**Guide for integrating epigenetics data into GeneWeaver**

**Data types**

GeneWeaver is a functional genomics database that couples curated genomic data with graph-theoretical tools for combinatorial analysis. Integration of data related to a trait of interest is done on a gene level in GeneWeaver. Currently, GeneWeaver is used effectively to integrate certain biological measurements including transcript abundance, protein abundance, and positional candidates, into the gene ID(s) corresponding to the statistically significant biological feature(s). However, there is an abundance of genomics data available related to traits that includes measuring epigenetic and non-coding RNA features. Being able to incorporate all the genomic data available for a given trait would allow more powerful analysis and better ability to generate a testable hypothesis relating that trait to a gene (or genes). Other data types include chromatin immunoprecipitation with DNA sequencing (ChIP-seq), DNA methylation assays including bilsulphite sequencing and methylated DNA immunoprecipitation with DNA sequencing (MeDIP-seq), and non-coding RNA sequencing/microarray. These biological features are all considered epigenetic features (although non-coding RNA is considered a separate mechanism by some people) and influence traits by regulating gene expression. Because these epigenetic features exert their influence on traits through gene regulation, we can represent these features in GeneWeaver as gene sets including all genes possibly regulated by the measured epigenetic feature.

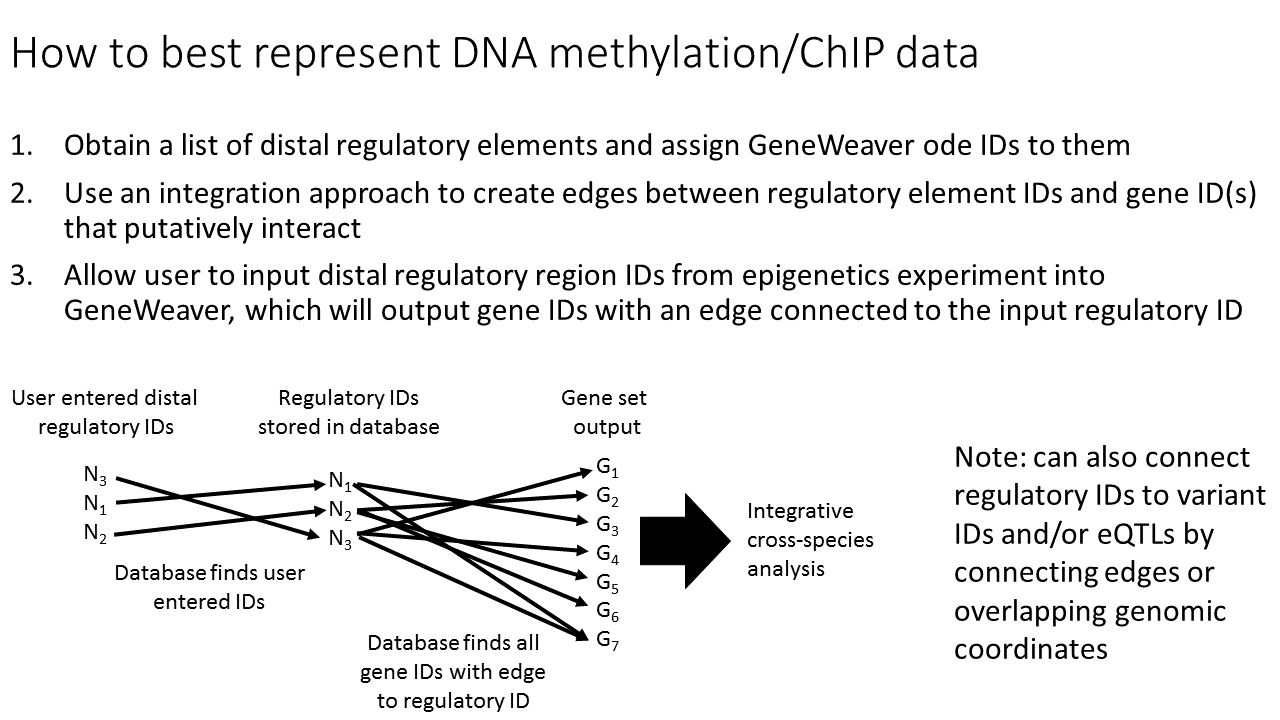
Representing these epigenetic features in GeneWeaver is advantageous for several reasons. More studies related to a given trait provide greater statistical power to discover genes involved in the trait. A greater data variety yields more information for hypothesis generation by relating a trait to a network of biological features. There is also room to expand GeneWeaver’s capabilities beyond epigenomic data. eQTL data and trait-associated variants including SNPs and indels can be integrated in GeneWeaver to obtain a more complete network of the factors influencing traits.

I will discuss how to integrate epigenetic features. Different epigenetic features have both similar and different mechanisms of influencing gene expression. DNA methylation and chromatin modifications work in concert to regulate gene expression by gene activation or heterochromatin silencing. The output of analysis pipelines for these epigenetic features is similar, and I will discuss them together. Non-coding RNAs fine tune the expression of mRNAs. This is another mechanism of gene regulation, and I will discuss integration to GeneWeaver in a separate section. In the future, eQTLs and variants can be incorporated into GeneWeaver through similar methods I am outlining. There are variant IDs which can be connected by edges to either a gene or regulatory ID if the variant lies within those genomic coordinates. eQTLs will need IDs (I’m not aware of a source of these) before they are easily integrated.

**DNA methylation and ChIP data**

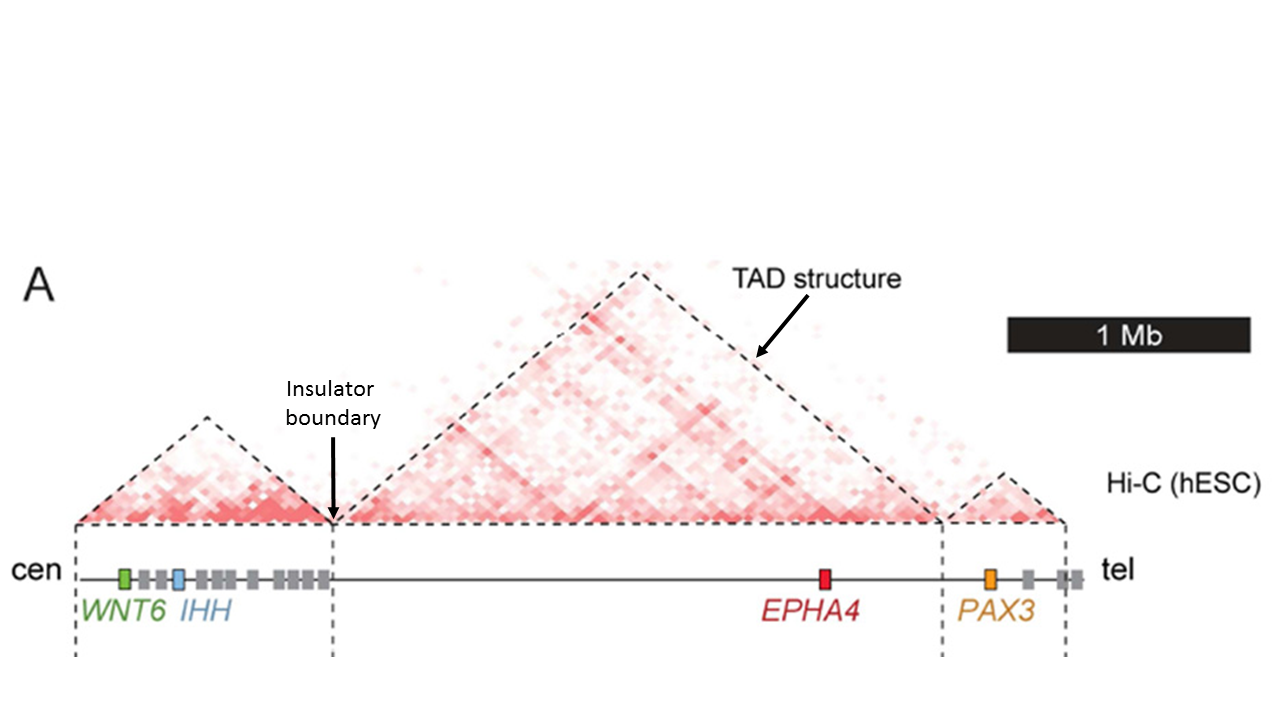
When measuring these epigenetic features, the read out is a sequence, or piece of DNA that is associated with the epigenetic modification. This genomic region is where an epigenetic modification is occurring that is correlated with a trait of interest. The challenge of representing an epigenetic feature as the genes possibly regulated by that feature is that DNA methylation and DNA associated protein modifications can occur across the genome. These epigenetic modifications can occur in open reading frames (ORFs), at promoters, or at intronic/intergenic regions. ORFs and promoters are directly related to the transcription of one gene, and the measured epigenomic feature can be assigned to that gene. The challenge of associating epigenetic features to genes comes when the read out indicates the epigenetic feature is being changed in an intronic/intergenic region.

The concept behind integrating biological data from intronic/intergenic into GeneWeaver relies on the use of an identifier list of regulatory features in the genome (an ID associated with the genomic coordinates of a regulatory region). A user would run their data analysis pipeline for their given epigenetic feature measured, which would output genomic coordinates (or peaks) associated with their measured feature. The user could use a genome browser to enter their file of peaks and determine the regulatory IDs associated with their intronic/intergenic lying peak. A user could then input the regulatory IDs they obtain into GeneWeaver. To represent these regulatory IDs as gene IDs, GeneWeaver will need to have a list of regulatory IDs which are connected individually by edges to the genes IDs that are targeted by the regulatory element (see below). The various methods of connecting regulatory IDs to gene IDs, and the data sources GeneWeaver could incorporate to perform this task, will be discussed next.



Schematic of workflow for GeneWeaver to take regulatory IDs and output Gene IDs

Intronic/intergenic regions have DNA elements such as enhancers, silencers, and insulators that impact gene expression. Enhancers bind transcription factors that help recruit/stabilize the PIC to increase gene expression. Silencers inhibit the ability of enhancers to function leading to decreased gene expression. All these interactions are occurring in the context of a 3D chromatin structure that is looping and allowing interactions between linearly distant regions of the chromosome. Insulators mediate the 3D interactions by creating chromatin looping boundaries. These boundaries are referred to as topologically associated domains (TADs) (see picture below). Chromatin segments within boundaries are significantly more likely to interact than those outside boundaries. These 3D chromatin interactions can occur from megabases away making non data-driven prediction a challenge. In WT or normal cells, TAD regions are remarkably stable from one cell type to the next. However, treatments such as alcohol could alter these TADs and chromatin interactions so predictions based on untreated cells could be invalid.



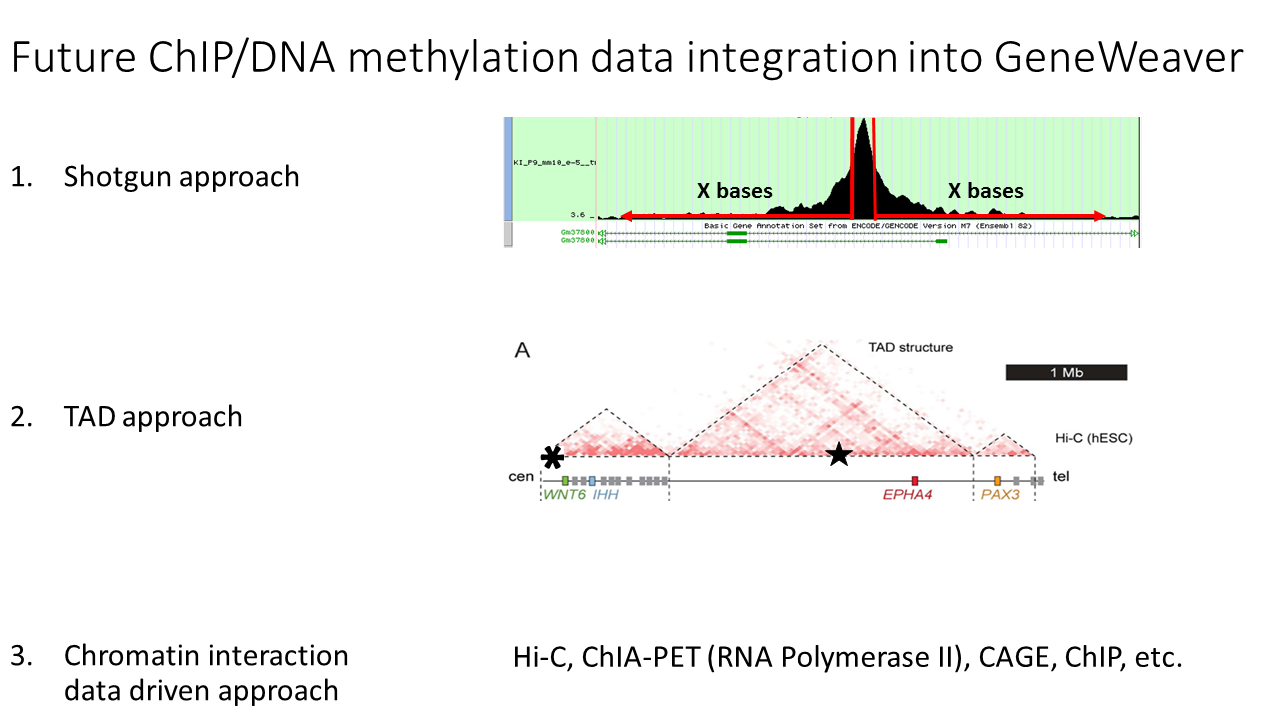
This figure from Lupiáñez et al., 2015 shows TADs as triangles with boundaries between the triangles. Darker red corresponds to increased frequency of chromatin interactions in Hi-C data. TADs have significantly more interactions within them than outside them, making them important when considering gene-enhancer interactions.

If an epigenetic feature that relates to a trait was found in an intronic or intergenic region, there are several ways to assign gene targets to intronic/intergenic epigenomic features (see below list for schematic):

1. The easiest way (would also generate the most false positives) would be to take the shotgun approach similar to old QTLs in GeneWeaver and include all genes within a certain megabase region both 5’ and 3’ to the measured epigenetic feature. It’s recently been suggested the average TAD is ~90kb (Ji, 2016). However, treatments such as alcohol could disrupt TADs and cause ectopic interactions. Therefore, to be safe, all genes within 5-10mb of the epigenetic feature should be included when taking this approach (this is open for debate though).

2. A more data oriented way to solve this problem would be a TAD approach. There are analysis pipelines for Hi-C data that predicts the genomic coordinates of TADs throughout the genome. We could design a way to take the coordinates of a TAD and find all regulatory IDs and gene IDs within that TAD. Any gene ID that lies within the same TAD as a regulatory ID can be connected with an edge.

3. The most data oriented way to solve this problem would be to take a direct chromatin interaction approach. This approach takes gene-enhancer predictions from (potentially) various sources. Enhancers from these predictions can be associated with regulatory IDs, and edges can connect the regulatory ID to the gene IDs that are predicted to interact by the data. Data includes Hi-C, ChIA-PET (for proteins such as RNA polymerase II), CAGE, and some ChIP.

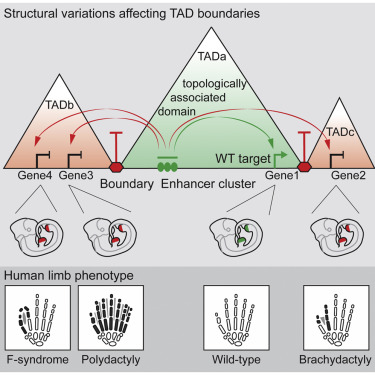


Representation of the various approaches to connecting edges between regulatory IDs and the gene IDs associated with the regulatory feature. For the TAD approach, the asterisk representing regulatory ID would be associated with WNT6 and IHH, whereas the star representing regulatory ID would be associated with EPHA4

Here is a table of various sources of data that could be used. Ensembl is critical to this idea because it has predicted regulatory elements represented with IDs. Users can use the UCSC genome browser to visualize what Ensembl regulatory IDs are associated with their data peaks. All other sources of data could be used by GeneWeaver to take Ensembl IDs and convert them into gene IDs possibly regulated by the Ensembl regulatory ID through one of the above methods.

|  |  |  |
| --- | --- | --- |
| **Source** | **Data Type** | **Species** |
| ENCODE/modENCODE | Predicted regulatory region coordinates | human, mouse, worm, fly |
| ENCODE/modENCODE | Gene-enhancer predictions | human |
| ENCODE/modENCODE | TADs of the genome | human, fly |
| Ensembl | Regulatory ID | human, mouse |
| Riken/FANTOM5 | Gene-enhancer predictions | human |
| Riken/FANTOM5 | Predicted regulatory region coordinates | human |
| Ren Lab "mouse encode" | Gene-enhancer predictions | mouse |
| Ren Lab "mouse encode" | TADs of the genome | mouse |

The caution against using data driven approaches to predicting gene-regulatory region interactions is that the data is all coming from WT or non-treated cells. Although TADs are very stable between normal cell types, treatments can affect TADs and change the interactions. Ectopic interactions can occur in organisms with diseased cells or following treatment with substances such as alcohol (see below). Ideally, Hi-C or similar chromatin interaction data would come from cells treated with (for instance) alcohol for predicting gene-regulatory region interactions in data sets following alcohol exposure (same applies for all conditions/treatments). This is one of several justifications to perform a Hi-C experiment on cells under conditions the lab is studying. For instance, it would be very helpful to do a Hi-C on alcohol treated neurons for the Chesler lab that looks at genes involved in alcohol response in GeneWeaver.

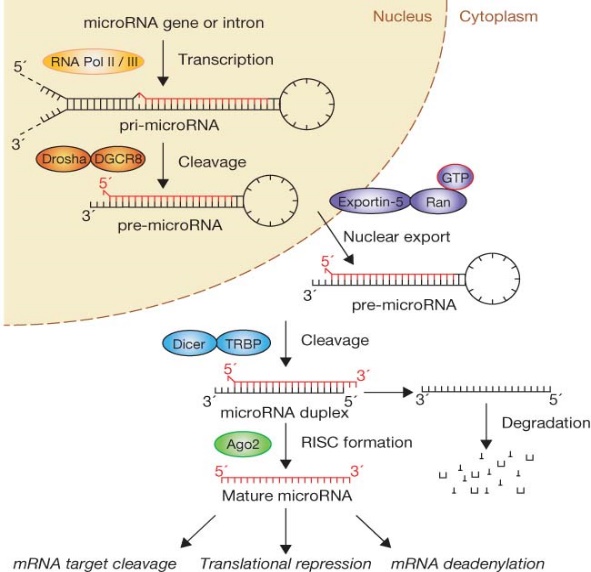


My personal recommendation would be to use a TAD driven approach with data under the same treatment/disease conditions if possible to integrate intronic/intergenic DNA methylation/ChIP data into GeneWeaver. Although TADs are stable, enhancers and regulatory elements can be very cell type specific, making the chromatin interaction driven approach risky as it might miss important interactions in different cell types. The TAD approach will generate several false positives per sample, but will not miss any important interactions. I feel this is the best approach to take for GeneWeaver using Ensembl regulatory IDs and Encode/Mouse Encode (Ren Lab) data currently available. Both Encode and Ren Lab have the pipelines they used to take Hi-C data and generate a file of predicted TADs. This pipeline could be applied to (for instance) an alcohol exposure Hi-C data set to generate a file of TADs for GeneWeaver integration.

**Non-coding RNA measurements**

For the non-coding RNA measurement section, I will be detailing only microRNA (miRNA) integration into GeneWeaver. Non-coding RNAs are an emerging field still and a lot of work needs to be put into giving them IDs and ascribing biological function to them. For example, published in Nature 6 days ago – “An atlas of human long non-coding RNAs with accurate 5′ ends” by the Riken/FANTOM5 consortium (Hon et al., 2017) is the most detailed description of non-coding RNAs to date. Yet, we still do not have IDs or know enough about the function of these RNAs to think about in GeneWeaver. These emerging fields must be followed so that information can be incorporated into GeneWeaver as it becomes available.

Currently GeneWeaver has incorporated pre-miRNA IDs from miRBase. A full list of both pre and mature miRNA IDs is available at miRBase and both need to be added (I submitted a request in GeneWeaver). Functionally, miRNAs impact gene expression at the mature miRNA stage, and so pre and mature miRNAs are functionally the same (see below). Therefore, in GeneWeaver, the list of pre-miRNAs and the list of mature miRNA IDs can be connected with edges because they ultimately will have the same effect on gene expression. Users can intersect their miRNA sets to find the most highly connected miRNAs related to a trait. However, not all miRNAs are conserved across species, and not all gene targets of homologous miRNAs are conserved across species. Integrating miRNAs related to a trait can be informative within a species, but ultimately entering miRNA data as gene sets is the best way to do cross-species analysis.



Schematic of miRNA biosynthesis pathway. Ultimately, the pre-miRNA is cleaved by Dicer and the mature miRNA associates with Argonaut2 in the RISC complex to regulate the expression of target genes.

There are several target predictors that provide gene targets of miRNAs. There is no prediction algorithm (or set of them) that is a gold standard. Most individual predictors have hundreds to thousands of predicted genes targeted by an individual miRNA. A set of ten differentially expressed miRNAs can be predicted to target half of all genes, which is not informative. To reduce the number of false positives, I used miRecords database. miRecords comprises 11 different predictive algorithms that all use different information to predict miRNA-mRNA binding and subsequent regulatory function. By setting a cutoff for number of algorithms that predicts an interaction, the false positive prediction noise can be dampened. I set cutoffs so for human miRNAs, at least 6 of 11 algorithms had to predict a gene target for me to include it in the GeneWeaver gene set related to a miRNA measurement. For mice miRNAs, at least 5 of 11 algorithms had to predict the miRNA gene target for inclusion. Other species must be explored individually. One potential issue is that a small portion of the miRNAs have 0 predicted gene targets at these cutoffs. Users risk losing valuable information with 0 predicted targets for some miRNAs in their set, but must balance this against having many thousand gene targets that has a lot of noise via false-positives.

**Note: this section can be included with the F1000 paper methods about how we successfully used target predictors for miRNAs to find genes possibly involved in alcohol response**