NetAffx™ Expression Array Comparison Tool: Instantaneous search of an expansive cross-species catalog

Gene Hsiao, Adam Tracy, Ron Shigeta, and Brant Wong
Affymetrix, Inc., Emeryville, CA 94608, USA

affymetrix
Biology for a better world

ABSTRACT

Affymetrix has an expansive and ever-growing catalog of more than 150 expression and DNA microarray products, spanning across more than 40 different genomes. We continue to support our catalog with rich annotations to reflect the current knowledge base of the scientific community, however yielding terabytes of annotation data. The sheer size and multi-dimensionality of this biological data makes online instantaneous search queries unachievable in a standard relational database framework.

We introduce here the NetAffx Expression Array Comparison Tool, built with the Apache Lucene and Solr indexing and search framework. This search engine technology together with client-side browser libraries allows prospective and current users to rapidly explore cross-sections of functional data relevant to their biological experiment. Users can simply input gene symbols, pathways, Gene Ontology (GO), or Medical Subject Heading (MeSH) terms into a single search box. In an instant, the tool returns tailored coverage statistics and an interactive visual interface where users can compare and contrast content. Additionally, the tool utilizes NCBI's Homologene homolog data enabling users to query across species thus aiding in translational studies.

The NetAffx Expression Array Comparison Tool is freely available at **http://www.affymetrix.com/estore/analysis/compare/index.affx**

Gene Annotations

For each NetAffx version release, annotations for the Affymetrix catalog begins with an en-masse download of **current transcript sequences from major public sources** including Genbank, Refseq, ENSEMBL, miRBase, lincRNAdb, and UCSC. These source data are analyzed and aggregated into representative transcript clusters from which array probes are sequence aligned with BLAT. The probesets are then associated with **up-to-date functional annotations** from sources including Entrez, Homologene, the GO consortium, UniProt, UniGene, OMIM, and GenMAPP pathways. This computational workflow encompasses a catalog of more than 150 catalog arrays spanning more than 40 species, and produces **terabytes of rich annotations** which then can be searched and mined.

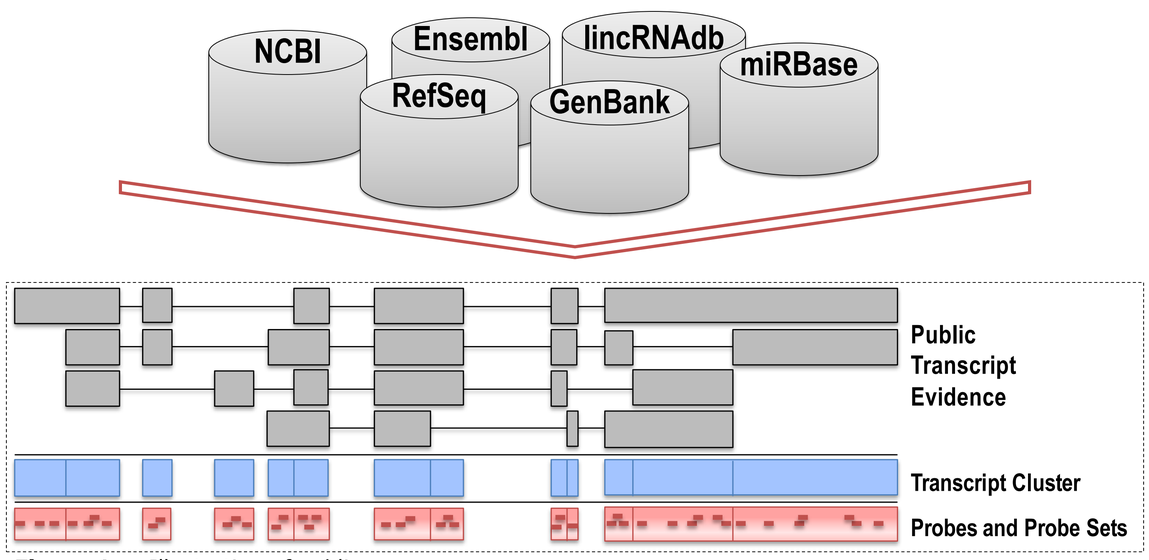


Figure 1.a: Illustration of public transcript evidence aggregated into transcript clusters. Probes and probesets are aligned and subsequently annotated.

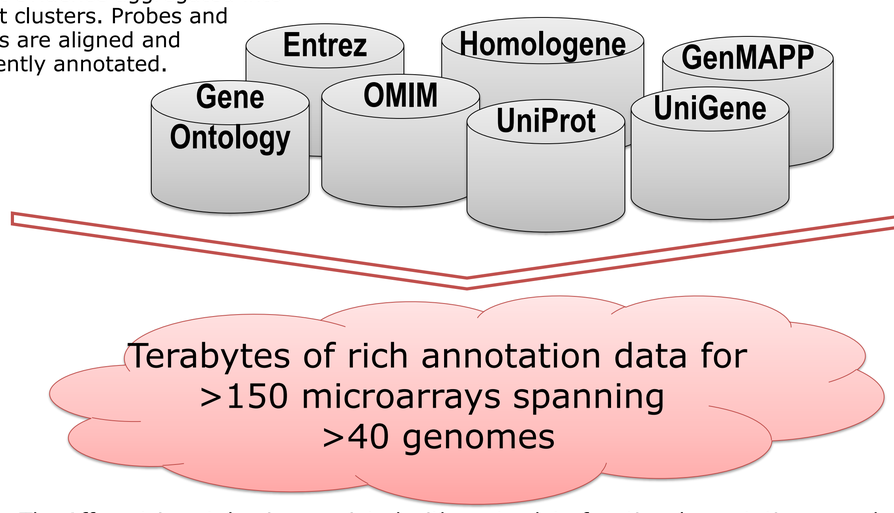


Figure 1: The Affymetrix catalog is associated with up-to-date functional annotations, producing massive amounts of data

Index Technology

To enable **instantaneous queries against terabytes of multi-dimensional biological annotation data**, we utilize the Apache Foundation sponsored open-source Lucene and Solr software framework[1]. The corpus of gene symbols, pathways, gene ontologies, homologene relationships, and other annotations data documents are analyzed and assembled into a Lucene inverted index. The Solr web application layer is tuned to reliably search, score, and facet documents that match a user's input query terms. These custom results are presented via client-side D3[2] and other javascript libraries for a visual and interactive user experience. This entire stack of search and aggregation is transparent to the user, such that the user need only to be concerned with their keywords of interest.

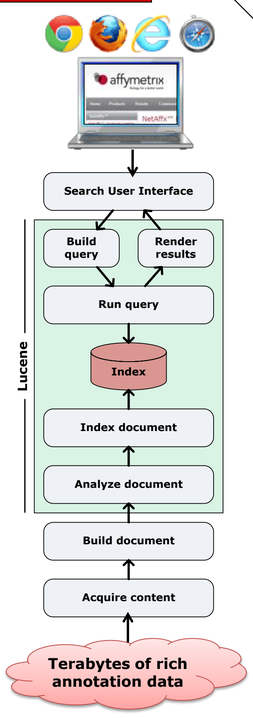


Figure 2, Lucene/Solr stack: Lucene indexing enables instant search of massive amounts of biological annotation data. Typical queries return in milliseconds

Instantaneous Search and Comparison

Search and compare content and biological coverage of the entire Affymetrix catalog of GeneChip® Expression arrays. Optionally input gene symbols, gene ontology, MeSH, and pathway terms for a tailored report; or, compare array content in their entirety.

Quickly compare arrays which span more than one species to answer which products best fit your cross-species translation study.

Drill down to individual genes and probesets for detailed information. Finally, Export your personalized results for offline viewing.

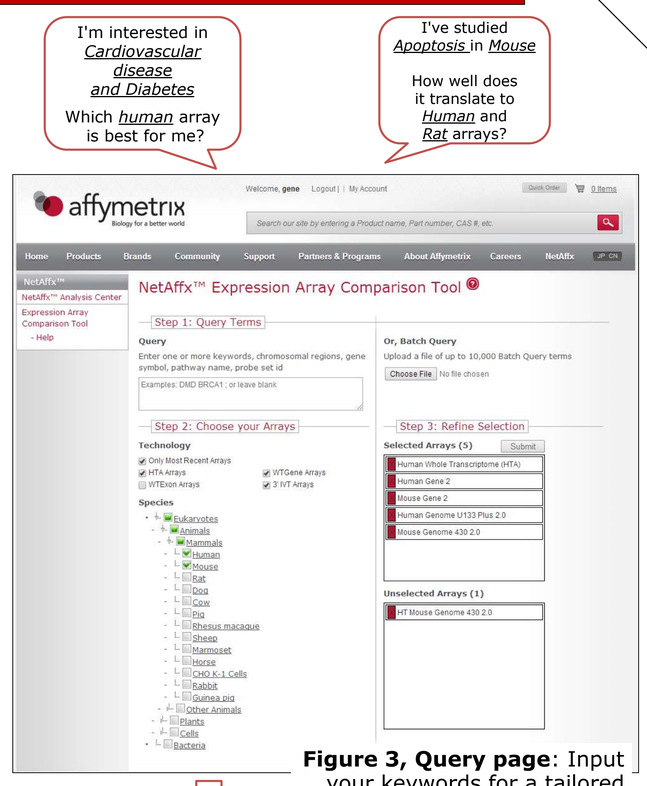


Figure 3, Query page: Input your keywords for a tailored array comparison report

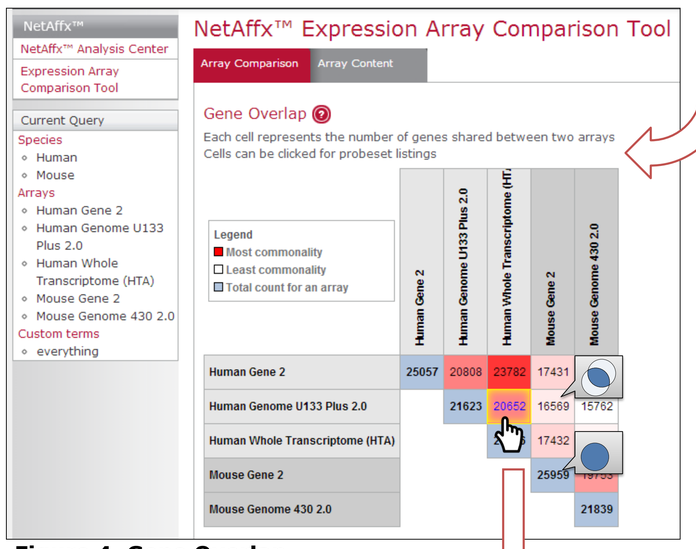


Figure 4, Gene Overlap Results: The heat map matrix displays the total number of genes represented by a single array, and the overlap between two arrays.

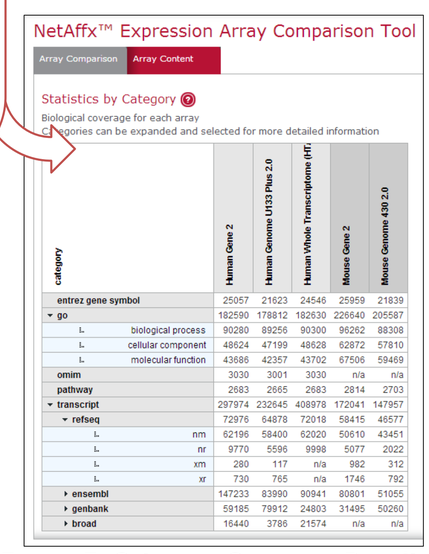


Figure 5, Category Coverage Results: Exact counts of annotation assignments for each array.

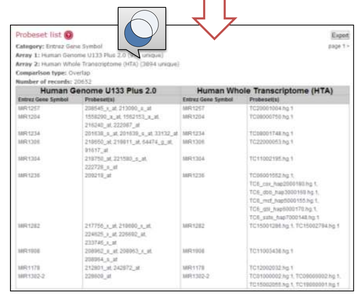
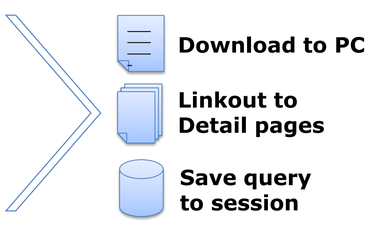


Figure 6, Detailed Report: Gene and probeset drill-down.



References:
[1] Apache Lucene. Web. 5 May 2014. <http://lucene.apache.org>
[2] Data-Driven Documents. Web. 5 May 2014. <http://d3js.org>
[3] NetAffx Analysis Center. Web. 5 May 2014. <http://www.affymetrix.com/estore/analysis/index.affx>
[4] Liu G., et al. *NetAffx: Affymetrix probesets and annotations*. Nucleic Acids Res 2003 Jan 1;31(1):82-6.
Acknowledgements: Special thanks to John Keefe, Michael Del Santo!