

# Before SNP mapping: Data preprocessing by fixed length genomic sequence patterns

Chia-Hao Ou , Ming-Jing Hwang

**Keywords:** SNP mapping, genomic sequence analysis, sequence patterns, data preprocessing

## 1 Introduction.

How to map SNP sequences onto their genomic positions and evaluate the mapping reliability is an issue of interest in current bioinformatics researches. The conventional approach to this problem is to use sequence alignment tools, such as BLAST [1], to evaluate the mapping results [2], but sequence alignment process demands a great deal of computational power, and repetitive DNA, such as repetitive elements and segmental duplications, could significantly complicate sequence alignment. Our previous work, the UniMarker (UM) method [3], employed an alignment-free sequence mapping program and achieved a high agree rate, 99.73%, with NCBI mapping results on dbSNP. To further improve on the efficiency and reliability of this method, in this work, we used hits of fixed length (14 base pairs) genomic sequence patterns as a data cleaning indicator to assign each SNP record to its most likely chromosomes and delegate only those of a low hit ratio to be mapped against all the chromosomes. This simple prescreening method was shown to identify some erroneous assignments of NCBI.

## 2 Methods.

When a SNP record is completely aligned to a genomic fragment, both sequences should share common subsequences. If there are mutations or indels in the SNP record, its hit rate (number of identical subsequent fragments) should be much lower than records that can be completely aligned. Since we can find unique sequences in most of the SNP records, we can use the hit ratio of each chromosome to find each SNP record's chromosome quickly and reliably.

In this work, we used a 14-mer overlapping window to generate 14-mer sequence patterns of each chromosome, with 50 arrays to store the segment patterns from all positive and negative strands of chromosomes. After that, we scanned each SNP record by the same 14-mer overlapping window and keep a chromosome-specific pattern count when there is a hit. Subsequences that contain an alphabet other than ATCG, such as N, will be counted by not\_ATCG counter. After scanning a SNP sequence, each chromosome's 14-mer hit ratio is calculated by the following equation:

$$\text{Hit ratio (\%)} = \text{Chromosome hits} * 100 / (\text{SNP sequence length} - \text{not\_ATCG count} - 13)$$

## 3 Results.

It takes 30 minutes to generate the 50 arrays that contain every chromosome's 14-mer patterns and another 4 hours to generate the hit ratio results for 3,365,561 SNP records (dbSNP build 115, the files named rs\_chNotOn.fas, rs\_chMulti.fas and rs\_chMasked.fas were excluded). We use the

following strategies to process the results: (1) Keep the chromosome identity for the SNP when the hit ratio is higher than 99%. (2) Keep the chromosome identity with the highest hit ratio (larger than 70%) when each hit ratio is lower than 99%. The results showed that 99.8% SNP records contain the same chromosome identity with the NCBI assignment. For SNP records satisfying either condition, the difference in the rate of the best and the second hit rate larger than 10% is 83.9%. Only 275 records (out of total 2,824,219 records) have chromosome assignments different from NCBI when the hit rate difference is larger than 10%. Checking with BLAST, BLAT, and SSAHA, we have found, for many of these records, the results of these three alignment based methods agreed with our pre-screening assignment and not with the assignment of NCBI.

## References

- [1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. 1990. Basic local alignment search tools. *J. Mol. Biol.* 215:403-410.
- [2] Kitts, A. and Sherry, S. The SNP Database of Nucleotide Sequence Variation. *The NCBI Handbook*.
- [3] Chen, Y.Y., Lu, S.H., Shih, S.C. and Hwang, M.J. 2002. Single Nucleotide Polymorphism Mapping Using Genome-Wide Unique Sequences. *Genome Research* 12-7:1106-1111.