

SNP-Flankplus: SNP ID-centric retrieval for SNP flanking sequences

Cheng-Hong Yang¹, Yu-Huei Cheng¹, Li-Yeh Chuang^{2*} and Hsueh-Wei Chang^{3,4,5}

¹Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Taiwan; ²Department of Chemical Engineering, I-Shou University, Taiwan; ³Faculty of Biomedical Science and Environmental Biology, Kaohsiung Medical University, Taiwan; ⁴Graduate Institute of Natural Products, College of Pharmacy, Kaohsiung Medical University, Taiwan; ⁵Center of Excellence for Environmental Medicine, Kaohsiung Medical University, Taiwan;

Li-Yeh Chuang * - Email: chuang@isu.edu.tw; * Corresponding author

received November 25, 2008; accepted December 01, 2008; published December 06, 2008

Abstract:

The flanking sequences provided by dbSNP of NCBI are usually short and fixed length without further extension, thus making the design of appropriate PCR primers difficult. Here, we introduce a tool named “SNP-Flankplus” to provide a web environment for retrieval of SNP flanking sequences from both the dbSNP and the nucleotide databases of NCBI. Two SNP ID types, *rs#* and *ss#*, are acceptable for querying SNP flanking sequences with adjustable lengths for at least sixteen organisms.

Availability: This software is freely available at <http://bio.kuas.edu.tw/snp-flankplus/>

Keywords: PCR; SNP; primer design; flanking sequences

Background:

Single nucleotide polymorphisms (SNPs) are the most commonly encountered genetic variants. Many kinds of primer design software tools, such as Primer 3 [1], provide the suitable polymerase chain reaction (PCR) primers for the PCR-based SNP genotyping methods. A longer template sequence is more helpful for optimal primer design; however, the SNP flanking sequences provided in NCBI dbSNP [2] are not always long enough for regular primer design.

Recently, FESD [3] designed a “SNPflank” function to identify flanking sequences for SNP IDs and provided customizable length with *rs#* input alone for human SNPs but is inaccessible recently. To offer longer template sequences for desired SNP for genotyping experiments, such as TaqMan real-time PCR [4], PCR-RFLP [5], and PCR-CTTP [6], we introduce the SNP-Flankplus for on-line retrieval of flanking sequences of target SNPs for sixteen organism genomes.

Methodology:

The system design, algorithm and database of the program are described below.

Algorithm

This program adopts the sequences of accession numbers of the corresponding SNPs and the SNP contig position to obtain desired flanking sequence with specific length. In order to save memory space during reading the sequence of

accession numbers, this system employs “block location way”, which splits the sequence of the accession numbers into multiple blocks. A specific block is loaded into the memory to search the required sequence and is hit by the algorithm 1 (under supplementary material).

When the flanking length exceeds a block, some nearby blocks are used, i.e. (block hit - d) or (block hit + d). d is the size of extending blocks and is calculated by the algorithm 2 (under supplementary material).

Database

The source databases are retrieved on-line and constantly updated from NCBI dbSNP and Nucleotide [4].

Result:

Input

The four main input interfaces in SNP-Flankplus are followed: (1) Single Reference cluster ID (*rs#*) input; (2) Single NCBI Assay ID (*ss#*) input; (3) Multiple SNP ID *rs#* and *ss#* input by pasting; and (4) Multiple *rs#* and *ss#* input through uploading a file (Figure 1a). Users are allowed to enter the SNP ID or multiple SNP IDs (*rs#* or *ss#*) for sixteen organisms when querying SNP information. When using the *ss#* input, the system will first query the corresponding *rs#*, and then retrieve SNP information related to this *rs#*. The SNP information contains allele information, submitted SNPs and other data for this RefSNP Cluster. Users can set the desired flanking length for the design of feasible primer sets. Two flanking length

options are available: the system can be either set to default lengths of 300 ~ 1000 bps, or alternatively, the length can be set to the maximum length of the corresponding contig accession (Figure 1b).

Output

The flanking sequence output is shown in fasta format with on-line representation and file and/or text. It contains SNP ID (rs#), allele name, chromosome position of SNP, contig position of SNP, organism source, contig accession and

sequence corresponding position, SNP type, sequence type, and case sensitivity. This information is separated by the “|” symbol. Its limitation of maximum flanking length is dependent on the corresponding contig accession number. Three types of flanking sequences are able to adjustable in real-time, such as: (1) SNP types contain general nucleotides, alleles, and IUPAC formats, (2) sequence types contain original, reverse, complementary, antisense sequences, and (3) case sensitive types contain upper case and lower case (Figure 1c).

(a) Input interfaces:

- Input Reference cluster ID(rs#) to get SNP flanking sequence
- Input NCBI Assay ID(ss#) to get SNP flanking sequence
- Paste multiple Reference cluster ID(rs#) and/or NCBI Assay ID(ss#) to get SNP flanking sequences
- Upload file including multiple Reference cluster ID(rs#) and/or NCBI Assay ID(ss#) to get SNP flanking sequences

(b) SNP information and adjustable flanking length:

SNP Flanking length set

Sequence length: Select to flanking sequence length: 300 Input to flanking sequence length: 300

1. refSNP ID: rs11878200 RefSNP Alleles: C/T Variation Class: SNP: single nucleotide polymorphism
submitted SNP(ss#): 1. ss17612126 2. ss41030732 3. ss80577258 4. ss90918682

Organism	Molecule Type	Ancestral Allele	Clinical Association	Created/Updated in build	Map to Genome Build
human (Homo sapiens)	Genomic	Not available	unknown	120/129	36.3

2. refSNP ID: rs62062585 RefSNP Alleles: A/G Variation Class: SNP: single nucleotide polymorphism
submitted SNP(ss#): 1. ss90524423

Organism	Molecule Type	Ancestral Allele	Clinical Association	Created/Updated in build	Map to Genome Build
human (Homo sapiens)	Genomic	Not available	unknown	129/129	36.3

(c) SNP flanking sequence:

SNP flanking information set

SNP type: General nucleotide Alleles IUPAC

Sequence type: Original Reverse Complementary Antisense

Case sensitive: Upper case Lower case

SNP flanking sequence output

Output:

```
>rs11878200|alleles=[C/T]|chr=19|ctg_pos=8970009|chr_pos=17568207|organism=human(Homo sapiens)
|NT_011295.10:8969769-8970249|snp_type=general|seq_type=original|case_sensitive=uppercase
TTAGTATCC CCACCTGCTC AGTTCCTTA TTAGGCCGAG ACACCTTAA TAAATATCT
GCTTCCCTGA CTATTCCTAG GTGACAGCCA CACCACATTG CCACCTTTTC CCCAGTTCA
AAGCCTCCTT CACATCTCC CTTGTATCT CCCACCCTTA ACCCACAAGT ATAAGACACC
TCTACTCCCT CTTAGCGAG CGATCATGCA CCCCCTTACCA TCCCATTAAA ACCTAATCAC
C
CTTACCCGGA TCAATGCCAA TATCCCATCC CACAGCACAC TTGAAAGGA TTAACCTG
TTATCATTCT CCGTTCACAG CATGGCCTTT TAAAGCCTAT AAACCCTCCT TACCATTCCC
TCATTTTACC TGTCCTAAAA CCAGACGATG CTTACAGGTT AGTTCAGGAT CTGGCCCTTA
TCAACCAAT TGTTTTGCTT ATCCACCCCG TGGTGCCAAA CCCATATACT CTCTATCCT
>rs62062585|alleles=[A/G]|chr=17|ctg_pos=7184430|chr_pos=7527806|organism=human(Homo sapiens)
|NT_010718.15:7184190-7184670|snp_type=general|seq_type=original|case_sensitive=uppercase
CGACTGGCCT ATTTATCCTT TTTTCTATAT TACCAGGTTT GGTITGCTAA AATTGGTTAG
CTGTTGCTAG TCTATGCTAA CAGGAATATT GGTCTATATT TTCTTTCTT ATAATGCTCT
TGTTTTGTTT TGGTACGAG ATTATGCTGG CTTCGAAAAC AAGITGGGAA ATATTCCTCT
ATTTTTCTT TCTTTCTTTT TTTTCTGAG ACAGGGTCTC ACTCTGTTGC CCAGGCTGGA
G
TGCAAGGGG CAATCTGGC TCATGCAAC CTCCTGCTCC CAGGTTCAA GGATCTTGT
GCCTCAGCCT CCTGAGTAA TGGCATTACA GGTATGTGTC ACCACGCTA GCTACTTTT
GTATTTTAG TAGAGATGG GTTTCGCCGT GTTGCCAGG CTGGTCTCGA ACTCCGACC
TCAAATGATC CCTGCCTCA GCGTACCAA GTGCTGAGAT TACAGGTATG AGCCACCGG
```

Figure 1: A web snapshot. (a) Four input interfaces. (b) SNP information and adjustable flanking length. (c) File or text output.

Conclusion:

SNP-Flankplus provides a real-time update mechanism is employed, and two SNP ID types (rs# and ss#) for sixteen organisms can be entered to obtain the latest SNP information and sequence. A maximum flanking length can be retrieved based on the corresponding contig accession number.

Acknowledgment:

This work was partly supported by the National Science Council in Taiwan under grant NSC97-2622-E-151-008-CC2, NSC96-2221-E-214-050-MY3 and KMU-EM-97-1.1b.

References:

- [01] S. Rozen *et al.*, *Methods Mol Biol*, 132: 365 (2000) [PMID: 10547847]
- [02] S. T. Sherry *et al.*, *Nucleic Acids Res.*, 29: 308 (2001) [PMID: 11125122]
- [03] H. J. Kang *et al.*, *Nucleic Acids Res.*, 33: D518 (2005) [PMID: 15608252]
- [04] F. M. De la Vega *et al.*, *Mutation Res.*, 573: 111 (2005) [PMID: 15829242]
- [05] H. W. Chang *et al.*, *BMC Genomics*, 7: 30 (2006) [PMID: 16503968]
- [06] N. Hamajima, *et al.*, *J Mol Diagn*, 4: 103 (2002) [PMID: 11986401]

Edited by P. Kanguane

Citation: Yang *et al.*, *Bioinformatics* 3(4): 147-149 (2008)

License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.

Supplementary material

Algorithms

```
If (SNP contig position % m * n == 0)
    block hit = SNP contig position / m * n;
else
    block hit = contig position / m * n + 1;
→ (1)
```

where m is the line length of the sequence of accession numbers in the fasta format and n is the block size having split. The symbols ‘%’, ‘/’, and ‘*’ represent to get the remainder after division, the division operation, and the multiplication, respectively.

```
If (flanking length / 2 > (SNP position in the block))
    d = (flanking length / 2) / m * n;
if ((SNP position in the block - 1) < (flanking length / 2) % (m * n))
    d++;
→ (2)
```