v1.1 long

GeneMerge— post-genomic analysis, data mining, and hypothesis testing

Cristian I. Castillo-Davis and Daniel L. Hartl

Department of Organismic and Evolutionary Biology, Harvard University, Biological Laboratories, 16 Divinity Avenue, Cambridge, MA 02138, USA

Correspondence:

Cristian I. Castillo-Davis

Department of Organismic and Evolutionary Biology

Harvard University

16 Divinity Avenue

Cambridge, MA 02138, USA

ccastillo-davis@oeb.harvard.edu

(617) 496-5540

ABSTRACT

Motivation: In the face of ever-growing genomic and proteomic data researchers are shifting their attention to post-genomic analysis— the interpretation and synthesis of thousands of data points from a chemical, biological, or evolutionary perspective. While freely available through public databases, different sets of genomic data are often difficult to integrate into a given study because no common platform exists for such analysis. This problem is compounded by the continual release of new genomic and proteomic datasets. Simple and flexible software that can take advantage of diverse genomic and proteomic data for both data mining and hypothesis testing is required.

Results: A program to analyze a range of genomic and proteomic data is presented called GeneMerge. Given a set of study genes, GeneMerge retrieves functional genomic data for each gene and provides statistical rank scores for over-representation of functions or categories within the set of study genes. GeneMerge can perform analyses on a wide variety of data quickly and easily and facilitates both data mining and hypothesis testing. Two datasets are analyzed to illustrate the power and flexibility of GeneMerge. First, we perform a functional analysis of published yeast microarray expression data from *Snf/Swi* deletion mutants. Second, using publicly available deletion viability data we test the population genetic hypothesis that "dispensable" genes are more likely to be polymorphic in natural yeast populations.

Availability: GeneMerge is available over the web and for download free of charge for academic use from: http://www.oeb.harvard.edu/hartl/lab/publications/GeneMerge.html.

Supplementary Material:

http://www.oeb.harvard.edu/hartl/lab/publications/GeneMerge.html

INTRODUCTION

As genomic technology and protocols improve, researchers will be dedicating more and more of their time to post-genomic analysis. Functional and categorical genomic analysis has already become as important as the identification of differentially expressed genes, especially in fields where more than a list of candidate genes is desired.

A great deal of genomic and proteomic information is available. For many genes, something is known about their molecular and biological function, pathway membership, physical chromosomal location, level of polymorphism, RNAi phenotypes, disease phenotypes, and rate of molecular evolution. For non-coding regions, data are often available concerning the presence of known or putative transcription binding sites, levels of DNA methylation or acetylation, and GC content.

Unfortunately, this information is often difficult to integrate into a given genomic analysis since it is available in disparate and complicated formats from multiple sources; no common platform exists for analysis of such data. Ideally, genomic information should be available for analysis within a unified framework where sets of genes from any experiment can be interrogated easily. Currently, such a framework does not exist. We have developed software in order to meet this need called GeneMerge.

GeneMerge is a web-based or stand-alone program that returns functional genomic data of all kinds (Table 1) for a given set of study genes and provides rank scores for over-representation of particular functions in the dataset, if present. A study set of genes may comprise genes significantly up-regulated in a microarray experiment, genes identified through chromatin immunoprecipitation (ChIP), or a collection of

fast-evolving genes identified through direct sequencing. GeneMerge can be used to analyze such genes from a functional, evolutionary, biochemical, genetic, or clinical perspective, and can easily incorporate future genomic and proteomic data.

Are functional pathways up-regulated during bacterial infection also up-regulated during fungal infection? What are they? Are fast evolving genes preferentially located in areas of high recombination? Do co-regulated genes show non-random enrichment for a certain family of transcription factors? These and other questions can be answered quickly and easily with GeneMerge.

While other programs provide portions of GeneMerge's capability, none can utilize as broad and varied a set of genomic and proteomic data, and none can easily incorporate recently available data. For example, dCHIP (Li and Wong, 2001) returns uncorrected P-values for high-level (most specific function only) GO ontology terms. Pathway Processor (Grosu *et al.*, 2002) provides an uncorrected P-value for enrichment in a particular metabolic pathway. As will be seen, GeneMerge carries out the analyses of these programs and more, including analyses utilizing the gene-association data in Table 1 and beyond. The simplicity and extensibility of GeneMerge means that it can be used on almost any genomic or proteomic data set. Here we analyze published and unpublished array data from *Saccharomyces cerevisiae* to illustrate the flexibility and functionality of the GeneMerge framework.

Snf/Swi **deletion experiment**

Using DNA micorarrays, Sudarsanam *et al.* (2000) examined gene expression in yeast carrying deletions in the *Snf/Swi* complex which has been implicated in chromatin

remodeling and transcription activation. Analysis of gene expression levels was performed on both rich and minimal media, however, no genome-scale functional analysis was carried out on the differentially expressed genes identified. Here, using GeneMerge, we carry out a functional and pathway analysis of genes up-regulated in *Snf/Swi* mutants grown on both rich and minimal media. Several potentially interesting biological processes and pathways are identified.

Polymorphism and deletion viability

Recent analysis of functional genomic data suggests that protein evolution is related to protein effects on organismic fitness; specifically, it has been shown that proteins that cause lowered fitness when deleted in yeast (so-called "non-dispensible" genes) tend to evolve more slowly (Hirsh and Fraser, 2001). While amino acid substitution rates tend to be higher in "dispensable" genes over long evolutionary distances (Hirsh and Fraser, 2001), it is not known whether these genes also tend to be more polymorphic in natural populations. Given that selection against deleterious mutations is expected to operate at the population level, coupled with the observation that variation among natural populations is ultimately transformed into variation among species, it is predicted that dispensable genes will be more polymorphic within populations. Using GeneMerge we perform a simple test of this hypothesis, that population-level polymorphism is more likely to occur in dispensable genes.

SYSTEM AND METHODS

Overview

GeneMerge returns descriptive information regarding genes under investigation and statistically-based rank scores regarding potential over-representation of descriptors in a given set of genes. Functional or categorical descriptive data is associated with genes in *gene-association* files. These text files link each gene in a genome with a particular datum of information. For example, the name of a gene and its chromosomal location, molecular function, or its identity as over-expressed in a particular type of cancer. Some currently available gene-association files are listed in Table 1. Full information on gene-association files is available in the Supplementary Material.

Gene-association data are many and varied and will undoubtedly grow as genomic and proteomic investigations accelerate (Table 1). To deal with this explosion of data requires both a clear analytical framework and the flexibility to incorporate new data as soon as they become available. GeneMerge addresses the first requirement by performing a simple statistical test to answer a straightforward question, Are particular functions or categories over-represented in the study dataset? GeneMerge addresses the second requirement, easy incorporation of newly available data, with its simple gene-association file format. GeneMerge gene-association files are easy to create such that almost any worker can generate an association file for use in their study. This means that as new information about genes and proteins is generated (publicly or privately) it can be quickly and easily incorporated into an analysis.

GeneMerge takes 4 input files:

1. study set gene file
2. population set gene file
3. gene-association file
4. description file

The study set is comprised of genes that are currently under investigation. The population set is comprised of those genes from which the study set was drawn, often a genome. The gene-association file links gene names with a particular datum of information using a shorthand identifier (ID). Finally, the description file contains human-readable descriptions of gene-association IDs.

For example, in a microarray experiment, the *study set* of genes may be those that are significantly up-regulated; the *population set* is thus comprised of all genes detected on the array. The *gene-association file* and *description file* depend on the analysis being performed. For example, to analyze chromosomal clustering in the study set of genes, one would use a chromosomal location gene-association file and its associated description file.

Output is a tab-delimited text file that can be opened in most spreadsheet programs. It contains functional or categorical data associated with each gene in the study set and rank scores for over-represented functions/categories, as well as other pertinent data (see Supplementary Material).

Statistics

Rank scores for functional or categorical over-representation within the study set of genes is obtained using the hypergeometric distribution (1). The hypergeometric distribution describes the discrete probability of selecting *r* items of one kind in a sample of size *k* from a population of size *n*, where *p* is equal to the proportion of *r*-type items in the population, and sampling is without replacement (Sokal & Rohlf, 1995).

$$\Pr(r \mid n, p, k) = \frac{\binom{pn}{r}\binom{(1-p)n}{k-r}}{\binom{n}{k}} \quad (1)$$

The hypergeometric thus gives a quantification of the level of one's "surprise" at finding over-representation for a particular item in a given sample of size *k* drawn from a larger population, size *n*. In GeneMerge, *k* is always the study set of genes and *n* is the population set, the set from which k is drawn, usually a genome or all genes on a particular DNA array. The study set *k* may be genes found to be significantly up or down-regulated in a microarray experiment or a list of genes deemed interesting for another reason. Genes in the sample *k* are associated with particular identifiers, for example functions, processes, or states. The number of genes with a particular identifier is *r*. *p* is the fraction of genes in the population *n* associated with the particular identifier under investigation. The hypergeometric gives the exact probability of drawing *r* genes with a particular identifier from a sample of size *k* from a population of size *n* given that the identifier exists in fraction *p* in the population set of genes.

In the case of the *Snf/Swi* data, n is the number of valid genes on the array and k is the number of genes over-expressed by the *Snf/Swi* deletion mutant. k is considered the study set and r is the number of genes among the study set that have a particular GO or KEGG term associated with them. For example, 10 of the 34 genes up-regulated in *Snf/Swi* mutants grown on rich media are involved in amino acid metabolism (GO term GO:0006520), thus $r = 10$. p is the proportion of genes on the array with this particular GO term. Since 0.013 of genes on the array are involved in amino acid metabolism, $p = 0.013$. Summing over the tail of the hypergeometric for all less likely cases, yields a P-value of 1.23×10^{-11} for over-representation of the term “amino acid metabolism” in the study set.

Because GeneMerge assesses over-representation for all categories within a given study set of genes, a correction is necessary to account for over-representation that will invariably occur by chance when multiple tests are carried out. A strict correction for multiple tests is the Bonferroni correction (Sokal and Rohlf, 1995). Here we use a modified Bonferroni correction based on the number of terms examined in each analysis.

Not all terms associated with genes are scored since they may represent "singletons" in either the study set or the population set. For instance the term "saccharopine dehydrogenase (NADP+, L-glutamate forming)" (GO:0004755), is associated with only one gene in the *Saccharomyces* genome (population set). Likewise, the molecular function "protein-methionine-S-oxide reductase" (GO:0008113) is associated with only one gene among those up-regulated in the *Snf/Swi* study set. In such cases, over-representation of the particular function or category is not possible and over-representation scores are not calculated.

It might be argued that such terms should not contribute to the Bonferroni correction since scores for these terms are not calculated— in effect no "tests" are carried out. However, there is a crucial distinction between population and study set singletons. In the former case, over-representation is a logical impossibility and scoring is ruled out before the analysis begins. In the latter case, the possibility of over-representation is contingent, and the decision to score or not to score is made only after examining the data. Therefore, even though scores for terms that appear only once in the *study set* of genes are not calculated, we apply them conservatively to the Bonferroni correction term since their ex post facto exclusion is not blind.

Uncorrected and corrected scores are called raw e-scores (e_r) and e-scores (e_s) respectively, as a reminder that, in some cases, these values will not reflect true P-values. In some types of association data, a gene may belong to multiple categories simultaneously, for example, in the case of genetic pathway data, a gene may often function in several biochemical cascades. In such cases, e-scores will not correspond to P-values. However, for one-to-one gene association data, which make up a majority of association data (chromosomal location data, GO data, deletion viability data, etc.), e-scores correspond exactly to P-values.

Snf/Swi functional and metabolic pathway analysis

Genes over-expressed in *Snf/Swi* deletion mutants grown on rich and minimal media were identified from the raw data of Sudarsanam *et al.* (2000) using the program Bayesian Analysis of Gene Expression Levels (BAGEL) (Townsend *et al.*, submitted) implementing the criterion of non-overlapping 95% credible intervals among

differentially expressed genes. Functional and categorical data for molecular function, biological process, and cellular component categories were obtained from the Gene Ontology Consortium (2000) database (http://www.geneontology.org/). Metabolic pathway data were obtained from the Kyoto Encyclopaedia of Genes and Genomes database (Kanehisa *et al.*, 2002) (http://www.genome.ad.jp/kegg/).

Data in GO are organized as a directed acyclic graph and include both high-level (most-specific) and low-level (more general terms). For example the gene *YAT1* is annotated as a carnitine O-acetyltransferase, its most specific annotation. However, carnitine O-acetyltransferase is also a specific instance of the more general category "O-acetyltransferase", which is an instance of the category "acetyltransferase", etc., up through to the very general category "transferase" and ultimately to the most general biological category "enzyme." Gene-association files of both high-level terms and complete-level terms (all levels for every gene) were constructed and used for functional analyses.

Deletion viability and polymorphism analysis

Polymorphic genes were identified using genomic hybridizations to Affymetrix S98 oligonucleotide arrays containing 126,645 unique 25mer yeast probes to identify polymorphisms among 14 strains of laboratory and natural yeast (Winzeler *et al.*, submitted). The array is sensitive to the detection of single nucleotide polymorphisms. Unfortunately, distinction between synonymous and nonsynonymous substitutions is not possible with these arrays. Genes with at least one detected polymorphism among the 14 strains were considered polymorphic. Genes with no polymorphism were considered non-

polymorphic. Among the 2991 genes probed on the chip, 1874 were polymorphic by this criterion. To create a deletion-viability gene-association file, lists of genes that result in inviablity or are viable when deleted were obtained from the Saccharomyces Genome Database (http://genome-www.stanford.edu/Saccharomyces/) based on the data of Winzeler *et al.* (1999) and the Giaever *et al.* (2002). 4713 genes were listed as having a deletion-viable phenotype and 1115 genes an inviable deletion phenotype. 413 genes had no data available concerning deletion phenotype.

IMPLEMENTATION

GeneMerge has been implemented in PERL as a stand-alone program that can be run from a command line interface and as a web-based package that can be run on a web server with a simple HTML GUI. Datasets can be analyzed over the web at http://www.oeb.harvard.edu/hartl/lab/publications/GeneMerge.html

DISCUSSION

Snf/Swi

Examination of molecular function, biological process, cellular component, and pathway data of genes up-regulated in *Snf/Swi* mutants in both rich and minimal media yielded several interesting results. First, *Snf/Swi* deletion mutants grown on minimal media showed up-regulation and significant over-representation of genes involved in protein biosynthesis (82 of 473 genes) and nitrogen starvation response (4 of 473 genes)

(Table 2). In contrast, *Snf/Swi* mutants grown on rich media showed up-regulation of genes involved in amino acid metabolism, in particular methionine metabolism and arginine biosynthesis (Table 3). Complete results for all three GO categories, "molecular function", "biological process", and "cellular component" are available at http://www.oeb.harvard.edu/hartl/lab/publications/GeneMerge.html.

Results based on different sets of gene-association data were often corroboratory in nature. For example, in *Snf/Swi* minimal media mutants, over-representation of the biological process "nitrogen starvation response" (GO identifier, GO:0006995) was indicated when using GO gene-association data (Table 3). A separate analysis of metabolic pathways using KEGG gene-annotation data indicated over-representation of genes involved in "nitrogen metabolism"(KEGG pathway ID, sce00910) (Table 2). In the former case, four genes in the data set were associated with the GO biological process "nitrogen starvation response", whereas in the latter, eight genes (the same four plus four additional genes) were associated with "nitrogen metabolism" according to KEGG annotation. The ability to interrogate the data from multiple perspectives with GeneMerge makes such corroboration possible and may help lend support to hypotheses under investigation.

Both high-level (most specific term only) and complete-level analysis of GO terms yielded similar results (Table 2 & 3). However, the more general search provided more biologically useful information since more gene-association was considered. For example, in the high-level GO analysis for *Snf/Swi* deletion mutants grown on rich media (Table 3), three biological process terms are identified as enriched: methionine metabolism, arginine biosynthesis, and sulfate assimilation. In the complete-level

analysis, 16 processes are reported as significantly enriched and include genes not identified as significant in the previous analysis.

For instance, 7 genes are listed under "amino acid biosynthesis" (Table 3), three of which are not involved in arginine biosynthesis. These genes, YER081W, YCR005C, YGL125W, are involved in serine, glutamate, and sulfur amino acid biosynthesis, respectively. Complete-level analysis reveals that YER069W is also involved in arginine biosynthesis raising the number of genes in this category to 4. In addition, enrichment of a new, potentially interesting biological function, urea cycle intermediate metabolism, is revealed by complete-level analysis.

One potential cost associated with undertaking complete-level analysis with GO association data and other gene-association data like it results from the large number of scores examined. While providing a much richer biological picture, these multiple tests can result in large Bonferroni corrections that may obscure marginally significant results that appear in a high-level analysis. For example, the e-score for the term "sulfate assimilation" (Table 3) changes from 0.014 to 0.058 in the complete-level analysis. At the same time, clearly significant functions and categories tend to remain so under complete analysis. For example the e-score for "protein biosynthesis" in *Snf/Swi* minimal media mutants identified with 82 of 473 up-regulated genes (Table 2) changes from 1.00×10^{-35} to 1.57×10^{-34}. Note also that P-values for hierarchically nested terms are not independent. The discretion of the investigator is required to weigh the cost of potential loss of statistical power against the benefits of increased information for these types of gene-association data.

Deletion viability and polymorphism

The hypothesis that population-level polymorphism is more likely to occur in dispensable genes is supported by the data. Among *S. cerevisiae* genes categorized as polymorphic, more are viable upon deletion than is expected by chance. Of the 1874 genes categorized as polymorphic, 1454 were deletion viable, representing an enrichment in this class of genes ($P < 0.005$) (Supplementary Material). Thus selection against deleterious mutations in potentially more "important" genes appears to result in visibly lower levels of polymorphism at the population level. While this result is preliminary and would benefit from further investigation, it provides an excellent example of how GeneMerge can be used to test a novel hypothesis quickly and easily.

Summary

GeneMerge provides a simple and powerful platform from which to interrogate a range of genomic data. Functional and metabolic pathway analysis, tests of population genetic hypotheses, cross-experiment comparisons, tests of chromosomal clustering and much more are possible with GeneMerge. Functional analysis of *Snf/Swi* deletion data (Sudarsanam *et al.*, 2000) yielded immediate and biologically relevant results not reported in the original investigation and demonstrates the power of cross-data corroboration. Further, utilizing publicly available deletion data in conjunction with known polymorphic genes, GeneMerge facilitated a rapid test of an explicit, evolutionary hypothesis. While GeneMerge is not meant to be a total solution to genomic data mining and hypothesis testing, it is a robust statistical tool with which to perform analyses on a wide variety of data, quickly and easily. This flexibility is paramount as the quantity of

genomic and proteomic data expands at ever-increasing rates. The simplicity of GeneMerge insures that despite a deluge of data, data can be explored and hypotheses tested at a genomic level by any worker.

ACKNOWLEDGEMENTS

C.I.C-D would like to thank Peter Bouman, Mark Smith, Jeff Townsend, Alex Platt, and Yun Tao for statistical discussion; David Emmert, Han Xie, Midori Harris, Aubrey de Grey, Xin Lu, Paul Grosu, and Chris Dagdigian for providing files and technical help; Jose Ranz for beta-testing early versions of the program, Dan Neafsey for reading the manuscript, and Jun Liu, Erin Conlon, X. Shirely Liu, Laura Garwin, Debbie Marks, and Andrew Murray for support and encouragement. Special thanks to Josh Cherry for help in software optimization.

REFERENCES

Cherry, J. M., Ball, C., Dolinski, K., Dwight, S., Harris, M., Matese, J. C., Sherlock, G., Binkley, G., Jin, H., Weng, S., and Botstein, D. "Saccharomyces Genome Database" ftp://genome-ftp.stanford.edu/pub/yeast/SacchDB/ (July 18, 2002).

The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. 2000. *Nat. Genet.* 25: 25-29

Giaever G., Chu A.M., Ni L., Connelly C., Riles L., Veronneau S., Dow S., Lucau-Danila A., Anderson K., Andre B., Arkin A.P., Astromoff A., El Bakkoury M., Bangham R., Benito R., Brachat S., Campanaro S., Curtiss M., Davis K., Deutschbauer A., Entian K.D., *et al.* 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature.* 418: 387-391.

Grosu P., Townsend, J.P., Hartl, D.L., Cavalieri, D. 2002. Pathway Processor: a tool for integrating whole-genome expression results into metabolic networks. *Genome Res.* 12: 1121-1126.

Hirsh, A.E., H.B. Fraser. 2001. Protein dispensability and rate of evolution. *Nature.* 411:1046-1049.

Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. 2002. The KEGG databases at GenomeNet. *Nucleic Acids Res.* 30: 42-46

Li, C. and W. Hung Wong. 2001. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol.* 2(8): RESEARCH0032.

Sokal R.R., F.J. Rohlf. 1995. *Biometry: The Principles and Practice of Statistics in Biological Research.* W.H. Freeman and Co., New York, Third Edition.

Sudarsanam P., Iyer, V.R., Brown, P.O., Winston, F. 2000. Whole-genome expression analysis of snf/swi mutants of *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci.* 97: 3364-3369

Townsend, J.P, D.L. Hartl. 2002. Bayesian Analysis of Gene Expression. *Genome Biol.* (Submitted)

Winzeler, E.A., Castillo-Davis, C.I., Oshiro, G., Liang, D., Richards, D.R., Zhou, Y., Hartl, D.L. 2002. Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays. *Genetics* (submitted).

Winzeler E.A., Shoemaker D.D., Astromoff A., Liang H., Anderson K., Andre B., Bangham R., Benito R., Boeke J.D., Bussey H., Chu A.M., Connelly C., Davis K., Dietrich F., Dow S.W., El Bakkoury M., Foury F., Friend S.H., Gentalen E., Giaever G., Hegemann J.H., Jones T., Laub M., Liao H., Davis R.W., *et al.* 1999. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science.* 285: 901-906.

Gene-association Data
· deletion viability†
· alternative splicing†
· KEGG metabolic pathway†
· KEGG developmental pathway†
· GO molecular function†
· GO biological process†
· GO cellular component†
· RNAi phenotype†
· chromosomal location†
· knock-out phenotype
· disease phenotypes
· polymorphic / non-polymorphic locus
· local recombination rate
· transcription binding site
· transcription binding site family
· DNA methylation
· acetylation
· GC content
· male specific
· female specific
· ortholog in clade X
· rate of molecular evolution
· tissue-specific expression
· over/under-expression in experiment X

Table 1

Publically available† gene-association data included in Supplementary Material and includes data for the following species: *D. melanogaster*, *C. elegans*, *S. pombe*, *O. sativa*, *H. sapiens*, *S. cerevisiae*, *M. musculus*, *A. thaliana*, *V. cholerae*, and *R. norvegicus*.

Swi/Snf minimal media – GO Biological Process (high-level)					
GMRG Term	**Pop frac**	**Study frac**	**e_r**	**e_s**	**Description**
GO:0006412	0.0352	82/473	9.09E-38	1.00E-35	protein biosynthesis
GO:0006530	0.0007	4/473	3.37E-5	0.004	asparagine catabolism
GO:0006995	0.0007	4/473	3.37E-5	0.004	nitrogen starvation response
GO:0006096	0.0039	9/473	3.82E-5	0.004	glycolysis
GO:0007532	0.0011	5/473	4.72E-5	0.005	mating-type specific transcriptional control
GO:0006360	0.0040	9/473	5.56E-5	0.006	transcription, from Pol I promoter
GO:0006094	0.0031	7/473	3.266E-4	0.036	gluconeogenesis
GO:0000154	0.0023	6/473	3.413E-4	0.038	rRNA modification
GO:0006333	0.0019	5/473	0.001	0.142	chromatin assembly/disassembly

Swi/Snf minimal media – GO Biological Process (complete-level)					
GMRG Term	**Pop frac**	**Study frac**	**e_r**	**e_s**	**Description**
GO:0006412	0.0441	91/473	4.89E-37	1.57E-34	protein biosynthesis
GO:0009059	0.0519	93/473	2.61E-32	8.42E-30	macromolecule biosynthesis
GO:0009058	0.0777	114/473	6.32E-31	2.04E-28	biosynthesis
GO:0006411	0.1236	105/473	3.11E-10	1.00E-7	protein metabolism and modification
GO:0008151	0.3759	234/473	2.85E-8	9.17E-6	cell growth and/or maintenance
GO:0008152	0.2749	179/473	2.20E-7	7.09E-5	metabolism
GO:0006528	0.0006	4/473	3.37E-5	0.011	asparagine metabolism
GO:0006530	0.0006	4/473	3.37E-5	0.011	asparagine catabolism
GO:0006995	0.0006	4/473	3.37E-5	0.011	nitrogen starvation response
GO:0009065	0.0011	5/473	4.72E-5	0.015	glutamine family amino acid catabolism
GO:0007532	0.0011	5/473	4.72E-5	0.015	mating-type specific transcriptional control
GO:0007530	0.0011	5/473	4.72E-5	0.015	sex determination
GO:0007531	0.0011	5/473	4.72E-5	0.015	mating-type determination
GO:0006360	0.0162	20/473	5.22E-5	0.017	transcription, from Pol I promoter
GO:0006096	0.0040	9/473	5.56E-5	0.018	glycolysis

Swi/Snf minimal media – KEGG Pathway					
GMRG Term	**Pop frac**	**Study frac**	**e_r**	**e_s**	**Description**
sce03010	0.0221	80/473	2.59E-55	1.58E-53	Ribosome
sce00910	0.0028	8/473	1.44E-5	0.001	Nitrogen metabolism
sce00010	0.0074	12/473	1.22E-4	0.007	Glycolysis / Gluconeogenesis
sce03020	0.0045	8/473	0.001	0.053	RNA polymerase
sce00071	0.0029	6/473	0.002	0.099	Fatty acid metabolism

Table 2

Swi/Snf rich media – GO Biological Process (high-level)					
GMRG Term	**Pop frac**	**Study frac**	**e_r**	**e_s**	**Description**
GO:0006555	0.0019	4/34	3.65E-7	6.20E-6	methionine metabolism
GO:0006526	0.0010	3/34	3.00E-6	5.10E-5	arginine biosynthesis
GO:0000103	0.0013	2/34	0.001	0.014	sulfate assimilation
GO:0000004	0.1101	2/34	0.902	1.000	biological_process unknown
GO:0006067	0.0002	1/34	NA	NA	ethanol metabolism

Swi/Snf rich media – GO Biological Process (complete-level)					
GMRG Term	**Pop frac**	**Study frac**	**e_r**	**e_s**	**Description**
GO:0006520	0.0137	10/34	1.40E-11	1.00E-9	amino acid metabolism
GO:0006519	0.0149	10/34	3.14E-11	2.26E-9	amino acid and derivative metabolism
GO:0008652	0.0087	7/34	1.16E-8	8.34E-7	amino acid biosynthesis
GO:0009084	0.0032	5/34	5.39E-8	3.88E-6	glutamine family amino acid biosynthesis
GO:0006526	0.0016	4/34	1.56E-7	1.12E-5	arginine biosynthesis
GO:0009064	0.0046	5/34	3.98E-7	2.87E-5	glutamine family amino acid metabolism
GO:0000051	0.0021	4/34	5.25E-7	3.78E-5	urea cycle intermediate metabolism
GO:0006525	0.0021	4/34	5.25E-7	3.78E-5	arginine metabolism
GO:0006555	0.0023	4/34	7.32E-7	5.27E-5	methionine metabolism
GO:0000096	0.0029	4/34	2.20E-6	1.59E-4	sulfur amino acid metabolism
GO:0009066	0.0037	4/34	6.25E-6	4.50E-4	aspartate family amino acid metabolism
GO:0006790	0.0013	2/34	8.04E-4	0.058	sulfur metabolism
GO:0000103	0.0013	2/34	8.04E-4	0.058	sulfate assimilation
GO:0006791	0.0013	2/34	8.04E-4	0.058	sulfur utilization

Swi/Snf rich media – KEGG Pathway					
GMRG Term	**Pop frac**	**Study frac**	**e_r**	**e_s**	**Description**
sce00220	0.0024	4/34	9.94E-7	3.78E-5	Urea cycle and metabolism of amino groups
sce00330	0.0039	4/34	7.47E-6	2.84E-4	Arginine and proline metabolism
sce00290	0.0026	3/34	8.08E-5	0.003	Valine, leucine and isoleucine biosynthesis
sce00252	0.0044	3/34	4.05E-4	0.015	Alanine and aspartate metabolism
sce00620	0.0055	3/34	0.001	0.031	Pyruvate metabolism
sce00920	0.0018	2/34	0.002	0.059	Sulfur metabolism

Table 3