Bioinformatics Advance Access published August 4, 2015

Genome analysis

ERC Analysis: web-based inference of gene function via Evolutionary Rate Covariation

Nicholas W. Wolfe and Nathan L. Clark*

Department of Computational and Systems Biology, University of Pittsburgh, 3501 Fifth Avenue, BST3 - 3064, Pittsburgh, PA 15260, USA

Associate Editor: Prof. Alfonso Valencia

ABSTRACT

Summary: The recent explosion of comparative genomics data presents an unprecedented opportunity to construct gene networks via the Evolutionary Rate Covariation (ERC) signature. ERC is used to identify genes that experienced similar evolutionary histories, and thereby draws functional associations between them. The ERC Analysis website allows researchers to exploit genome-wide datasets to infer novel genes in any biological function and to explore deep evolutionary connections between distinct pathways and complexes. The website provides 5 analytical methods, graphical output, statistical support, and access to an increasing number of taxonomic groups.

Availability and Implementation: All ERC-based analyses and data are available at http://csb.pitt.edu/erc_analysis/

Contact: nclark@pitt.edu

1 INTRODUCTION

Genes carry out their functions in a complex network of interactions, both physical and genetic. These interactions have effects on gene evolution for multiple reasons. First, genes performing a common function will experience shared adaptive and conservative evolutionary pressures, so that fluctuations in these pressures affect genes throughout the network. Second, changes in one gene product can affect the function of physically interacting gene products and thereby encourage compensatory changes. Both effects are hypothesized to cause genes participating in a common function to experience evolutionary rates that covary over time. This signature of co-functionality is quantified by the evolutionary rate covariation (ERC) statistic.

ERC measures the gene-by-gene correlation of evolutionary rates over a phylogeny of species, allowing the extraction of genes that have experienced parallel evolutionary histories. Generally, ERC values are significantly elevated between genes that encode proteins in a common complex, metabolic cascade, or genetic pathway, and this observation extends across divergent taxonomic groups including prokaryotes, fungi, insects, and mammals (Juan *et al.*, 2008; Clark *et al.*, 2013; Findlay *et al.*, 2014). In practical application, ERC can be used to discover new functionalities via a sort of "guilt by association." Recently, ERC signatures were used to expand the gene network controlling mating response in *Drosophila* (Findlay *et al.*, 2014). In this study, ERC signatures between known pathway genes and an initial pool of 664 candidates were used to identify 6 new genes in the pathway. A similar study used ERC in yeast and *Drosophila* to discern the functional roles of homologous recombination genes (Godin *et al.*, 2015). For human genes, ERC has been shown to efficiently identify causal disease genes out of chromosomal regions or other large sets of candidate genes (Priedigkeit *et al.*, 2015). These successful applications make a compelling case for public access to ERC datasets so that geneticists and evolutionary biologists may similarly infer functional relationships between genes.

We have generated genome-wide ERC datasets focused on 3 taxonomic groups including humans and popular genetic models. The size of these datasets prohibits casual browsing and analysis, and requires custom computational tools and statistical tests in order to be used effectively. The aim of the ERC Analysis website is to allow users to quickly perform custom ERC-based inferences for specific gene sets of interest.

2 IMPLEMENTATION OF ERC ANALYSIS

2.1 Datasets: scope and generation

The website, written in PHP and Perl, provides rapid, custom analysis of ERC datasets in 3 taxonomic groups – mammals, *Drosophila*, and yeasts – with near genome-wide coverage at 87%, 63%, and 67% gene coverage, respectively, and with the capacity to host additional taxonomic groups as their datasets are generated. The mammalian dataset employs 33 species, *Drosophila* 12 species, and yeasts 18 species, and these numbers will be augmented as more genomes become available. Protein-coding gene models and annotations were centered on a focal species for each taxonomic group. Mammals were based on human "UCSC genes" as annotated at the UCSC Genome Browser; *Drosophila melanogaster* annotations were as in FlyBase and yeast genes as for *Saccharomyces cerevisiae* in SGD (Cherry *et al.*, 2012; Dos Santos *et al.*, 2015; Karolchik *et al.*, 2014). The website provides statistical analysis with rapid, memory mapped retrieval of data and precomputed permutation test P-values fit to each application, dataset, and sample size.

The underlying ERC datasets were calculated as previously described (Clark *et al.*, 2012; Findlay *et al.*, 2014; Priedigkeit *et al.*, 2015). Generally, each protein was compared to its orthologs within its taxonomic group to estimate the rate of amino acid divergence along each branch of its phylogenetic tree. Each protein has a unique pattern of fast and slow branches relative to other proteins. Using these characteristic patterns, we calculated each rate

*To whom correspondence should be addressed.

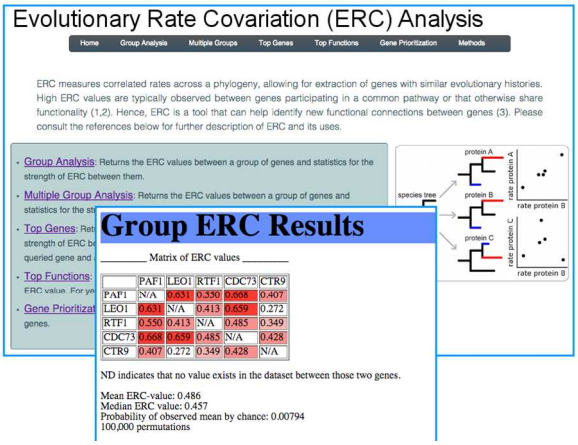

Fig. 1. The ERC Analysis website offers 5 functions and provides results complete with tables and statistical analysis for each custom query.

covariation value (ERC) as the Pearson correlation of branch-specific rates between a pair of proteins (Sato *et al.*, 2006).

2.2 Analysis Tools

Analytical tools were written to perform the tasks most commonly requested by geneticists and molecular, cellular, and evolutionary biologists. Broadly, these tasks i) describe evolutionary correlations between co-functional genes, ii) discover new candidate genes for a specific biological process, and iii) prioritize candidate genes for experimental validation (Fig. 1). Each of these analyses, detailed below, provides an easy-to-interpret results page complete with statistical support. When possible, output is available as a hyperlinked table or tab-delimited text to accommodate downstream computational analyses.

2.2.1 Group Analysis This tool allows the user to quickly measure the strength of the ERC signature within a group of genes known or suspected to be functionally related. This analysis is an important first step to assess the underlying strength of ERC in a given function before proceeding to subsequent analyses below. The results provide the pairwise matrix of ERC values between all listed genes along with a permutation-based P-value reflecting the elevation of the mean ERC within the gene set compared to random sets of matched size. An additional option can be chosen to cluster correlated genes within the matrix.

2.2.2 Multiple Groups The user submits multiple sets of genes, and the tool creates a matrix of all pairwise ERC values for each, their mean and median, their permutation P-value, and a Wilcoxon test to assess significant differences between groups.

2.2.3 Top Genes This popular analysis reveals the most correlated genes for a given input gene. It provides the biologist with a list of candidate genes enriched for functional association with the input gene. The tool returns an ERC-ranked list of genes along with their empirical rank-based P-value and a functional description of each gene.

Top Genes will search either genome-wide or within a user-defined set of query genes. The ability to search within a specific set of query genes based on secondary criteria has proven highly beneficial for ERC applications (Findlay *et al.*, 2014).

2.2.4 Top Functions Similar to Top Genes, this tool takes an input gene and returns the set of complexes or pathways associated with it via the ERC network. This tool is useful to provide candidate functions for poorly characterized genes.

2.2.5 Gene Prioritization When a biomedical study results in a list of candidate genes – from a mapping or association study for example – it is often difficult to experimentally test all candidates. This tool prioritizes those candidates for functional validation using ERC signatures with genes known to affect the phenotype in question. The user submits a set of training genes – from a particular disease or pathway for example – and a set of candidates. The tool then provides a ranked list of those candidates and statistics of their association with the training set. This type of analysis was first demonstrated in the context of human disease gene mapping (Priedigkeit *et al.*, 2015).

3 CONCLUSION

The evolutionary history of a gene can be a powerful predictor of its function and can rapidly provide novel candidates for genetics studies. The ERC Analysis website provides this information to all biomedical researchers in a simple interface. Furthermore, because the power of ERC analysis generally improves with more species, the website is continually updated with increasingly powerful datasets gained by incorporation of newly sequenced genomes.

Funding: This work was supported by a Charles E. Kaufman New Investigator research grant from The Pittsburgh Foundation (KA2014-73920) and by a Pilot Grant from the NIH-supported Pittsburgh Center for Kidney Research (P30 DK079307).

REFERENCES

Cherry,J.M. *et al.* (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.*, **40**(Database issue), D700–D705.

Clark,N.L. *et al.* (2012) Evolutionary rate covariation reveals shared functionality and co-expression of genes. *Genome Res.*, **22**(4), 714–720.

Clark,N.L. *et al.* (2013) Evolutionary rate covariation involving meiotic proteins results from fluctuating evolutionary pressure in yeasts and mammals. *Genetics*, **193**(2), 529–538.

Dos Santos,G. *et al.* (2015) FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res.*, **43**(D1), D690—D697.

Findlay,G.D. *et al.* (2014) Evolutionary rate covariation identifies new members of a protein network required for Drosophila female post-mating responses. *PLOS Genet.*, **10**(1), e1004108.

Godin,S.K. *et al.* (2015) Evolutionary and functional analysis of the invariant SWIM domain in the conserved Shu2/SWS1 protein family from *Saccharomyces cerevisiae* to *Homo sapiens*. *Genetics*, **199**(4), 1023-1033.

Juan,D. *et al.* (2008) High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc. Natl. Acad. Sci. USA*, **105**, 934-939.

Karolchik,D. *et al.* (2014) The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.*, **42**(1), D764-D770.

Priedigkeit,N.M. *et al.* (2015) Evolutionary signatures amongst disease genes permit novel methods for gene prioritization and construction of informative gene networks. *PLOS Genet.*, **11**(2), e1004967.

Sato,T. *et al.* (2005) The inference of protein- protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics*, **21**, 3482–3489.