Journal of Genetics and Molecular Biology
Vol. 17, No. 3, 156-161, September 1, 2006

Affyrmation: Online Real-time Affirmation and Annotation for Affymetrix SNP Information

Yan-Hau Chen, Ming-Fang Tsai, and Adam Yao*

Institute of Biomedical Sciences, Academia Sinica

Affyrmation, a web-based application, provides online affirmation and annotation for Affymetrix SNP information. Affymetrix SNP information such as dbSNP_ID, chromosome, physical position, flanking sequence and associated gene are verified with NCBI dbSNP, UCSC Genome Browser, and SNPper to make sure the information is up-to-date. In addition, Affyrmation digs out gene function, tissue specificity, disease, subcellular location, pathway, Gene Ontology, plus some useful URLs to HapMap, RefSeq, OMIN, etc. in a 1MB region flanking each submitted SNP. All results are graphically displayed for good visualization. Data can be exported in CSV or XLS format for further analysis. Affyrmation can be accessed at http://genepipe.ngc.sinica.edu.tw/affyrmation/.

Key words: ***SNP, annotation, web, verification, Affymetrix***

Introduction

DNA chip technology is becoming widely used in finding disease-associated genes by analyzing genetic information for gene expression and genotyping. Affymetrix 100K SNP chip [1] is one of the most popular technologies adopted by many researchers for the above purpose. Though, Affymetrix provides their SNP annotation files for download to help further analysis, the update interval of the Affymetrix SNP information is at least three months or longer. As a result, we have developed Affyrmation to make up the possible lag between the Affymetrix SNP information and the most up-to-date information in the public domain.

Affyrmation is a web application that lets researchers use from any computer having access to the Internet. Affyrmation shows a friendly result visualization interface with good self-explanatory design. The researchers can easily understand the results and know quickly how to use it after a few clicks.

The application verifies the Affymetrix SNP information such as dbSNP_ID, chromosome, physical position, flanking sequence and associated genes.

*Corresponding author: Dr Adam Yao, Inst of Biomedical Sciences,Academia Sinica, No. 128, Academia Rd, Sec. 2, Nankang 11529, Taipei, Taiwan. TEL: 886-2-2652-3091
e-Mail: adam@ibs.sinica.edu.tw

This article is available online at: http://www.ceps.com.tw

An example file is available for download from the website. While all the fields shown in the example file are required, only one of the two fields, [dbSNP_ID] and [Flanking Sequence] must be filled. The real-time SNP information used by Affyrmation for affirmation is extracted from three sources in the following order: NCBI [2], UCSC [3] and SNPper [4].

The researchers can either input the Affymetrix chip data in a file or download and modify the annotation files from Affymetrix. Then upload to Affyrmation for processing. For each submission, the maximum SNP number allowed for Affyrmation is 100.

IMPLEMENTATION

Stage I

When Affyrmation gets the submitted file, it will affirm the information on a row-by-row basis. If the dbSNP_ID field is filled, Affyrmation will query the information by dbSNP_ID in NCBI. Otherwise, it will align the flanking sequence with BLAT to get the dbSNP_ID from UCSC Genome Browser first before querying NCBI the SNP information. The returned SNP information from the query such as chromosome name, physical position, and flanking sequence is extracted to compare with the information from Affymetrix. Matched, mismatched, and unavailable data are displayed in green, red, and grey color, respectively.

Stage II

Affyrmation verifies 'associated gene' field in the file by queried their gene_ID from public NCBI website to check if all information about genes are matched or mismatched. If there's not enough information in NCBI then the process will turn to UCSC Genome Browser then to SNPper to get the relation.

Stage III

Finally, all of the SNPs are shown on the chromosome map for an overview of the verified results. It is easy to see the distribution of SNP and clearly to tell the consistency in the results by different colors (See Fig.1).

If all the data in this column of the file is exactly matched as the current version of the data, the match rate will be 100%. A summary chart is also available to display the match rates of all columns.

Clicking any SNP on the chromosome map will show information of the nearby gene(s) including gene function, tissue specificity, disease, sub-cellular location [*5*], pathway [*6*], and Gene Ontology [7] in a 1Mbp region nearby the SNP (See Fig.2). Links to SNP, HapMap, PubMed, etc. of the region are also provided to help the researchers prioritize candidate genes (See Fig.3). Once a candidate gene is selected, primers of the gene exons and promoter region can easily be generated with our PrimerZ (http://genepipe.ibms.sinica.edu.tw/primerz/) whose link is also included in Affyrmation.

The system provides output result in three formats, HTML, Excel, and CSV. However we recommend users to choose HTML format which can present better visual result with dynamic content and additional information. Excel and CSV formats are good for further data processing.

The software and development environment we used are as the followings:

1. Java, JDK: j2sdk1.5_06, Server VM
2. Struts Framework 1.2
3. Red Hat Enterprise Linux Academic Edition
4. Tomcat 5.5 web server
5. Kapow RoboSuite

Affyrmation was developed in Struts Framework. The web server is Tomcat 5.5 on Red Hat Linux platform. Kapow RoboSuite robots are mainly used to retrieve web pages and data for affirming the SNPs information

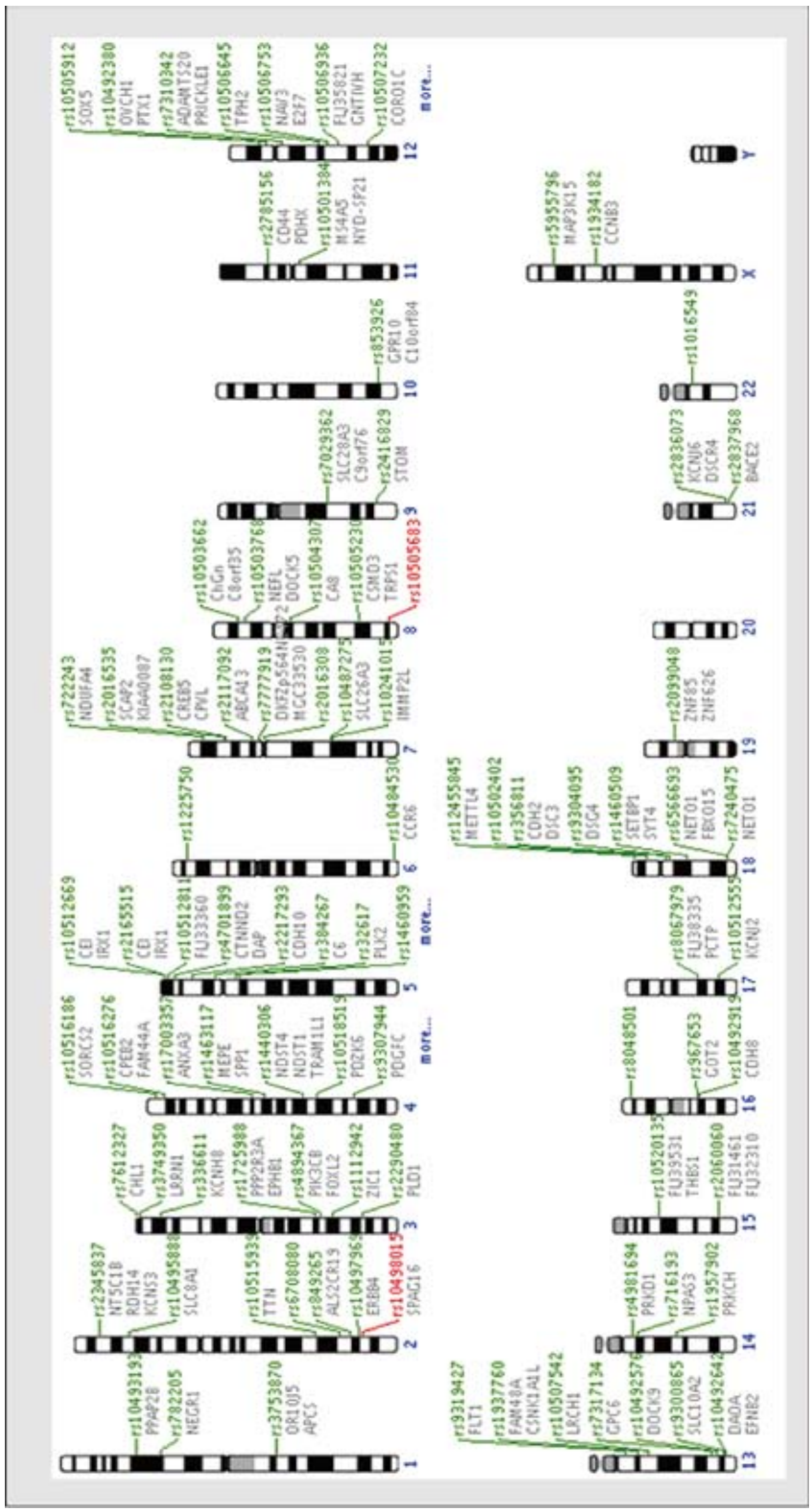

Fig. 1. An overview of the verified results on the chromosome diagram

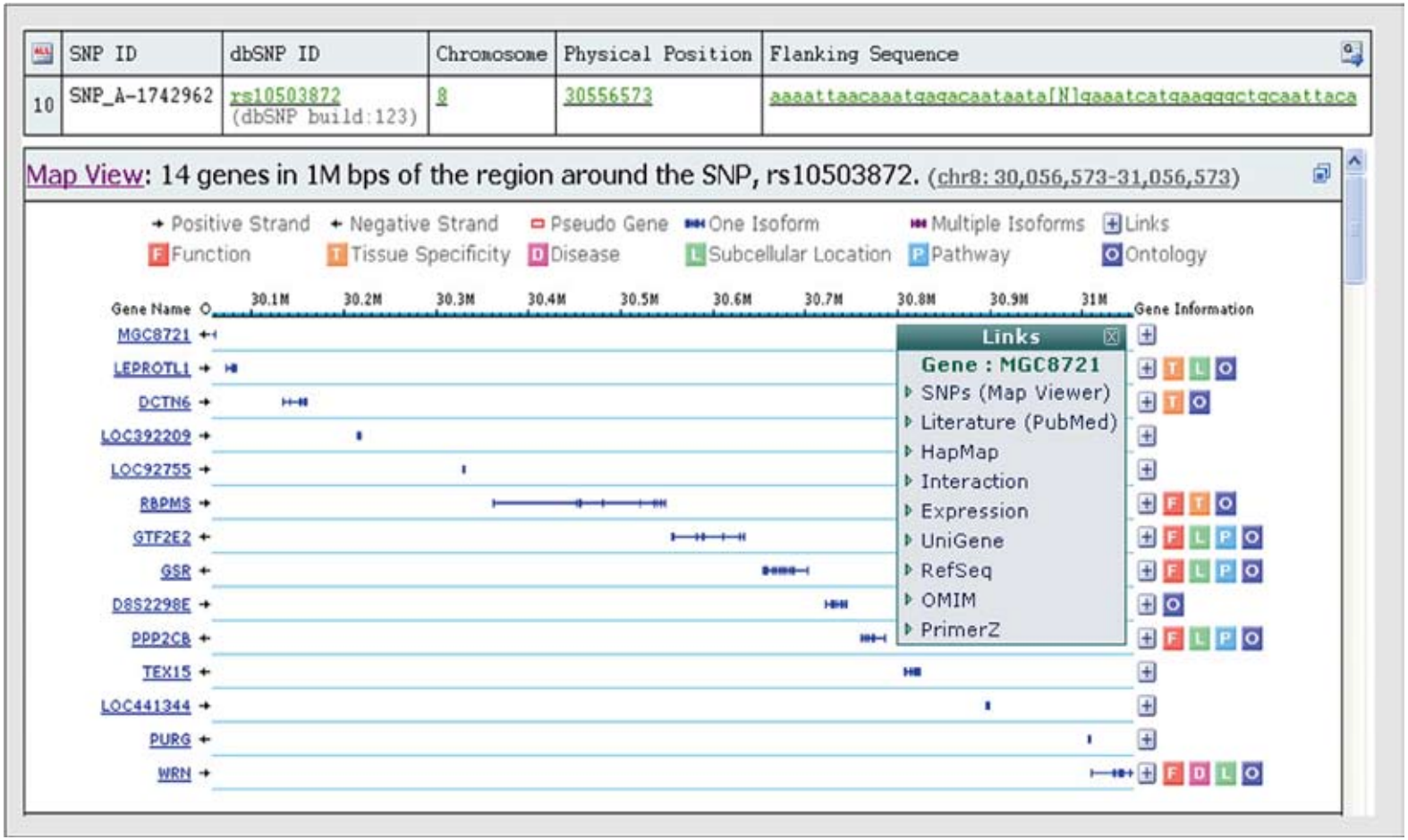

Fig. 2. The gene information including gene function, SNPs in the gene, tissue specificity, disease, sub-cellular location, pathway, and ontology in a 1Mbp region nearby the selected SNP.

Fig. 3. The text view shows the detail gene information

of a submitted file.

CONCLUSION

Affymetrix 100K SNP chip has been on the market for more than one year and its newly developed 500K SNP chip is coming out as well. More and more data generated from these chips need to be managed, processed, and kept up-to-date in the post-genomic era. The more up-to-date lab data is usually more reliable because of advanced technologies applied. However, frequent updating process can become very tedious for researchers utilizing data from many various sources. Affyrmation fluently simplifies the process by automatically verifying data between Affymetrix and SNP information sources.

In addition, Affyrmation also provides the gene information around each SNP. The visualization interface can help the researchers identify interesting information easily. Later on, we plan to add more features, such as verifying allele frequency and Ensembl gene, and links to microarray data and protein structure.

ACKNOWLEDGEMENTS

Thanks to Dr. Chien-Hsun Chen for the helpful discussion at the beginning of this work and Dr. Wen-Harn Pan for her suggestion on result visualization. The project is supported by Academia Sinica Life Science Grant No. 40-19(GSK)

REFERENCES

[1] Kennedy GC, Matsuzaki H, Dong S, Liu WM, Huang J, Liu G, Su X, Cao M, Chen W, Zhang J, et al.: Large-scale genotyping of complex DNA. Nature Biotechnology, 2003; 21: 1233–1237.

[2] Sherry, S.T., Ward, M.H., Kholodov,M., Baker,J., Phan,L., and Smigielski,E.M. et al.: dbSNP: the NCBI database of genetic variation. Nucleic Acids Res, 2001; 29: 308–311.

[3] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, and Haussler D: The human genome browser at UCSC. Genome Res, 2002; 12(6), 996-1006.

[4] Riva AA, and Kohane IS: SNPper: retrieval and analysis of human SNPs. Bioinformatics, 2002; 18: 1681-1685.

[5] Zdobnov EM, Lopez R, Apweiler R, Etzold T: The EBI SRS server-recent developments.

Bioinformatics, 2002; 18: 368-373.
[6] Kanehisa M, Goto S: The KEGG databases at GenomeNet. Nucleic Acids Res, 2002; 30: 42–46.
[7] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nature Genet, 2000; 25: 25-29.