

A gene set repository for model organisms

Josiah Altschuler¹, Eitan Rubin², Aviv Regev³

Keywords: gene sets, modules, expression compendia

1 Introduction.

Biological processes are coordinated by collections of gene products that act in concert to perform distinct functions. Recent research [1, 2, 3] has demonstrated that a higher order characterization of genome wide data, such as expression profiles, is both statistically more robust and often more biologically meaningful. Such characterization is typically initiated by combining a collection of gene sets (lists of genes that share a common attribute) with one or more genomics data sets - primarily compendia of expression profiles. While this approach is highly successful, it has been principally applied to date in a limited set of organisms. One of the limitations to rapid adoption of a gene-set level analysis by the wide biological community is the availability of an easily accessible resource of gene sets for a broad range of model organisms.

2 Software and files.

Here we present a publicly available online repository of gene sets (http://compbio.cgr.harvard.edu/genesets/gene_sets.htm) for ten model organisms, including bacteria, plants, yeasts, invertebrates and mammals. The repository provides access to two types of gene sets: Annotation-based and computationally-derived. Annotation-based gene sets are derived and regularly updated from public functional, pathway and genomics resources. These gene sets represent the biological community's current assignment of genes to functional and anatomical categories (based on the GO ontology), molecular pathways (KEGG, BioCyc, BioCarta and SuperArray resources), molecular complexes (GRID, MIPS and BIND), targets of transcription factors (based on ChIP-chip experiments [4, 5, 6]), cellular location [7], and cellular and clinical phenotype (Genetic Association DB [8], and genetic footprinting [9]).

To complement this knowledge-based view of biological processes, our repository includes a growing number of computationally-derived gene sets, automatically generated from genomics data. Here, we expect a shared molecular signature to reflect shared function. For example, such gene sets would include clusters of co-expressed genes, genes that share the same *cis*-regulatory element(s) in their promoters, genes whose expression changes in response to the same mutations [10], genes that encode for proteins with a shared domain [11], or genes that constitute a densely interconnected component in a molecular network (e.g. [12], [13] or [14])

The gene set files are updated regularly and automatically. An easy-to-use interface provides access to files in a simple format that facilitates input into any subsequent analysis algorithm. A collection of links to tools and sites that allow gene set level analysis is included in the repository, and extensive

¹ Bauer Center for Genomics Research, Harvard University, 7 Divinity Ave, Cambridge, MA, USA
E-mail: jaltschuler@cgr.harvard.edu

² Bauer Center for Genomics Research, Harvard University, 7 Divinity Ave, Cambridge, MA, USA
E-mail: erubin@cgr.harvard.edu

³ Bauer Center for Genomics Research, Harvard University, 7 Divinity Ave, Cambridge, MA, USA
E-mail: aregev@cgr.harvard.edu

expression compendia will be made available in the future as well.

Finally, an online form is available to allow users to request the addition of gene sets from additional resources or datasets. By providing a comprehensive library of gene sets to help researchers in analyzing their genomics data set of interest, we hope to help transform the analysis of genomics data from a gene-centric to a process-centric perspective.

3 References.

- [1] Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC. 2003. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet.* 34(3):267-73.
- [2] Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet.* 34(2):166-76.
- [3] Lamb J, Ramaswamy S, Ford HL, Contreras B, Martinez RV, Kittrell FS, Zahnow CA, Patterson N, Golub TR, Ewen ME. 2003. A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer. *Cell.* 114(3):323-34.
- [4] Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature.* 431(7004):99-104.
- [5] Basehoar AD, Zanton SJ, Pugh BF. 2004. Identification and distinct regulation of yeast TATA box-containing genes. *Cell.* 116(5):699-709.
- [6] Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, Wheeler R, Wong B, Drenkow J, Yamanaka M, Patel S, Brubaker S, Tammana H, Helt G, Struhl K, Gingeras TR. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell.* 116(4):499-509.
- [7] Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK. 2003. Global analysis of protein localization in budding yeast. *Nature.* 425(6959):686-91.
- [8] Becker, K.G., Barnes, K.C., Bright, T.J. & Wang, S.A. 2004. The Genetic Association Database. *Nature Genetics.* 36, 431-432.

- [9] Dunn B, Ferea T, Spellman P, Schwarz J, Terraciano J, Troyanovich J, Walker S, Greene J, Shaw K, DiDomenico B, Wang Q, Kaloper M, Metzner S, Chung E, Bondre C, Venteicher A, Botstein D, Brown P. 2004. Genetic footprinting: A functional analysis of the *S. cerevisiae* genome. *SGD Curated Paper, in preparation*.
- [10] Hughes TR, Roberts CJ, Dai H, Jones AR, Meyer MR, Slade D, Burchard J, Dow S, Ward TR, Kidd MJ, Friend SH, Marton MJ. 2000. Widespread aneuploidy revealed by DNA microarray expression profiling. *Nat Genet.* 25(3):333-7.
- [11] Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L, Copley R, Courcelle E, Das U, Durbin R, Fleischmann W, Gough J, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McDowall J, Mitchell A, Nikolskaya AN, Orchard S, Pagni M, Ponting CP, Quevillon E, Selengut J, Sigrist CJ, Silventoinen V, Studholme DJ, Vaughan R, Wu CH. 2005. InterPro, progress and status in 2005. *Nucleic Acids Res.* 33.
- [12] Kanehisa M. The KEGG database. 2002. *Novartis Found Symp.* 247:91-101; discussion 101-3, 119-28, 244-52. Review.
- [13] Karp PD, Riley M, Saier M, Paulsen IT, Paley SM, Pellegrini-Toole A. 2000. The EcoCyc and MetaCyc databases. *Nucleic Acids Res.* 28(1):56-9.
- [14] Segre D, Deluna A, Church GM, Kishony R. 2005. Modular epistasis in yeast metabolism. *Nat Genet.* 37(1):77-83.