GRAVI: Gene Regulatory Analysis Using Variable Inputs

Stephen M Pederson[1,2], Geraldine Laven-Law[1], Richard Iggo[1,3], Amy Dwyer[1], Theresa E Hickey[1] and Wayne D Tilley[1]

[1]Dame Roma Mitchell Cancer Research Laboratories, Adelaide Medical School, University of Adelaide
[2]Black Ochre Data Laboratories, Telethon Kids Institute [3]Bordeaux Institute of Oncology, University of Bordeaux

Aim: Identification of Key Regulatory Targets

The GRAVI workflow was built to integrate ChIP-Seq, HiC, RNA-Seq and Histone marks to identify key regulatory targets. The workflow was designed specifically for **multiple transcription factors** to discover **shared regulatory targets which respond to a treatment** and is available from `www.github.com/steveped/GRAVI`.

The motivating question was the joint response of the Androgen Receptor (AR) and Estrogen Receptor (ER) in response to DHT-treatment across multiple cell lines, along with additional transcription factors such as GATA3 and changes to the histone mark H3K27ac. This research currently offers an exciting new therapeutic approach to ER$^+$ breast cancer[1].

GRAVI Outline

The GRAVI workflow is written in `snakemake` and produces a set of linked `html` pages with all code visible and all results presented clearly in a manner inspired by the `workflowr` package[2]. This makes for completely reproducible results which are comparable across multiple datasets (e.g. multiple cell lines or tissues), and are easily accessible for collaborators. The minimal input is one ChIP-Seq target across two conditions with 1) Annotation Preparation, 2) Peak Calling, 3) Differential Binding being always performed.

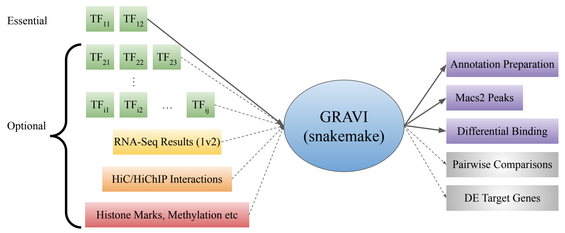

Fig. 1: **GRAVI Outline** ChIP-Seq inputs are shown in green

Incorporation of multiple ChIP targets will enable **pairwise comparisons** to be performed on differential binding results to find shared responsive regulatory regions.Likewise the addition of results from RNA-Seq analyses will provide direct evidence of changed gene expression from treatment responsive regulatory regions. Pathway enrichment is performed within each step of the analysis, with the workflow being heavily dependent on the `extraChIPs` Bioconductor package[3].

Annotation Preparation

A set of *non-overlapping genomic regions* is defined for the specific annotations utilised, and incorporating all **transcript-level** information with genomic distances defined by the user. Regions are defined as 1) Promoter, 2) Upstream Promoter, 3)Exon, 4) Intron, 5) Proximal Intergenic, and 6) Distal Intergenic. The relationship between defined regions and any provided genomic features (e.g. H3K27ac) data or any long-range interactions (e.g. HiC) is also assessed. Standardised colour schemes are also defined for propagation through the workflow.

Peak Calling

Macs2 callpeak[4] is used for identifying peaks within individual replicates and across all replicates within a treatment group. QC on each sample is performed identifying using common metrics such as FRIP with the relationship between all samples assessed using UpSet Plots and VennDiagrams. Consensus peaks are identified with treatment groups and across all samples with all regions exported as `bed` files.

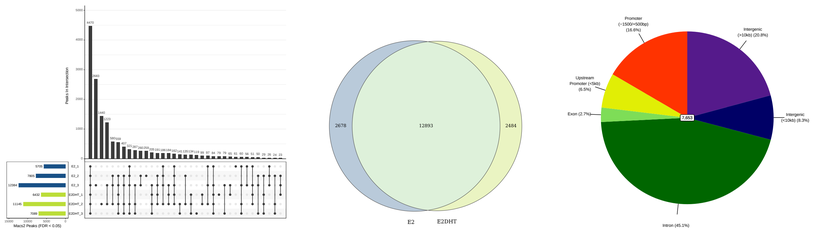

Fig. 2: Example outputs for ER binding showing all samples as an UpSet plots, common consensus peaks and the genomic distribution.

Differential Binding

Differential Binding analysis is performed using sliding windows[5] then converting retained windows to `logCPM` values and normalising using Smooth-Quantile Normalisation[6] (Figure 3) and assessing differential binding using `limma-trend`[7]. A range-based H_0[8] is used with the default set to a change in binding $< 20\%$ being not of interest. Representative windows for a combined region are chosen based on the window with maximal signal. Before final p-value adjustment, weighted p-values are calculated using Independent Hypothesis Weighting using either 1) Genomic Region, 2) External Features or 3) Detection of additional targets (Figure 4). The p-value adjustment method and significance threshold are able to be specified by the user in an experiment-wide manner. Standard sets of result tables are produced as interactive tables, along with MA-plots, profile heatmaps and genomic distributions etc.

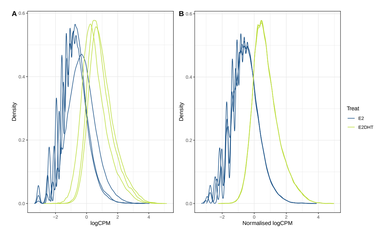

Fig. 3: Example showing logCPM values for AR before and after Smooth Quantile Normalisation. AR is primarily cytoplasmic under E2 then shifts to be nuclear under E2+DHT

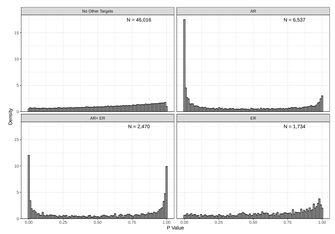

Fig. 4: Partitioned p-values for the GATA3 binding response to DHT based on co-detection of the additional ChIP targets AR and ER.

All regions are mapped to genes using `extraChIPs::mapByFeature()` which maps sites to genes based on which feature type it overlaps. **Any long-range interactions provided are used to provide more comprehensive mappings** at this point. If **RNA-Seq** results are provided an additional step is performed assessing the relationship between differentially expressed (DE) genes and any binding sites, directly identifying candidate sites with an observed response at the RNA level. Pathway enrichment analysis is also performed with and without RNA-Seq data.

Pairwise Comparisons

Given prior rigour in statistical testing, comparison across pairs of factors becomes more of a *site classification problem*. When a site is found to be responsive for one transcription factor a lower p-value threshold is applied to the second factor in any overlapping sites, ensuring **unchanged sites are classified more accurately** as these are important in a pairwise analysis. Each site is classified across both ChIP targets as *Up, Down, Unchanged or Undetected* yielding a set of pairwise classifications. Comparisons of changed binding (Figure 5) are generated across all sites, and separated by annotated region or external feature. Profile heatmaps are also created (Figure 6) and enrichment testing is performed on each set of regions. Example sites for each set of regions are also plotted by default using `extraChIPs::plotHFGC()` (Figure 7). If RNA-Seq data is provided, an analysis of DE genes by pairwise changes in ChIP target binding is also performed (Figure 8).

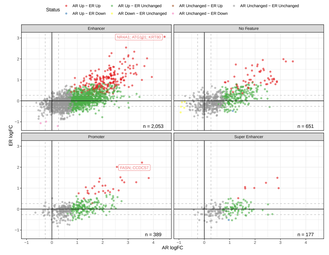

Fig. 5: Pairwise logFC values for AR and ER broken down by externally provided H3K27ac-defined features.

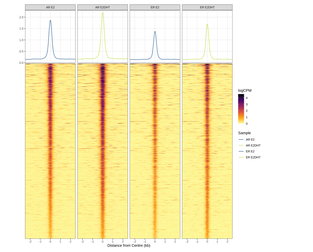

Fig. 6: Profile heatmap for sites showing increased binding for both AR and ER.

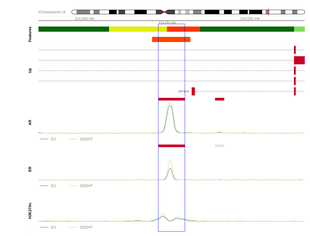

Fig. 7: *ZBTB16* is up-regulated, shows increased binding for both AR and ER, with an increase in H3K27ac signal

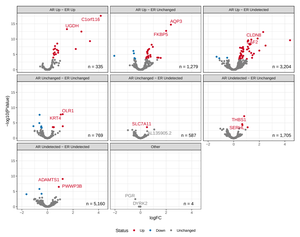

Fig. 8: DE Genes by AR and ER binding patterns. Here, the top row would be the most informative.

References

[1] Theresa E. Hickey et al. "The androgen receptor is a tumor suppressor in estrogen receptor–positive breast cancer". In: *Nature Medicine 2021 27:2* 27.2 (2021), pp. 310–320. ISSN: 1546-170X.

[2] John D Blischak, Peter Carbonetto, and Matthew Stephens. "Creating and sharing reproducible research code the workflowr way [version 1; peer review: 3 approved]". In: *F1000Research* 8.1749 (2019).

[3] Stephen Pederson. *extraChIPs: Additional functions for working with ChIP-Seq data*. R package version 1.2.2. 2022.

[4] Yong Zhang et al. "Model-based analysis of ChIP-Seq (MACS)". In: *Genome Biology* 9.9 (2008). ISSN: 14747596.

[5] Aaron T L Lun and Gordon K Smyth. "csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows". In: *Nucleic Acids Res.* 44.5 (2016), e45.

[6] Stephanie C Hicks et al. "Smooth quantile normalization". In: *Biostatistics* 19.2 (July 2017), pp. 185–198. ISSN: 1465-4644.

[7] Charity W Law et al. "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts". en. In: *Genome Biol.* 15.2 (Feb. 2014), R29.

[8] Davis J McCarthy and Gordon K Smyth. "Testing significance relative to a fold-change threshold is a TREAT". en. In: *Bioinformatics* 25.6 (Mar. 2009), pp. 765–771.