Gene expression

ArrayFusion: a web application for multi-dimensional analysis of CGH, SNP and microarray data

Tsun-Po Yang¹, Ting-Yu Chang², Chi-Hung Lin¹,²,⁴, Ming-Ta Hsu¹,³ and Hsei-Wei Wang¹,²,⁴,*

¹Microarray and Gene Expression Analysis Core Facility, VGH National Yang-Ming University Genome Research Center, ²Institute of Microbiology and Immunology, ³Institute of Biochemistry and Molecular Biology, National Yang-Ming University and ⁴Department of Teaching and Research, Taipei City Hospital, Taipei, Taiwan

Received on March 16, 2006; revised on July 11, 2006; accepted on August 21, 2006
Associate Editor: Joaquin Dopazo

ABSTRACT

Summary: ArrayFusion annotates conventional CGH results and various types of microarray data from a range of platforms (cDNA, expression, exon, SNP, array-CGH and ChIP-on-chip) and converts them into standard formats which can be visualized in genome browsers (Affymetrix™ Integrated Genome Browser and GBrowse in the HapMap Project). Converted files can then be imported simultaneously into a single genome browser to benefit a collective interpretation between different array results. ArrayFusion therefore provides a new type of tool facilitating the integration of CGH and array results to provide new experimental directions.

Availability: http://microarray.ym.edu.tw/tools/arrayfusion
Contact: hwwang@ym.edu.tw

1 INTRODUCTION

Microarray has been proven to be powerful on genome-wide experiments. Several types of array applications (such as RNA expression array, SNP array, exon array, CGH array and ChIP-on-chip) are developed and commercially available. The collective interpretation of experimental results from different types of array applications can sometimes yield novel research directions. For example, the combination of dynamic gene profiling data with the static ChIP-on-chip results forms the basis for the analysis of genetic network topology (Luscombe et al., 2004). Additional mapping of various types of array data onto their chromosomal locations also gives rise to new testable hypotheses: annotating SNP and gene expression microarray data onto their corresponding chromosomal locations led to the identification of LOH (loss of heterozygosity) regions and gene clusters, respectively (Lindblad-Toh et al., 2000; Mijalski et al., 2005), and LOH can then be used to explain the possible mechanism of gene expression changes. Therefore, tools for the recognition and incorporation of various types of microarray data from different platforms into one single genome-wide level and the presentation of integrated results on a chromosomal map will be biologically valuable.

Several tools were published for the integration of different types of genomic experiments. For example, MACAT (Toedling et al., 2005) and ChroCoLoc (Blake et al., 2006) were developed for the detection of gene clusters, dChipSNP for SNP LOH analysis (Lin et al., 2004), and eQTL for the integration of classical genetics information (Mueller et al., 2006). Most of them focus on one array application only. New tools capable of recognizing and summarizing more types of array platforms, including various home-made cDNA arrays, are still required.

There is also a need for the integration and visualization of classical CGH information in modern genome browsers. For decades CGH techniques have been applied for the study of genetic diseases and cancers, and a massive amount of CGH profiles are accessible in public databases (e.g. the Progenetix database, http://www.progenetix.de). Co-presentation of conventional CGH information with modern array data in genome browsers will expand the application of CGH results.

In this study we present ArrayFusion, a web application which can annotate and map different types of probe IDs onto genomic coordinates. ArrayFusion also supports query for cytological location, so it bridges the gaps between prior CGH records and current microarray data. The output results are converted into standard formats which can be visualized and explored in genome browsers. Converted files can be viewed together in a single genome browser, thereby assisting a multi-dimensional exploration between array results. ArrayFusion hence represents a value-added software layer that lies between a variety of data and several genome browsers to accelerate the discovery of new biological knowledge (Fig. 1).

2 IMPLEMENTATION

ArrayFusion is built in JaveServer Pages (JSP) by the implementation of Struts Action Framework and uses Apache Tomcat as the backend servlet container. The MVC (Model-View-Controller) Model 2 architecture is chosen as our design pattern with the intention of utilizing and separating each module into independent pieces, hence the whole system is flexible to grow and easy to maintain. Powered by MySQL database server, ArrayFusion maintains and comprises annotations from public databases (HGNCE and NCBI) and different chip suppliers (Affymetrix™ and Agilent™).

Currently the genome assembly for all human arrays is NCBI build 35. To ensure users can continuously analyze different datasets under the same assembly, we include a function to support the conversion between different human genome assemblies. This is achieved by the LiftOver’s chain conversion files from UCSC.

*To whom correspondence should be addressed.
Fig. 1. ArrayFusion recognizes CGH, cDNA and oligonucleotide microarray data from a variety of platforms, and generate converted formats which can be observed and evaluated in several genome browsers simultaneously, thereby benefiting a multi-dimensional exploration of array results to provide new experimental hypotheses.

Designed a database to store these chain files and developed a LiftOver-like module in Java to convert genome coordinates between assemblies.

As for user’s privacy, submitted IDs along with the output files are stored in the session, which is located in user’s local computer. No information is shared. Users can delete all of the output files from the ‘Query Results’ on the web interface.

Owing to the mobility of Java language, the whole software can be installed in different platforms without restraint. This has been tested in both Intel Pentium and AMD Athlon 64 CPUs on Windows 2003/XP, Fedora Linux 4, Red Hat Linux Enterprise 4 and SUSE Linux 9 operating systems.

3 USAGE AND DATA PRESENTATION

ArrayFusion can be accessed online freely without registration. A batch query function is available for NCBI cytological locations, HGNC gene symbols, GenBank or RefSeq mRNA accessions, Ensembl Gene IDs or commercial probe IDs. The required file format should be IDs separated by a new line. Users can just copy-and-paste gene or array identifiers in the text area or upload a file containing these data. Upon querying the database, annotations are generated by mapping queried IDs to their corresponding chromosomal locations. A few examples showing how ArrayFusion may aid in forming new biological hypotheses are available online.

In terms of gene symbol queries, the corresponding gene symbol may fall in either ‘Approved Symbol’, ‘Previous Symbols’ or ‘Aliases’ columns in the HGNC database. ArrayFusion will search sequentially until the corresponding gene symbol is found. After identified, ArrayFusion maps the genes to represent the RefSeq IDs. The mapping procedure follows the HGNC database design, starting from manually curated RefSeq IDs and then the Entrez Gene mapped entries. ArrayFusion also assigns the corresponding Affymetrix HG-U133 Plus 2.0 probeset IDs for queried gene symbols, so users may compare their cDNA array data with Affymetrix results.

The output formats include (1) a TXT tab-delimited annotation file, which includes combined information for queried IDs. For Affymetrix exon array annotations, ArrayFusion additionally parses the original ‘Gene Assignment’ columns from NetAffx Analysis Center (https://www.affymetrix.com/analysis/index.affx) and splits them into separate but handy columns; (2) an EGR format file for Affymetrix Integrated Genome Browser (IGB; http://www.affymetrix.com/support/development/tools/download_igb.affx). Data viewed in Affymetrix IGB can be further redirected to UCSC Genome Browser (http://genome.ucsc.edu/), which in turn expands the usage of ArrayFusion; (3) a GFF format file for Generic Genome Browser (GBrowse; http://www.gmod.org/gbrowse) applied by the HapMap project (http://www.hapmap.org), for UCSC Genome Browser, and for Ensembl’s KaryoView (http://www.ensembl.org/Homo_sapiens/karyoview). In addition to chromosome location information, haplotype information from the HapMap public data can also be considered together, thereby helping the interpretation of microarray data and the formation of new hypotheses. We recommend users to use IGB and/or GBrowse to start their analysis.

In summary, ArrayFusion can recognize and convert CGH records and various types of microarray data from different platforms into one single genome-wide level (Fig. 1), enabling a multi-dimensional interpretation of array data and the development of novel research hypotheses.

ACKNOWLEDGEMENTS

We thank Dr Chih-Hung Jen and Mr Chien-Yi Tung for critical reading of the manuscript and their inspiring comments. This work is supported by grants from the NRPGM office of the Nation Science Council (NSC), Taiwan (NSC-94-3112-B-010-015-Y) and in part by another grant from NSC (NSC-94-2321-B-010-013).

Conflict of Interest: none declared.

REFERENCES


