Genome Medicine

METHOD Open Access

Pathprinting: An integrative approach to understand the functional basis of disease

Gabriel M Altschuler[1], Oliver Hofmann[1,2,5], Irina Kalatskaya[3], Rebecca Payne[1], Shannan J Ho Sui[1,2], Uma Saxena[1], Andrei V Krivtsov[4], Scott A Armstrong[4,5], Tianxi Cai[1], Lincoln Stein[3] and Winston A Hide[1,2,5*]

Abstract

New strategies to combat complex human disease require systems approaches to biology that integrate experiments from cell lines, primary tissues and model organisms. We have developed Pathprint, a functional approach that compares gene expression profiles in a set of pathways, networks and transcriptionally regulated targets. It can be applied universally to gene expression profiles across species. Integration of large-scale profiling methods and curation of the public repository overcomes platform, species and batch effects to yield a standard measure of functional distance between experiments. We show that pathprints combine mouse and human blood developmental lineage, and can be used to identify new prognostic indicators in acute myeloid leukemia. The code and resources are available at http://compbio.sph.harvard.edu/hidelab/pathprint

Background

Complex human diseases arise from perturbations of the cellular system [1]. Defining these changes from a systems biology perspective provides the opportunity to relate the function of genes, pathways, and processes. The ability to compare experiments across model organisms and humans directly influences our capacity to determine the basis of disease [2-4], and the importance of cross-species data analysis has been well illustrated: human disease genes have been identified by large-scale meta-analysis of conserved human-mouse co-expression [5], gene-based cross-species distance metrics have highlighted diseases that activate similar human and mouse pathways [6], and oncogenetic expression signatures have been prioritized by comparing human cancer and mouse model expression profiles [7-9]. Gene expression provides the most extensive resource to profile functional changes, and the opportunity for large-scale meta-analyses has been made possible by the development of public data repositories such as the National Center for Biotechnology Information Gene Expression Omnibus (GEO) [10] and the European Bioinformatics Institute ArrayExpress [11]. Cross-study analysis and integration is an area of extremely active research; however, most gene-based approaches are confounded by the challenge of comparing gene activity between different platforms and species. Consistent and scalable methods for combining these data are now required so that researchers can perform comprehensive integration of existing knowledge with new experiments, identify consistent signals, compare heterogeneous data, and validate hypotheses.

Methods for cross-study integration of gene expression data have tended to focus on differential expression in well-matched control and experimental samples [12], because approaches based on correlation or absolute profiles [13] are dominated by laboratory and platform variability in cross-study analyses [14]. The ability to leverage public data to address platform-effects has been demonstrated most recently by the Gene Expression Barcode (GEB) and Gene Expression Commons, both of which define absolute gene expression scores based on a background distribution [15,16]. However, by virtue of their reliance on gene level comparisons, these compelling simplifying approaches are restricted to selected platforms, and so do not address global comparison of biological function across experiments and species.

We sought to develop a new function-based approach for comparing profiles, which can truly scale across the diversity of available experiments, platforms, and species. Expression of biological functions across batches and divergent expression platforms shows higher concordance than across genes [17], and assigning genes to pathways

* Correspondence: whide@hsph.harvard.edu
[1]Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, USA
Full list of author information is available at the end of the article