

Genome Biol. 2009;10(11):R130. Epub 2009 Nov 17.

## **BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources.**

Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, Hodge CL, Haase J, Janes J, Huss JW 3rd, Su AI.

Genomics Institute of the Novartis Research Foundation, 10675 John Jay Hopkins Dr, San Diego, CA 92121, USA. cwu@gnf.org.

**ABSTRACT** : Online gene annotation resources are indispensable for analysis of genomics data. However, the landscape of these online resources is highly fragmented, and scientists often visit dozens of these sites for each gene in a candidate gene list. Here, we introduce BioGPS <http://biogps.gnf.org>, a centralized gene portal for aggregating distributed gene annotation resources. Moreover, BioGPS embraces the principle of community intelligence, enabling any user to easily and directly contribute to the BioGPS platform.

BMC Genomics. 2009 Jan 14;10:22.

## **BioMart--biological queries made easy.**

Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A.

Ontario Institute for Cancer Research, MaRS Centre, 101 College Street, Toronto, Ontario, Canada. damian@ebi.ac.uk

**BACKGROUND**: Biologists need to perform complex queries, often across a variety of databases. Typically, each data resource provides an advanced query interface, each of which must be learnt by the biologist before they can begin to query them. Frequently, more than one data source is required and for high-throughput analysis, cutting and pasting results between websites is certainly very time consuming. Therefore, many groups rely on local bioinformatics support to process queries by accessing the resource's programmatic interfaces if they exist. This is not an efficient solution in terms of cost and time. Instead, it would be better if the biologist only had to learn one generic interface. BioMart provides such a solution. **RESULTS**: BioMart enables scientists to perform advanced querying of biological data sources through a single web interface. The power of the system comes from integrated querying of data sources regardless of their geographical locations. Once these queries have been defined, they may be automated with its "scripting at the click of a button" functionality. BioMart's capabilities are extended by integration with several widely used software packages such as BioConductor, DAS, Galaxy, Cytoscape, Taverna. In this paper, we describe all aspects of BioMart from a user's perspective and demonstrate how it can be used to solve real biological use cases such as SNP selection for candidate gene screening or annotation of microarray results.

CONCLUSION: BioMart is an easy to use, generic and scalable system and therefore, has become an integral part of large data resources including Ensembl, UniProt, HapMap, Wormbase, Gramene, Dictybase, PRIDE, MSD and Reactome. BioMart is freely accessible to use at <http://www.biomart.org>.

Cancer Inform. 2007 Feb 4;3:11-7.

## **Analysis of Gene Expression Data Using BRB-Array Tools.**

Simon R, Lam A, Li MC, Ngan M, Menenzes S, Zhao Y.

Biometric Research Branch, National Cancer Institute, 9000 Rockville Pike, Bethesda MD.

BRB-ArrayTools is an integrated software system for the comprehensive analysis of DNA microarray experiments. It was developed by professional biostatisticians experienced in the design and analysis of DNA microarray studies and incorporates methods developed by leading statistical laboratories. The software is designed for use by biomedical scientists who wish to have access to state-of-the-art statistical methods for the analysis of gene expression data and to receive training in the statistical analysis of high dimensional data. The software provides the most extensive set of tools available for predictive classifier development and complete cross-validation. It offers extensive links to genomic websites for gene annotation and analysis tools for pathway analysis. An archive of over 100 datasets of published microarray data with associated clinical data is provided and BRB-ArrayTools automatically imports data from the Gene Expression Omnibus public archive at the National Center for Biotechnology Information.

Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W449-56.

## **Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments.**

Khatri P, Bhavsar P, Bawa G, Draghici S.

Department of Computer Science, Wayne State University, 431 State Hall, Detroit, MI 48202, USA.

The Onto-Tools suite is composed of an annotation database and five seamlessly integrated web-accessible data mining tools: Onto-Express (OE), Onto-Compare (OC), Onto-Design (OD), Onto-Translate (OT) and Onto-Miner (OM). OM is a new tool that provides a unified access point and an application programming interface for most annotations available. Our database has been enhanced with more than 120 new commercial microarrays and annotations for *Rattus norvegicus*, *Drosophila melanogaster* and *Carnorhabditis elegans*. The Onto-Tools have been redesigned to provide better biological insight, improved

performance and user convenience. The new features implemented in OE include support for gene names, LocusLink IDs and Gene Ontology (GO) IDs, ability to specify fold changes for the input genes, links to the KEGG pathway database and detailed output files. OC allows comparisons of the functional bias of more than 170 commercial microarrays. The latest version of OD allows the user to specify keywords if the exact GO term is not known as well as providing more details than the previous version. OE, OC and OD now have an integrated GO browser that allows the user to customize the level of abstraction for each GO category. The Onto-Tools are available online at <http://vortex.cs.wayne.edu/Projects.html>

Nat Genet. 2006 May;38(5):500-1.

## **GenePattern 2.0.**

Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP.

Nat Protoc. 2009;4(1):44-57.

## **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.**

Huang da W, Sherman BT, Lempicki RA.

Laboratory of Immunopathogenesis and Bioinformatics, Clinical Services Program, SAIC-Frederick Inc., National Cancer Institute at Frederick, Frederick, Maryland 21702, USA.

DAVID bioinformatics resources consists of an integrated biological knowledgebase and analytic tools aimed at systematically extracting biological meaning from large gene/protein lists. This protocol explains how to use DAVID, a high-throughput and integrated data-mining environment, to analyze gene lists derived from high-throughput genomic experiments. The procedure first requires uploading a gene list containing any number of common gene identifiers followed by analysis using one or more text and pathway-mining tools such as gene functional classification, functional annotation chart or clustering and functional annotation table. By following this protocol, investigators are able to gain an in-depth understanding of the biological themes in lists of genes that are enriched in genome-scale studies.

Genome Res. 2005 Oct;15(10):1451-5. Epub 2005 Sep 16.

## **Galaxy: a platform for interactive large-scale genome analysis.**

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A.

Center for Comparative Genomics and Bioinformatics, Huck Institutes for Life Sciences, Penn State University, University Park, Pennsylvania 16802, USA.

Accessing and analyzing the exponentially expanding genomic sequence and functional data pose a challenge for biomedical researchers. Here we describe an interactive system, Galaxy, that combines the power of existing genome annotation databases with a simple Web portal to enable users to search remote resources, combine data from independent queries, and visualize the results. The heart of Galaxy is a flexible history system that stores the queries from each user; performs operations such as intersections, unions, and subtractions; and links to other computational tools. Galaxy can be accessed at <http://g2.bx.psu.edu>.

Bioinformatics. 2009 Jun 1;25(11):1363-9. Epub 2009 Apr 8.

## **CloudBurst: highly sensitive read mapping with MapReduce.**

Schatz MC.

Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA. [mschatz@umiacs.umd.edu](mailto:mschatz@umiacs.umd.edu)

**MOTIVATION:** Next-generation DNA sequencing machines are generating an enormous amount of sequence data, placing unprecedented demands on traditional single-processor read-mapping algorithms. CloudBurst is a new parallel read-mapping algorithm optimized for mapping next-generation sequence data to the human genome and other reference genomes, for use in a variety of biological analyses including SNP discovery, genotyping and personal genomics. It is modeled after the short read-mapping program RMAP, and reports either all alignments or the unambiguous best alignment for each read with any number of mismatches or differences. This level of sensitivity could be prohibitively time consuming, but CloudBurst uses the open-source Hadoop implementation of MapReduce to parallelize execution using multiple compute nodes. **RESULTS:** CloudBurst's running time scales linearly with the number of reads mapped, and with near linear speedup as the number of processors increases. In a 24-processor core configuration, CloudBurst is up to 30 times faster than RMAP executing on a single core, while computing an identical set of alignments. Using a larger remote compute cloud with 96 cores, CloudBurst improved performance by >100-fold, reducing the running time from hours to mere minutes for typical jobs involving mapping of millions of short reads to the human genome. **AVAILABILITY:** CloudBurst is available open-source as a model for parallelizing algorithms with MapReduce at (<http://cloudburst-bio.sourceforge.net/>).

PLoS Comput Biol. 2009 Jun;5(6):e1000369. Epub 2009 Jun 26.

## **Managing and analyzing next-generation sequence data.**

Richter BG, Sexton DP.

Enterprise Research IS and Informatics, Brigham and Women's Hospital, Massachusetts General Hospital, and Partners Healthcare, Boston, Massachusetts, United States of America. [brichter@partners.org](mailto:brichter@partners.org)

Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50. Epub 2005 Sep 30.

## **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.**

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP.

Broad Institute of Massachusetts Institute of Technology and Harvard, 320 Charles Street, Cambridge, MA 02141, USA.

Comment in:

Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15278-9.

Although genomewide RNA expression analysis has become a routine tool in biomedical research, extracting biological insight from such information remains a major challenge. Here, we describe a powerful analytical method called Gene Set Enrichment Analysis (GSEA) for interpreting gene expression data. The method derives its power by focusing on gene sets, that is, groups of genes that share common biological function, chromosomal location, or regulation. We demonstrate how GSEA yields insights into several cancer-related data sets, including leukemia and lung cancer. Notably, where single-gene analysis finds little similarity between two independent studies of patient survival in lung cancer, GSEA reveals many biological pathways in common. The GSEA method is embodied in a freely available software package, together with an initial database of 1,325 biologically defined gene sets.