Heat*seq: an interactive web tool for high-throughput sequencing experiment comparison with public data

Guillaume Devailly*, Anna Mantsoki and Anagha Joshi*

Department of Developmental Biology, The Roslin Institute, University of Edinburgh, Easter Bush Campus, Midlothian EH25 9RG, UK

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on April 18, 2016; revised on June 6, 2016; accepted on June 20, 2016

Abstract

Summary: Better protocols and decreasing costs have made high-throughput sequencing experiments now accessible even to small experimental laboratories. However, comparing one or few experiments generated by an individual lab to the vast amount of relevant data freely available in the public domain might be limited due to lack of bioinformatics expertise. Though several tools, including genome browsers, allow such comparison at a single gene level, they do not provide a genome-wide view. We developed Heat*seq, a web-tool that allows genome scale comparison of high throughput experiments chromatin immuno-precipitation followed by sequencing, RNA-sequencing and Cap Analysis of Gene Expression) provided by a user, to the data in the public domain. Heat*seq currently contains over 12,000 experiments across diverse tissues and cell types in human, mouse and drosophila. Heat*seq displays interactive correlation heatmaps, with an ability to dynamically subset datasets to contextualize user experiments. High quality figures and tables are produced and can be downloaded in multiple formats.

Availability and Implementation: Web application: http://www.heatstarseq.roslin.ed.ac.uk/. Source code: https://github.com/gdevailly.

Contact: Guillaume.Devailly@roslin.ed.ac.uk or Anagha.Joshi@roslin.ed.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

High throughput sequencing is now becoming routine for many biological assays including transcriptome analysis through RNA-sequencing (RNA-seq), or transcription factor (TF) binding sites identification through chromatin immuno-precipitation followed by sequencing (ChIP-seq). Additionally, collaborative projects such as Bgee (Bastian et al.,), ENCODE (Bernstein et al., 2012) and Roadmap Epigenomics (Kundaje et al., 2015) have generated genome-wide datasets across hundreds of cell types or tissues. Despite this large data being freely available in the public domain, the lack of computational tools accessible to experimental scientists with no or elementary computational skills prohibits the use of this data to its full potential for discovery.

Though genome browsers, including summary tracks provided by many consortia, are extremely useful to study a few genes, promoters or single nucleotide polymorphisms, they lack the genome-wide overview. Only a few public resources such as the CODEX database (Sánchez-Castillo et al., 2015a) and the BLUEPRINT GenomeStats tool (Zerbino et al., 2014) allow a genome-wide comparison with the user data. We therefore developed Heat*seq, a free, open source, web application providing fast and interactive comparison against high throughput sequencing experiments in the public domain. Users can upload a processed text file containing either gene expression value (Fragments Per Kilobase of transcript per Million (FPKM) or Tags Per Million (TPM)), peak coordinates or peak coordinates and corresponding expression value for CAGE/Cap Analysis of Gene Expression.
Expression). The application provides clustered correlation heatmaps, summarising global similarities between all samples in the dataset and the user sample. Heat*seq provides over 12,000 publicly available genome-wide experiments in human, mouse, and drosophila for fast and interactive comparison. In summary, Heat*seq is an interactive web tool that allows users to contextualize their sequencing data with respect to vast amounts of public data in a few minutes without requiring any programming skills.

2 Methods
2.1 Data collection
We collected gene expression data (RNA-seq), TF ChIP-seq data and CAGE data (over 4000 individual experiments) from Bgee (Bastian et al., 2008), Blueprint epigenome (Pradel et al., 2015), CODEX (Sánchez-Castillo et al., 2015b), ENCODE (Bernstein et al., 2012), FANTOM5 (Forrest et al., 2014), FlyBase (Atrill et al., 2016), GTEx (Lonsdale et al., 2013), modENCODE (Celniker et al., 2009) and Roadmap Epigenomics (Bernstein et al., 2010), in human, mouse, and drosophila. Data formatting was done using R (scripts available on GitHub). The source for each dataset is listed in Supplementary Table S1. Heatmaps represent Pearson’s correlation values between experiments calculated using a Gene x Experiment numeric matrix with gene expression values for expression data (log scaled), a Genomic regions × Experiments binary matrix indicating presence or absence of a peak for TF ChIP-seq data and a Genomic regions × Experiments numeric matrix of expression values for CAGE data (log scaled). Importantly, we constructed a metadata table which provides a web-link to original data and allows users to sub select each dataset.

2.2 Web-application development
Heat*seq is an R shiny open source interactive tool which computes correlation values between the user file and each experiment in a dataset.

Detailed user instructions are on the application website.

3 Results
3.1 Application description
Heat*seq tool supports three data types: HeatRNAseq, HeatChIPseq and HeatCAGEseq. Data upload, correlation calculation and heatmap generation takes about a minute. Importantly, users can interactively sub select relevant experiments using the metadata information (e.g. cell type, TF name). The interactive heatmap also allows selecting different clustering methods as well as zooming in and out on the heatmap. The high resolution figures and tables can be downloaded in multiple formats. Thus, Heat*seq provides global overview of relationships between public experiments and the user data. Four user scenarios are discussed below.

3.2 User scenarios
3.2.1 User data quality control
We compared a Neocortex, 10 days post-partum (Ray et al., 2015) RNA-seq sample with Bgee mouse RNA-seq data using HeatRNAseq. The top five correlation values (Pearson Correlation Coefficient > 0.9) correspond to Bgee brain samples (Supplementary Table S2). Thus, Heat*seq can be used as a fast data quality check for next-generation sequencing data.

3.2.2 Cell context identification
An oestrogen receptor (ER) alpha ChIP-seq in MCF7 cells (Zhuang et al., 2015) comparison to the ENCODE TFBS dataset by sub-selecting ENCODE ER ChIP-seq experiments revealed that the binding pattern of ERx in MCF7 cells was more similar to its binding pattern in T-47D cells than in ECC-1 cells (Fig. 1A). MCF7 and T-47D were derived from mammary tumours while ECC-1 is an endothelial cell line.

3.2.3 New hypotheses by data integration
CpG islands (CGI) from the UCSC (Karolchik et al., 2004) comparison to HeatChIPseq found that RNA polymerase II and TAF1 (Supplementary Table S4) were enriched at CGIs, as ~50% of human gene promoters contain a CGI (Illingworth and Bird, 2009). Interestingly, we identified factors avoiding CGIs including MAFK, GATA3 and ZNF274.

Similarly, tRNA promoters were highly correlated with RNA polymerase III, and its co-factors BDPI, RPC155 and BRF1 (Supplementary Table S4) using HeatChIPseq. Interestingly, comparison with BRF family data revealed that BRF1, but not BRF2 was bound at tRNA genes (Supplementary Fig. S1B).

3.2.4 Public data assessment
Heat*seq can be used to assess data in the public domain, highlighted by two examples below amongst others:
A MYC ChIP-seq in H1-hESC cells does not cluster with other ENCODE MYC ChIP-seq experiments (Fig. 1C), including H1-hESC sample from a different experimental group (Deavily et al., 2015).

Two out of seven erythroblast RNA-seq samples from the Blueprint Epigenome consortium are more correlated with endothelial cells than with the rest of the erythroblast samples (Fig. 1D).

4 Conclusion
With Heat*seq, comparing RNA-seq, ChIP-seq or CAGE experiments to hundreds of publicly available datasets becomes a trivial task. Researchers can now investigate the relationships between...
various high-throughput sequencing experiments fast and inter-
actively without requiring any programming skills. Such analysis can
assess data quality, cell variability and generate novel regulatory
hypotheses.

Acknowledgements
We would like to thank Barry Horne for the R shiny server set-up and admin-
istration and the Edinburgh R user group (EdinR) for their advice and support.

Funding
A.J. is a Chancellor’s fellow and AJ lab is supported by institute strategic
funding from BBSRC and Biological Sciences Research Council
(BBSRC, BB/J004233/1). G.D. is funded by the People Programme (Marie
Curie Actions FP7/2007-2013) under REA grant agreement No PCOFUND-
GA-2012-600181.

Conflict of Interest: none declared.

References
Attrill,H. et al. (2016) FlyBase: establishing a Gene Group resource for
Drosophila melanogaster. Nucleic Acids Res., 44, D786–D792.
Bastian,F. et al. (2008) Bgee: integrating and comparing heterogeneous transcrip-
tome data among species. In: Barroch,A., Cohen-Boulakia,S. and
Frodevaux,C. (Eds.) Data Integration in the Life Sciences. Berlin, Heidelberg:
Springer Berlin Heidelberg, DILS 2008, LNBI 5109, pp. 124–131.
Bernstein,B.E. et al. (2010) The NIH Roadmap Epigenomics Mapping
Consortium. Nat. Biotechnol., 28, 1045–1048.