METHOD ARTICLE

A multi-tool recipe to identify regions of protein-DNA binding and their influence on associated gene expression [version 1; referees: 1 approved, 1 approved with reservations]

Daniel Carlin [1], Kassi Kosnicki[1], Sara Garamszegi[2], Trey Ideker[1], Helga Thorvaldsdóttir[2], Michael Reich[1], Jill Mesirov[1,3]

[1]The University of California, San Diego School of Medicine, 500 Gilman Dr, La Jolla, CA, 92093, USA
[2]Broad Institute, Cambridge, MA, 02142, USA
[3]Moores Cancer Center, University of California, San Diego, La Jolla, CA, 92093, USA

v1 **First published:** 06 Jun 2017, **6**:784 (doi: 10.12688/f1000research.11616.1)
Latest published: 06 Jun 2017, **6**:784 (doi: 10.12688/f1000research.11616.1)

Abstract

One commonly performed bioinformatics task is to infer functional regulation of transcription factors by observing differential expression under a knockout, and integrating DNA binding information of that transcription factor. However, until now, this this task has required dedicated bioinformatics support to perform the necessary data integration. GenomeSpace provides a protocol, or “recipe”, and a user interface with inter-operating software tools to identifying protein occupancies along the genome from a ChIP-seq experiment and associated differentially regulated genes from an RNA-Seq experiment. By integrating RNA-Seq and ChIP-seq analyses, a user is easily able to associate differing expression phenotypes with changing epigenetic landscapes.

This article is included in the Galaxy gateway.

This article is included in the GenomeSpace collection.

Open Peer Review

Referee Status: ? ✓

	Invited Referees 1	2
version 1 published 06 Jun 2017	? report	✓ report

1 **Isha Sethi**, Dana Farber Cancer Institute & Harvard School of Public Health, USA

2 **Andrew D Sharrocks**, University of Manchester, UK
Munazah Andrabi, University of Manchester , UK

Discuss this article

Comments (0)

Corresponding author: Daniel Carlin (dcarlin@ucsd.edu)

Competing interests: No competing interests were disclosed.

How to cite this article: Carlin D, Kosnicki K, Garamszegi S *et al.* **A multi-tool recipe to identify regions of protein-DNA binding and their influence on associated gene expression [version 1; referees: 1 approved, 1 approved with reservations]** *F1000Research* 2017, **6**:784 (doi: 10.12688/f1000research.11616.1)

Grant information: This work was funded by the National Human Genome Research Institute, NIH U41HG007517.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 06 Jun 2017, **6**:784 (doi: 10.12688/f1000research.11616.1)

Introduction

The genetic make-up of an organism plays a key role in gene regulation, especially during early cell differentiation and development. We can observe this phenomenon in siblings who possess different eye and hair color as a result of differing genetic code. However, epigenetic mechanisms, such as histone modifications, transcription factor binding and DNA methylation, also contribute to the complexity of individuals' phenotypes as is observed in identical twins who possess the same genetic code while having slightly different features. Phenotypic differences associated with disease and varying stages of development have been mapped to changing patterns in gene regulation; and phenotype can often be attributed to a changing epigenetic landscape rather than hard-coded genetic features.

In order to decode these epigenetic differences, biologists often turn to an analysis based on two experimental assays; RNA sequencing (RNA-Seq) (Nagalakshmi *et al.*, 2008; Wilhelm *et al.*, 2008), which quantifies the amount of (usually messenger) RNA in a cell, and Chromatin Immuno-precipitation sequencing (ChIP-seq) (Johnson *et al.*, 2007; Robertson *et al.*, 2007), which shows where a particular protein binds the genome. Commonly, this protein is expected to have some influence on the mRNA expression of nearby genes (i.e., it is a transcription factor). Thus, by knocking out the gene that codes for the DNA binding protein and observing changes in mRNA expression, the biologist can infer the direct effect of the protein on expression.

When analyzing genomic data, today's computational biologist may utilize a variety of different tools specific to each step of their analysis process. Not only must they be able to create the perfect marriage between the type of data and the tool, but they must be able to correctly manipulate the output, both for interpretation and for format conversion between tools. For the non-programming biologist, smooth integration of many of these tools is provided through GenomeSpace (Qu *et al.*, 2016, www.genomespace.org) and its user-friendly "recipes" (recipes.genomespace.org). GenomeSpace is a web-based visual workbench that supports a diverse range of bioinformatics tools and data resources popularly used in genomic analyses. Because GenomeSpace provides the ability to reformat data as it moves between software tools, one can create easy to use step-by-step workflows specific to a given analysis task. We refer to these published workflows as "recipes".

We present one such recipe, currently available in GenomeSpace, which identifies differentially expressed genes between two samples, and compares that gene list with differential transcription factor occupancy from a ChIP-Seq experiment. This recipe is designed to elucidate which DNA-protein binding events are responsible for an observed change in mRNA expression. By identifying protein occupancies throughout the genome and comparing them to observed differences in mRNA expression, we can support hypotheses of functional regulation.

Methods

This recipe takes as input the aligned reads from a differential RNA-seq transcription factor knockout experiment, and aligned reads from a ChIP-Seq experiment for the transcription factor that was knocked out. The output is a visualization of the genomic regions containing both differentially expressed genes and a binding site for the transcription factor. Since all tools used in this recipe are hosted remotely, running the recipe has no system requirements beyond an internet connection. We describe the individual steps of the recipe here.

Obtaining and loading data

We start by obtaining a reference genome matching our model organism and aligning RNA-seq reads from two or more conditions (e.g. experimental and control) and ChIP-Seq reads from at least two samples, an input control and an experiment. In ChIP-Seq, the input control is a sample that has been run through all of the same preparatory and sequencing steps as the experiment, except for the antibody binding. This controls for the natural background of reads that are not selected by the binding of the target protein. Both RNA-seq and ChIP-Seq read data are uploaded to GenomeSpace in the BAM (Binary sequence Alignment MaP) format and the reference genome in the GTF (Gene Transfer Format).

Differential gene expression analysis

We next perform differential expression analysis using GenePattern (Reich *et al.*, 2006, genepattern.broadinstitute.org), which can be launched from the GenomeSpace user interface. We use GenePattern's *Cuffdiff* module to identify genes with differential expression between samples, measured by their FPKM (Fragments Per Kilobase of transcript per Million mapped reads) value. For each condition, we input the read data for an individual sample followed by the GTF reference genome. The output of the differential analysis is exported to GenomeSpace in *Cuffdiff's* tabular format.

Filtering and formatting differential gene expression data

We next launch Galaxy (Afgan *et al.*, 2016; Giardine *et al.*, 2005, galaxyproject.org), again available through the GenomeSpace interface, and import RNA-seq reads from both conditions along with a file containing differential expression for each gene. This data is directly available through GenomeSpace. Using a Galaxy workflow, we filter genes that are significantly (q-value < 0.05) differentially expressed between the experiment (in this case a knockout) and control samples and extract their chromosome number, gene region start, gene region end, and gene symbol. Next we use Galaxy's SAMtools (Li *et al.*, 2009) *Filter* subtool, which extracts this data from the original RNA-seq reads in the BAM format. We convert the BAM files to the bigWig format so that they can be viewed in the Integrative Genomics Viewer (IGV) (Robinson *et al.*, 2011; Thorvaldsdottir *et al.*, 2013).

Identifying transcription factor binding sites

Next, we use GenomeSpace to import the ChIP-seq files from both the input control and experimental samples to Galaxy. Using Galaxy's MACS2 (Feng *et al.*, 2012) *callpeak* subtool, we obtain a bedGraph file containing peak-enrichment data of both our experimental and input control files. Additionally, we use the

MACS2 *callpeak* tool to identify differential peaks along the genome, indicative of transcription factor binding sites, and output this data as a bedGraph file. The two bedGraph files are converted in Galaxy to the bigWig format for visualization in IGV.

Visualizing transcription factor binding sites and expression of associated genes

We next launch IGV through the GenomeSpace user interface. We select the appropriate reference genome included in IGV, and load all gene expression and peak-enrichment Bigwig files from GenomeSpace. Tracks are then scaled by group so their track heights are adjusted accordingly for better visualization.

Use case

We applied the recipe described above to an example dataset from Laurent *et al.* (2015), accession GSE6328, from NCBI's Gene Expression Omnibus (GEO) database (Barrett *et al.*, 2013; Edgar *et al.*, 2002). We can identify the interplay between the epigenetics and transcriptomics of mouse embryonic stems cells by observing how the binding of the transcription factor, *Prep1*, influences gene expression. *Prep1* is known for its contribution in embryonic development (Laurent *et al.*, 2015). In comparing genome-wide maps of mouse embryonic cells expressing *Prep1* to those that do not, we can identify potential target genes that are being differentially regulated by these binding events. One such example of this is illustrated in Figure 1. Here, the transcription factor binding site has been identified and shown to up-regulate the expression of the gene *Igf2*.

Variations of this recipe

This recipe can be used, not only to identify the regulation of genes by transcription factor binding, but also to identify any epigenetic mechanism that can be analyzed by ChIP-sequencing. For example, we can identify regions in the genome where histone modifications have occurred, and match those regions to observed changes in expression presumably resulting from the histone modifications. However, we must consider the nature of the data when selecting parameters in the MACS2 tool in Galaxy. For example, when performing peak enrichment on histone modification occupancies, a user must select an advanced option to include broader regions, since histone modifications are represented by a much broader peak area along the genome.

Figure 1. Epigenetic landscape of Prep1 binding and associated regulation of Igf2. The left panel illustrates the binding of the Prep1 transcription factor. In the right panel, we see the up-regulation of the gene, Igf2, as a result of this binding event.

Data availability

The original ChIP-seq and RNA-seq data of this experiment have been deposited in GEO, with accession number GSE63282. The recipe providing all the detailed steps and corresponding videos associated with this process is accessible at: http://recipes.genomespace.org/view/69.

Author contributions

DC, KK and SG designed the software protocol. KK and SG prepared a first draft of the manuscript. DC and JM finished the manuscript. JM, TI, and HT oversaw the administration and management of this project.

Competing interests

No competing interests were disclosed.

Grant information

This work was funded by the National Human Genome Research Institute, NIH U41HG007517.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

Additional members of the Mesirov Lab, specifically Ted Liefeld and Clarence Mah at the University of California- San Diego, aided in the testing and editing of this protocol.

References

Afgan E, Baker D, van den Beek M, *et al.*: **The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update.** *Nucleic Acids Res.* 2016; **44**(W1): W3–W10.
PubMed Abstract | Publisher Full Text | Free Full Text

Barrett T, Wilhite SE, Ledoux P, *et al.*: **NCBI GEO: archive for functional genomics data sets--update.** *Nucleic Acids Res.* 2013; **41**(Database issue): D991–D995.
PubMed Abstract | Publisher Full Text | Free Full Text

Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res.* 2002; **30**(1): 207–210.
PubMed Abstract | Publisher Full Text | Free Full Text

Feng J, Liu T, Qin B, *et al.*: **Identifying ChIP-seq enrichment using MACS.** *Nat Protoc.* 2012; **7**(9): 1728–1740.
PubMed Abstract | Publisher Full Text | Free Full Text

Giardine B, Riemer C, Hardison RC, *et al.*: **Galaxy: a platform for interactive large-scale genome analysis.** *Genome Res.* 2005; **15**(10): 1451–1455.
PubMed Abstract | Publisher Full Text | Free Full Text

Johnson DS, Mortazavi A, Myers RM, *et al.*: **Genome-wide mapping of *in vivo* protein-DNA interactions.** *Science.* 2007; **316**(5830): 1497–1502.
PubMed Abstract | Publisher Full Text

Laurent A, Calabrese M, Warnatz HJ, *et al.*: **ChIP-seq and RNA-seq analyses identify components of the Wnt and Fgf signaling pathways as Prep1 target genes in mouse embryonic stem cells.** *PLoS One.* 2015; **10**(4): e0122518.
PubMed Abstract | Publisher Full Text | Free Full Text

Nagalakshmi U, Wang Z, Waern K, *et al.*: **The transcriptional landscape of the yeast genome defined by RNA sequencing.** *Science.* 2008; **320**(5881): 1344–1349.
PubMed Abstract | Publisher Full Text | Free Full Text

Qu K, Garamszegi S, Wu F, *et al.*: **Integrative genomic analysis by interoperation of bioinformatics tools in GenomeSpace.** *Nat Methods.* 2016; **13**(3): 245–247.
PubMed Abstract | Publisher Full Text | Free Full Text

Reich M, Liefeld T, Gould J, *et al.*: **GenePattern 2.0.** *Nat Genet.* 2006; **38**(5): 500–501.
PubMed Abstract | Publisher Full Text

Robertson G, Hirst M, Bainbridge M, *et al.*: **Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.** *Nat Methods.* 2007; **4**(8): 651–657.
PubMed Abstract | Publisher Full Text

Robinson JT, Thorvaldsdóttir H, Winckler W, *et al.*: **Integrative genomics viewer.** *Nat Biotechnol.* 2011; **29**(1): 24–26.
PubMed Abstract | Publisher Full Text | Free Full Text

Thorvaldsdóttir H, Robinson JT, Mesirov JP: **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.** *Brief Bioinform.* 2013; **14**(2): 178–192.
PubMed Abstract | Publisher Full Text | Free Full Text

Wilhelm BT, Marguerat S, Watt S, *et al.*: **Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution.** *Nature.* 2008; **453**(7199): 1239–1243.
PubMed Abstract | Publisher Full Text

Open Peer Review

Current Referee Status: ? ✓

Version 1

Referee Report 22 August 2017

doi:10.5256/f1000research.12547.r24958

✓ **Andrew D Sharrocks** , **Munazah Andrabi**
Faculty of Life Sciences, University of Manchester, Manchester, UK

The authors present a workflow (which they refer to as a “recipe”) for the integration of RNA-Seq and ChIP-Seq experiments to find associations between genomic binding of TFs and their potential direct effects on the mRNA expression using the web-based work-bench GenomeSpace. The workflow enables integration of differential gene expression analysis following transcription factor knockdown (RNA-seq data) with binding data for the same transcription factor (ChIP-seq data). While there are many useful pipelines available for analysing and integrating sequencing data, GenomeSpace and its associated “recipes” make the analysis and integration of the data less daunting for a biologist with little or no programming experience.

Overall this “recipe” will likely be useful for biologists. However, we would like to make a few comments:

1. The authors do not make it clear whether the user has to pre-align the files to the reference genome before uploading them to GenomeSpace or the alignment itself can be done via their recipe. If not then it will be useful to provide the necessary tools and guidance required for alignment given that aligning the reads to genome is one of the most memory and time consuming steps.

2.The authors use the Cuffdiff module from the Cufflinks package to perform differential expression. Although the goal of this recipe is providing user-friendly and simplified workflow for integration of data the authors should mention the advantages of alternate tools such as EdgeR and DeSeq for identifying differential expression. These tools are available in GenePattern therefore it will be sensible to provide the user with all options. Especially since these tools are known to have better normalisation techniques and perform a more robust and reliable identification of differentially expressed genes compared to Cuffdiff.

3. The authors should not confuse the *Enriched peak* data obtained by using the MACS callpeak tool on the ChIP data and its input with *Differential peak*s. Differential peaks are obtained between two experimental conditions and not between the ChIP experiment and its input. This language can be misleading especially for beginners.

4. A flowchart of the analysis steps in the paper would be highly useful to get started.

5. Figure 1 is not entirely clear as presented. It is not clear why two separate panels are provided rather than a single panel that shows the location of the ChIP-seq peaks relative to the gene expression changes. Also, indicating what the colours represent in the gene expression data. The track labelling on the left is also not clear, and presumably “overlay” is the RNAseq data and “peaks” the ChIP-seq data.

While the current labelling is presumably driven by the naming of the original files, better labelling is suggested in the context of this figure so it is clear to the reader.
Finally, looking at the coordinates provided, we are not convinced that this is a good example to provide. The TF binding event Appears to be over 2.5Mb from the TSS of the putative target gene, meaning that any links here would be fairly low confidence. This would of course become more obvious if displayed on a single panel.

6. To make the “recipe” really useful, it would be good to have outputs beyond simple genome browser views. Having a tabular output of differentially expressed genes and the relative location(s) (and coordinates) of any binding peaks for the transcription factor in question would be useful to have.

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Referee Expertise: Gene regulation, including RNAseq and ChIPseq analysis

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 11 July 2017

doi:10.5256/f1000research.12547.r23943

? **Isha Sethi**
Dana Farber Cancer Institute & Harvard School of Public Health, Boston, MA, USA

The authors have made a pipeline integrating differential RNA-Seq expression analysis with ChIP-Seq analysis and implemented it through the GenomeSpace platform. Though as mentioned by the authors in the paper: this is a commonly performed bioinformatic task, their aim is to make this integrated analysis easily accessible to non-bioinformatic users. For this purpose their workflow on the web-based workbench involves integrating multiple tools like Cuffdiff module in GenePattern (for Differential

RNA-Seq analysis) with MACS2 in Galaxy (for ChIP-Seq analysis).

Though the workflow presented by the authors seems easy to use by any biologist, it also appears to be severely limited not just in its scope of application but also in its choice of tools which are hardcoded. For example the authors use "CuffDiff" for Differential RNA-Seq expression analysis. The authors do not state why they chose this particular method or even why they chose its GenePattern module and not the Galaxy implementation. Though admittedly this is a popular tool and has the advantage of transcript level analysis, it also suffers from known limitation of underestimating the number of differential genes. Other count based method like DESeq2 tool (also implemented in Galaxy) might be better suited for most gene-level differential RNA-Seq analysis. Also, the authors do not clearly explain why their workflow is better or easier for a biologist to implement than using the same tools through Galaxy directly (which has been made for a non-coding biologist). I would argue that working directly on Galaxy even if slightly more complicated would be more rewarding to users as it offers not just greater flexibility of tools but also the option to select different parameters than default.

Hence in conclusion, to make this manuscript better the authors should 1) provide a clearer explanation for their choice of tools and why is it easier/better to use their pipeline than the same tools on Galaxy directly, 2) If possible the authors should try to expand their workflow to provide a greater flexibility to the user to choose their tools for RNA-Seq and ChIP-Seq analysis.

Is the rationale for developing the new method (or application) clearly explained?
Partly

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Referee Expertise: Next-Generation Sequencing, Genomics, Epigenomics, Transcriptomics, Chromatin

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.