372lAcsl372lAcsl373l</td>
</tr>
<tr>
<td>(+) = 3</td>
<td>4:5'GGACGAC3'</td>
<td>IUPAC</td>
<td>...</td>
</tr>
<tr>
<td>(-) = 0</td>
<td>1:5'ACCT3'</td>
<td>Abu AluI</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>2:5'CGCCTG3'</td>
<td>BamHI RadRawnI</td>
<td>...</td>
</tr>
<tr>
<td>(-) = 1</td>
<td>3:5'ATGAA3'</td>
<td>TspDI</td>
<td>...</td>
</tr>
<tr>
<td>(-) = 2</td>
<td></td>
<td>IUPAC</td>
<td>...</td>
</tr>
<tr>
<td>(-) = 3</td>
<td></td>
<td>IUPAC</td>
<td>...</td>
</tr>
</tbody>
</table>
4. Function: Multiple SNPs within one sequence

4.1 Overview

In the function for multiple SNPs within one sequence, up to 50 SNPs representing in the [dNTP1/dNTP2] or IUPAC formats within the input sequence are acceptable for analysis. The flanking sequences for two nearby SNPs should not be overlapped within 6 nucleotides. There are two input areas in the function: one is “Description” which is a title for describing input sequence and the other is “Sequence” which is input multiple SNPs sequence.

Fig. 4-1 and Fig. 4-2 are two SNPs within one sequence in the [dNTP1/dNTP2] and IUPAC formats, respectively.

Fig. 4-3 and Fig. 4-4 are three SNPs within one sequence in the [dNTP1/dNTP2] and IUPAC formats, respectively.

Fig. 4-5 and Fig. 4-6 are four SNPs within one sequence in the [dNTP1/dNTP2] and IUPAC formats, respectively.

4.2 Example: Multiple SNPs within one sequence input

![Fig. 4-1](image1.png)

**Fig. 4-1. Two SNPs within one sequence in the [dNTP1/dNTP2] format.** When the sequence is too long, the part of the sequence will list in the next line automatically. The system is able to regard them as the single sequence input. If the SNP is marked in IUPAC code, the sequence is continued as shown in Fig. 4-2.
Fig. 4-2. Two SNPs within one sequence in the IUPAC format.

Fig. 4-3. Three SNPs within one sequence in the [dNTP1/dNTP2] format. If the SNP is marked in IUPAC code, the sequence is continued as shown in Fig. 4-4.
Fig. 4-4. Three SNPs within one sequence in the IUPAC format.

Fig. 4-5. Four SNPs within one sequence in the \([dNTP1/dNTP2]\) format. When the sequence is too long, the part of the sequence will list in the next line automatically. The system is able to regard them as the single sequence input. If the SNP is marked in IUPAC code, the sequence is continued as shown in Fig. 4-6.
After inputting multiple SNPs within one sequence, users can click the “Query” button, this system will separate multiple SNPs into specific SNP fasta sequence (Fig. 4-7). User can select the interested SNPs to perform SNP-RFLP analysis after clicking the button “Analyze sequence”. The following steps are the same as the function of (3) SNP in fasta sequence input (page 19) and they are shown here.

In Fig. 4-7, the “All selected” button provides convenience to check all SNPs and the “Clear” button provides cancel all SNP checked.

---

**Fig. 4-6. Four SNPs within one sequence in the IUPAC format.**

---

**Fig. 4-7. Four SNPs within one sequence in the IUPAC format.**
5. Function: GenBank accession no. for input

5.1 Overview

SNPs within the sequence information for the input accession no. of GenBank includes reference SNP ID (rs#), submitter SNP ID (ss#), accession version, HUGO gene name, local link ID (gene ID), and Local SNP ID. The classification of dbSNP in NCBI for function class (coding nonsynonymous, reference, intron, coding synonymous, locus region, mRNA UTR, and splice site), SNP class (heterozygous, indel, mixed, multinucleotide polymorphism, named locus, no variation, and snp), and heterozygosity are selectable. Furthermore, the TaqMan for non-limit, SNP500Cancer, ABI, and SNP500Cancer/ABI are available. All the information from GenBank is retrieved online for all available species (Fig. 5-1 and Fig. 5-2).

5.2 Example: Use HUGO gene name “TP53” to analyze

![Accession Key](image)

**Fig. 5-1. This input options of GenBank Accession.**

![Limits](image)

**Fig. 5-2. The options for organism, limits and accession in this function of GenBank Accession.**
After inputting accession no. (for example, TP53), users can click the button “Query” and it will be redirect to a waiting page (Fig. 2-12). Waiting for a moment, the results will be shown as Fig. 5-3 including: the key for input (here is Gene Name: TP53), the pages options, items for SNPs, function class, SNP class, Heterozygosity, TaqMan, and SNPs information which the same as (2) SNP ID input (Fig. 2-13) (page 11). The following operation (not shown) is similar as described above.

![Query SNPs Results: 20](image)

1. rs26160806 [Pan troglodytes]
   - TAGCCAGGCTGGTTTCACAACCTGCACTTCCACCACCTGCCACC
   - MapView: No VarVU, No Published
   - GeneView: SeqView
   - No 3D: No DMM

2. rs26062790 [Pan troglodytes]
   - TCCGAGATCATCATATACAAAGATGAA[AT]CTCATTCCAAGGCTGGGGGTGGG
   - MapView: No VarVU, No Published
   - GeneView: Not on mRNA
   - No 3D: No DMM

3. rs26897734 [Pan troglodytes]
   - TGGGTTGGGGGGGTTGGGGCTGCC[C/T]TTTTCAATGGATTTCCACACAGTTTC
   - MapView: No VarVU, No Published
   - GeneView: SeqView
   - No 3D: No DMM

![Fig. 5-3](image) The results for SNPs information query using HUGO gene name TP53 (here limit only show 20 SNPs and three SNPs is picked in the figure)
6. Function: TagSNP from HapMap

6.1 Overview

To reduce the necessary amount of SNPs for genotyping, it is believed that a subset of the SNP in a region (tagSNPs) can be chosen to represent most of the remaining SNP variants. The International HapMap Project is very helpful in providing the tagSNPs in the human genome for several races including YRI (Yoruba in Ibadan, Nigeria), JPT (Japanese in Tokyo, Japan), CHB (Han Chinese in Beijing, China), and CEU (CEPH; Utah residents with ancestry from northern and western Europe).

As shown in Fig. 6-1 and Fig. 6-2, the HapMap database versions, population, pairwise methods (tagger pairwise or tagger multimarkers), R square cut off, and MAF (minor allele frequency) cut off are adjustable. Position between chromosome, accession number, gene name, cytoband position, and ENCODE (ENCyclopedia Of DNA Elements) region can be queried. The tagSNPs information from HapMap is retrieved online and the mining function of RFLP restriction enzymes for the tagSNP is implemented. The current online linked tagSNP database is the HapMap Data Rel 23a/phaseII Mar08, on the NCBI B36 assembly, dbSNP b126. When users input query term (for example, Gene Name: ‘BRCA2’) and click the button “Query”, this system will query tagSNP from HapMap and sent a waiting page (Fig. 2-12). Then, the results for HapMap database versions, population, pairwise methods, R square cut off, and MAF cut off, and tagSNP sequence information will be shown (Fig. 6-3). The following operation is similar as described above.

6.2 Example: Use HUGO gene name ‘BRCA2’ to analyze

![Fig. 6-1. The results for SNPs information query using HUGO gene name ‘BRCA2’ in the setting of the HapMap database of “HamMap Data Rel 22/phasel2 Apr07, on NCBI B36 assembly, dbSNP b126 and other parameters as shown in this figure. The result is shown in Fig. 6-3.](image-url)
### The tagSNPs provided by the HapMap are selected from different blocks. In the same block, the picked tagSNP may be different from time to time at some intervals, such as:

- Interval-1: same as top view;
- Interval-2: partly different to Interval-1;

Since the SNP-RFLPing 2 is retrieved online to the HapMap, therefore, it has the same condition as shown in the HapMap.

---

**Fig. 6-2.** The results for SNPs information query using HUGO gene name ‘BRCA2’ (here limit only show 20 SNPs and three SNPs is picked in the figure).

**Fig. 6-3.** The results for tagSNPs information query using HUGO gene name ‘BRCA2’ as the setting for Fig. 6-1. The ten tagSNPs retrieved in SNP-RFLPing 2 (top view) is the same as shown in HapMap (bottom view). (Please click the “Animation” box to demonstrate)
7. Function: Transcript ID/miRNA

7.1 Overview

MicroRNAs (miRNAs) are a family of small RNAs that complement to the transcripts of protein coding genes, repress translation, or decrease mRNA stability. The dysfunction of miRNA can impact many targets and, thus, affect cell physiology and cancer progression. Polymorphisms in the miRNA pathway may affect gene expression, lead to change complex phenotypes, and have the potential to be disease markers for personalized medicine. In this function, the polymorphism in microRNA Target Site (PolymiRTS) database is applied to provide transcript ID and miRNA number of human and mouse to query SNPs (Fig. 7-1).

7.2 Example: Use transcript ID and miRNA number to analyze

![Fig. 7-1. The interface for Transcript ID/miRNA.](image)

When users click the button “Query”, this miRSNP information will be queried, and the waiting page will be shown (Fig. 7-2). After the query finished, the results for miRSNP will be shown. Fig. 7-3 and Fig. 7-4 show the result for query transcript ID “NM_015658” of human and the results for query transcript ID “miR-198” of human, respectively. The provided information contains transcript ID, SNP ID (rs#), Location (SNP position in transcript ID), Chr Pos (chromosome position), Ancestral Allele, Allele, miR ID (miRNA ID), Support, miR Site (miRNA site), and FuncClass (function class).

![Fig. 7-2. Wait for miRSNP information query.](image)
### Fig. 7-3. The result for query transcript ID “NM_015658” of human.

<table>
<thead>
<tr>
<th>ID</th>
<th>Transcript ID</th>
<th>SNP ID</th>
<th>Location</th>
<th>Chr Pos</th>
<th>Ancestral Allele</th>
<th>Allele</th>
<th>miR ID</th>
<th>Support</th>
<th>miRSite</th>
<th>FuncClass</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NM_015658</td>
<td>rs28319</td>
<td>2651</td>
<td>869549</td>
<td>-</td>
<td>G</td>
<td>miR-507</td>
<td>0</td>
<td>tcTTTGAAAat</td>
<td>O</td>
</tr>
</tbody>
</table>

### Fig. 7-4. The results for query miRNA number “miR-198” of human. (This figure show 8 records of 10 records)

<table>
<thead>
<tr>
<th>ID</th>
<th>Transcript ID</th>
<th>SNP ID</th>
<th>Location</th>
<th>Chr Pos</th>
<th>Ancestral Allele</th>
<th>Allele</th>
<th>miR ID</th>
<th>Support</th>
<th>miRSite</th>
<th>FuncClass</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NM_198576</td>
<td>rs2799073</td>
<td>6632</td>
<td>980688</td>
<td>G</td>
<td>A</td>
<td>miR-198</td>
<td>0</td>
<td>cTCTGGACcdtcgc C</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>NM_022114</td>
<td>rs12735779</td>
<td>5760</td>
<td>3342087</td>
<td>G</td>
<td>C</td>
<td>miR-198</td>
<td>0</td>
<td>tcactCTGGGACA C</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>NM_199454</td>
<td>rs12735779</td>
<td>5703</td>
<td>3342087</td>
<td>G</td>
<td>G</td>
<td>miR-198</td>
<td>0</td>
<td>tcactCTGGGACA C</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>NM_000778</td>
<td>rs11211405</td>
<td>2312</td>
<td>47167933</td>
<td>A</td>
<td>G</td>
<td>miR-198</td>
<td>0</td>
<td>tgTCTGGACgtgc C</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>NM_000566</td>
<td>rs16865014</td>
<td>1163</td>
<td>148029723</td>
<td>C</td>
<td>C</td>
<td>miR-408</td>
<td>2</td>
<td>atcgATCTGGGAcc atcgATCTGGGACc D</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>NM_021133</td>
<td>rs641265</td>
<td>4093</td>
<td>180809448</td>
<td>T</td>
<td>T</td>
<td>miR-198</td>
<td>0</td>
<td>cCTGGACAtgta C</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>NM_001002295</td>
<td>rs12782127</td>
<td>2197</td>
<td>8156297</td>
<td>T</td>
<td>G</td>
<td>miR-198</td>
<td>10</td>
<td>ataTACTGGCAaatg C</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>NM_002051</td>
<td>rs12782127</td>
<td>2194</td>
<td>8156297</td>
<td>T</td>
<td>G</td>
<td>miR-198</td>
<td>10</td>
<td>ataTACTGGCAaatg C</td>
<td></td>
</tr>
</tbody>
</table>
In order to query the SNP sequence information for SNP-RFLP analysis, users can check the box in the front of transcript ID and click the button “Search SNP Sequence”. A waiting page for SNPs information will be displayed. Then, users can reselect SNP once again for SNP-RFLP analysis (Fig. 7-5). The following operation is similar as described above (not shown here).

Fig. 7-5. The SNP information results for query two transcript IDs “NM_198576” and “NM_199454” of miRNA number “miR-198” of human.
8. Function: Gene Ontology-based annotation for SNPs

8.1 Overview

The Gene Ontology Browser (GO Browser; [http://cgap.nci.nih.gov/Genes/GOBrowser](http://cgap.nci.nih.gov/Genes/GOBrowser)) providing the annotation for human and mouse genes by molecular function, biological process, and cellular component. GO IDs and vocabulary terms may be input to find the genes with an interested function as well as their corresponding SNPs. In this function, CGAP database (human and mouse), Chromosomes, and GO term are available for input (Fig. 8-1). In order to get more GO terms, users can link to CGAP GO Browser ([http://cgap.nci.nih.gov/Genes/GOBrowser](http://cgap.nci.nih.gov/Genes/GOBrowser)) (Fig. 8-2).

8.2 Example: Query Gene Ontology vocabulary term

Fig. 8-1. The interface for the input of “Gene Ontology-based annotation for SNPs”.

![CGAP GO Browser Interface](image.png)
Fig. 8-2. More GO terms can be found in CGAP GO Browser. Users can copy-and-paste the interested GO term to the window of GO term to search the SNP information for specific GO term. (Please click the “Animation” box to demonstrate)

When users input GO term (for example, positive regulation of interleukin-2 biosynthetic process), the waiting page for querying CGAP GO information will be shown (Fig. 8-3). Waiting for a moment, the results which contain UniGene, Gene Symbol, Locus ID, CytoBand, OMIM, sequence, and SNP Information for the query GO term will be presented (Fig. 8-4). In SNPs information, two types for SNP query are available: one is NCBI SNPs, and the other is HapMap tagSNP. Through clicking the SNPs and tagSNPs images, the SNPs information will be inquired. The SNP query function is similar as described above.

Fig. 8-3. Wait for CGAP GO information query.
Fig. 8-4. The results for query Gene Ontology vocabulary term “positive regulation of interleukin-2 biosynthetic process”.

<table>
<thead>
<tr>
<th>ID</th>
<th>UniGene</th>
<th>Gene Symbol</th>
<th>Locus ID</th>
<th>Cytoband</th>
<th>OMIM</th>
<th>sequence</th>
<th>SNPs Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Hs.838</td>
<td>CD80</td>
<td>H041</td>
<td>3q13.3-q21</td>
<td>112203</td>
<td>NM_005191</td>
<td><img src="tagSNPs" alt="SNPs" /></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Hs.1722</td>
<td>IL1A</td>
<td>3552</td>
<td>2q14</td>
<td>147760</td>
<td>NM_000575</td>
<td><img src="tagSNPs" alt="SNPs" /></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Hs.49195</td>
<td>GLMN</td>
<td>11146</td>
<td>1p22.1</td>
<td>601749</td>
<td>NM_053274</td>
<td><img src="tagSNPs" alt="SNPs" /></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Hs.126756</td>
<td>IL1B</td>
<td>3553</td>
<td>2q14</td>
<td>147720</td>
<td>NM_000576</td>
<td><img src="tagSNPs" alt="SNPs" /></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Hs.171182</td>
<td>CD86</td>
<td>942</td>
<td>3q21</td>
<td>601020</td>
<td>NM_175862</td>
<td><img src="tagSNPs" alt="SNPs" /></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Hs.401013</td>
<td>IRF4</td>
<td>3662</td>
<td>6p25-p23</td>
<td>601900</td>
<td>NM_002460</td>
<td><img src="tagSNPs" alt="SNPs" /></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
9. Function: File upload for input

9.1 Overview
This function is similar to (2) SNP ID input, and (3) SNP in fasta sequence input functions. SNP ID (rs# and ss#) and SNP in fasta sequence format file (*.txt) are acceptable to query the SNP-RFLP information. User can select either “SNP IDs” or “SNP fasta sequences” option, and then through the button “browse…” to get an available input file in local machine (Fig. 9-1). Finally, click the button “Put In” to perform this function. For more detail formats and steps for operating, please reference to (2) SNP ID input, and (3) SNP in fasta sequence input functions.

9.2 Example: Upload interface for SNP IDs and SNP fasta sequences

![Fig. 9-1. The interface for “File upload” function.](image-url)
10. Appendix

10.1 Abbreviations

- **SNP**: Single nucleotide polymorphism
- **RFLP**: Restriction Fragment Length Polymorphism
- **CGAP**: Cancer Genome Anatomy Project
- **PolymiRTS**: Polymorphism in microRNA Target Site database
- **HUGO**: Human Genome Organisation
- **IUPAC**: The International Union of Pure and Applied Chemistry
- **YRI**: Yoruba in Ibadan, Nigeria
- **JPT**: Japanese in Tokyo, Japan
- **CHB**: Han Chinese in Beijing, China
- **CEU**: CEPH (Utah residents with ancestry from northern and western Europe
- **MAF**: Major Allele Frequency
- **miRNA**:
- **GO**: Gene Ontology
- **ENCODE**: ENCyclopedia Of DNA Elements

10.2 Related links

- **PolymiRTS**: [http://compbio.utmem.edu/miRSNP/](http://compbio.utmem.edu/miRSNP/)
- **CGAP GO browser**: [http://cgap.nci.nih.gov/Genes/GOBrowser](http://cgap.nci.nih.gov/Genes/GOBrowser)
- **ENCODE**: [http://genome.ucsc.edu/ENCODE/](http://genome.ucsc.edu/ENCODE/)
- **Seq-SNPing**: [http://bio.kuas.edu.tw/Seq-SNPing/](http://bio.kuas.edu.tw/Seq-SNPing/)
- **SNP-Flankplus**: [http://bio.kuas.edu.tw/snp-flankplus/snpin.jsp](http://bio.kuas.edu.tw/snp-flankplus/snpin.jsp)
- **LD2SNPing**: [http://bio.kuas.edu.tw/LD2SNPing/](http://bio.kuas.edu.tw/LD2SNPing/)
- **SNP ID-info**: [http://bio.kuas.edu.tw/snpid-info](http://bio.kuas.edu.tw/snpid-info)